Ilja van Beest

Faculty of Social and Behavioral Sciences

Tilburg University

P.O. Box 90153  
5000 LE Tilburg  
The Netherlands

Prof. Dr. Nico van Yperen  
Academic Editor, PLOS ONE  
Rijksuniversiteit Groningen

Manuscript ID PONE-D-15-02806  
  
Dear Prof Dr. van Yperen,  
  
Thank you for your kind words about our work and for the opportunity to submit a revised version of our paper “The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies" to PLOS ONE. As you recommended, we have carefully considered each of the comments made by the reviewers, paying special attention to those highlighted by you in your letter. A detailed overview of our revisions is included below. For your convenience we copied the three reviews and added a detailed description of how we made the appropriate changes immediately below each comment.

We believe that the changes we made have substantially improved the manuscript and made our contribution stronger. We warmly thank you for your help in achieving this and look forward to your final decision.

Kind regards, also on behalf of Chris Hartgerink, Jelte Wicherts, and Kip Williams

Ilja van Beest

**Reviewer #1:** The authors have conducted a meta-analysis of studies using the Cyberball game, which manipulates the degree of social inclusion versus ostracism experienced by participants. Particular focus in this meta-analysis is on the immediate and delayed effects of the experimental manipulation and on examining whether immediate or delayed effects are more susceptible to the moderating influence of other factors.  
  
In general, I think that PLoS One is an appropriate outlet for this meta-analysis and I would support its publication. However, below I list a number of general issues, concerns, comments, and appeals for clarification that I think the authors need to address first.  
  
General Issues / Concerns / Comments:  
  
**#1**

page 7: Author predictions were used to determine how an interaction should be coded. Was there always a clear prediction given by authors so that this decision could be made unambiguously? If not, was the intercoder reliability of these assessments measured? Aside from this, we know that some 'predictions' are actually generated post-hoc, after the results have become available. That is a limitation that should be acknowledged.

**Answer**

*Of the 120 studies that were investigated, 52 studies contained an interaction. The prediction in these 52 studies, was based on the explicit prediction of the authors of the manuscript. Moreover, the first authors (Chris and Ilja) checked and discussed each paper until consensus was reached. We did not record these discussions and intercoder reliability cannot be assessed. However, their general impression was that the coding was fairly consistent….*

*We acknowledge that the predictions of the authors could be post-hoc and this is now acknowledged in the revised manuscript. We now say*

A potential limitation of our decision to follow the prediction of the authors is that the predictions may have been generated post-hoc on the basis of observed outcomes.

*on page 7 line 135.*

**#2**

page 10, line 208: In studies with more than one additional factor (besides the ostracism factor), the authors "collapsed effect sizes across the factor that authors expressed least interest in." I can imagine that this decision cannot always be made with 100% certainty. Did the authors attempt to estimate the intercoder reliability of these assessments?

**Answer**

*Seventeen of the 52 studies with a cross-cutting variable involved designs that were more complex than the 2x2 design. For these 17 studies, we chose…. Again, these issues were resolved by discussion and intercoder reliability was hence not assessed.*

**#3**

page 10, line 224: I know from personal experience that one of the last things that authors of a meta-analysis want to hear is: Your search is outdated. Indeed, a meta-analysis may go through through several (re)submission rounds before being accepted/published and the date of the search then increasingly falls further behind. There is in principle no need to demand an update, so I will not insist. However, are the authors aware of any additional studies that have become available after their search was concluded?

**Answer**

*We wholeheartedly agree with this reviewer that the search for studies in a meta-analysis is quite time-consuming and that one should always chose a moment to stop updating the database. Nevertheless, we conducted another search in Web of Knowledge for Cyberball studies, which resulted in 71 hits for 2013-2015 (searched on March 17, 2015). After inspecting which of these studies would have met our inclusion criteria, 29 remained after our previous end date (April 2013). These 29 references are available* [*here*](file:///C:\Users\chjh\Dropbox\shared\search.ciw) *(EndNote format). Of these 29, we already included 2 studies that were not published when we collected them, and 14 contained a cross-cutting variable. Given the current size of the database and the sample sizes in these new studies, we do not expect them to significantly change any of our core conclusions. Hence, we decided not to redo all of the analyses using this updated database.*

**#4**

page 13: I am wondering about the selection/coding of the first and last measure. Was there never any ambiguity regarding the order in which instruments were administered? Also, if authors said that they used measures X, Z, and Z after the game, the actual order may have been different.

**Answer**

*We based the coding of the first- and last measure on the information presented in the paper describing the primary study. This information was straightforward and we did not encounter ambiguity regarding the order in which the instruments were administered. We acknowledge that people may have included more measures than reported and that unreported measures remain unaccounted for, such that the estimate for time between the first and last is a crude one. In other words, we could not get better information than that reported in the paper, which is why we retain the information reported in the paper as the most viable situation.*

***#5***

page 13: First and last measures were classified into four categories (interpersonal, intrapersonal, fundamental needs, or model correspondence). So, if I understand the authors correctly, first a measure was chosen as being first/last and then this classification was made (so there is always exactly one first measure and if the study applied multiple/delayed assessments, there is always exactly one last measure). Can the authors please confirm/clarify this?

**Answer**

*We hereby confirm that every study contained a first measure and if present, a last measure. Table 2 illustrates this, where some studies do not contain an effect on the last measure.*

**#6**

page 13: Also, does that imply that the first measure may have assessed, for example, intrapersonal effects, while the last measure may have assessed, for example, interpersonal effects? Or in other words, is it possible that the effect size estimates in Table 2 (d\_T1 and d\_T2, and similarly, Delta-d\_T1 and Delta-d\_T2) actually reflect different measurement types? This needs to be clarified, since this has major implications for the interpretation of the results reported on pages 25 to 27.

**Answer**

*Yes, this is correct. Figure 2 separates the effects per type of measure and shows that results are consistent across the different types of dependent variables, except for interpersonal behavior (as mentioned in the text).*

**#7**

page 18, line 350: I am not sure if "standardized simple effects across the ostracism factor" is appropriate terminology here (and elsewhere in the paper). In a two-way factorial design, a "simple effect" is the effect of one factor \*within\* one of the levels of the other factor. So, if that other factor has two levels, then there would be two simple effects. That would apply to each time point, so in a 2x2 design with multiple measures (one of which is the first and one is the last measure), there would be 4 (not 2) simple ostracism effects. However, if I understand the authors correctly, they are not computing simple effects here, but marginal/main effects for the first and for the last measure (i.e., the difference between the ostracism and inclusion levels averaged over any other factors). Please clarify this (and the terminology throughout the manuscript).

**Answer**

*We did intend simple effects, as we calculated four simple effects for the ostracism factor (one in the moderated conditions, one in the non-moderated conditions, for both first and last measure). The reviewer refers to the set of 52 studies where a second factor is included, where we calculated the simple effect of ostracism within the non-moderated level. We clarified this in the revised manuscript. Specifically, we now write:*

Standardized effects were calculated across the ostracism factor, where the 52 studies with a cross-cutting variable were included as a simple effect of ostracism within the non-moderated level.

*On page 18 ~ line 349. Additionally, we deleted*

Non-factorial studies delivered only simple effects for the first and last measure, and no interactions

*From 355-356 to prevent confusion.*

**#8**

page 18: The description of the interaction effect given here (and on the previous pages and also the appendix) suggests that moderators of the ostracism effect can take on only two values/levels. However, was that always the case?

**Answer**

*Moderator factors could include more levels, in which case we selected the two conditions that were the farthest apart in design. For example, if a study included an ostracism factor (included or ostracized) and a players factor (3, 5, 10, 15 players) as a moderator, we used the 3 and 15 player levels. Selection based on the factorial levels occurred in 10 studies.*

Table 2:

**#9**

1) I see many rows where "First author" and "Year" is identical. Can the authors explain how this arises?

**Answer**

*We thank the reviewer for this comment. The reason is that papers may contain multiple studies. To clarify this, we now added a note.*

Multiple rows for the same first author and year is possible due to multiple studies in one paper or multiple papers in one year

**#10**

2) In the table notes, the authors write: "Non-integer Ns arise from division of full sample N for included conditions, appropriate due to random assignment." I don't understand what the authors mean by this (and I could find no further discussion of this in the paper).

**Answer**

*Ns of for example 12.333 arise from a 3-condition design, where random assignment was used. If N per condition was not given, we divide total N (e.g., 37) by the number of conditions (3) to come to a condition N estimate. To clarify we added an example in the table note:*

(e.g., two conditions out of 3, when sample is 56: (56 / 3) × 2 = 37.333)

**#11**

3) It appears that multiple estimates are often obtained from the same study. Given that "N" differs for these rows, these effects seem to be based on different samples, so within a particular study, the estimates may be independent. However, that still does not preclude the possibility that multiple estimates obtained from the same study are more similar to each other than estimates obtained from different studies. In other words, the data seem to have a multilevel structure, which would imply the need to employ an appropriate multilevel meta-analysis model that accounts for such dependencies (e.g., by adding a random effect at the study level to the current model).

**Answers**

*The reviewer notes that the data may be interdependent within an analysis; this is incorrect. Effects that go into the same meta-analysis are independent (i.e., one effect per study): every row is an independent study, which also explains the difference in N. However, the reviewer is correct in stating that from one paper multiple independent studies can be included. This multilevel modeling is therefore not necessary.*

**#12**

page 25: I assume the authors applied the version of Egger's regression test that relates the effect size estimates to their standard errors. For standardized mean differences, the standard error depends on the size of the effect, which can cause spurious associations especially when effects are large. Similar deficiencies of the test have been observed when using effect size measures based on dichotomous data (e.g., risk/odds ratios or risk differences). For a more appropriate version of the test, the authors should use some measure of precision that does not depend on the size of the effect, the obvious choices being the sample size, the inverse sample size, or square-root transformations thereof.

**Answer**

*As requested by the reviewer, we conducted these regression tests with 1/N as predictor. Results are the same as the Egger’s test with standard error as predictor and is therefore not adjusted further in the manuscript. We include a footnote on this in the manuscript that reads:*

Due to the dependency between the standardized effect size and the standard error, we ran an alternative version of the Egger’s test that regresses on 1/N. This analysis yielded highly similar results.

**#13**

page 25: Coding the estimated time between exclusion and the moment at which the last measure was taken in \*seconds\* seems artificially precise. Did the authors calculate the intercoder reliability for these estimates based on independent coders? Also, please rescale this moderator into some larger units (e.g., minutes) which avoids the extremely small coefficient (.0001). In addition, since this is one of the primary hypotheses tested in the paper, please provide a scatterplot of the time variable against the effect size estimates.

**Answer**

*Following the suggestion of the reviewer we rescaled the time estimate into minutes. The results have been adjusted accordingly.*

*Also note that the time estimation was based on the number of items times the six second rule, plus any additional time mentioned in the paper. This information was readily available in all manuscripts although we acknowledge that it is possible that not all dependent variables were disclosed in a paper describing the study (see also our answer reviewer 1, #4). As mentioned in the 68 studies without cross-cutting variable were coded by Chris Hartgerinck, the 52 with a cross-cutting variable were coded by both Chris Hartgerink and Ilja van Beest. Consensus was readily reached and we did not collect quantitative information to calculate intercoder reliability.*

*Following the suggestion of the reviewer, we now provide scatterplots of time versus effect (simple and interaction on timepoint two) in the Supplemental Materials of the revised manuscript.*

***#*14**

page 27: Same issues apply here. I cannot imagine that two independent coders would ever come to the exact same assessment when coding time in seconds. Also, please rescale time to avoid the overly small coefficient. And please provide a scatterplot.

**Answer**

*See answer (reviewer #1, answer #13).*

***#*15**

page 28 and Figure 2: As far as I can tell, here the authors are indeed talking about simple effects (e.g., "the between-subjects effect of being ostracized with no moderator present, whereas moderated ostracism effect refers to being ostracized with a moderator present"). Earlier, the authors also talked about "simple effects" (which I think are actually main effects -- see my earlier comment -- but maybe I am misunderstanding what the authors did). Please clarify this.

**Answer**

*See answer (reviewer #1 answer #7).*

**#16**

Also, if I understand Figure 2 correctly, I would assume then that the \*difference\* between, let's say, the points for "All" in panels (1) and (2) is equal to the \*difference\* between the points for "All" in panels (5) and (6) (since the difference between the two simple effects for factor A within the two levels of factor B must be equal to the difference between the two simple effects for factor B within the two levels of factor A). However, visual inspection suggests that this may not the case. Can the authors clarify?

**Answer**

*We are not sure whether we understand the question. It seems that the reviewer postulates that the difference in the simple effects for ostracism on the different moderator levels is supposed to be equal to the difference in simple effects for the moderator levels on the ostracism levels. Below we provide an example that this would be incorrect and that simple effects do differ.*

|  |  |  |
| --- | --- | --- |
|  | N-mod | mod |
| Ostr | 5 | 7 |
| Incl | 2 | 3 |

*In this case, the simple effect of ostracism is 5-2 = 3 for the non-moderator level and 7-3 = 4 for the moderated level. For the simple effect of moderator within the ostracism level, 5-7 = -2 and within the included level 2-3 = -1. Correspondingly, simple effects all differ and are not required to be equal, as the reviewer proposes.*

**#17**

page 30, line 514: "Model indicates" -- which model?

**Answer**

*The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2*

Subset All contains all measures. Subset Fundamental contains only fundamental need measures. Subset Intrapersonal contains all intrapersonal measures. Subset Interpersonal contains all interpersonal measures. Subset Model contains those where first measures is immediate and last measure is delayed. See Supplement S4.

*On page 28 this was clarified under the heading Measures, where the subsets are named.*

**#18**

page 30, lines 515-516: I don't understand what the authors mean by "listwise deletion for equal ks across time points". Please clarify.

**Answer**

*To clarify what we mean by listwise deletion we adjusted the sentence as follows:*

Listwise deletion ensures that estimates are made on full rows in the data. Listwise deletion was applied in all the subsets, which only altered results for interpersonal measures.

**#19**

page 30, line 520: What estimates did the authors use for these analyses? The estimates shown in Table 2 or the "simple effects" that went into the analyses that led to Figure 2? I assume the former values were used, but please clarify this. Also, if my assumption is correct, then as far as I can tell, listwise deletion (due to incomplete information about the predictor variables) led to the removal of 120 - 45 = 75 estimates for T1 and 95 - 41 = 54 estimates for T2. Is that correct? If so, then this should be mentioned as a limitation.

**Answer**

*The analyses were based on the ostracism effect across all 120 studies (as in Table 2 column d T1). However, due to listwise deletion the number of effects indeed reduced the number of effects included and now reads:*

To inspect for structural and sampling effects of the studies, we ran mixed-effect models on the 120 ostracism effects, on both the first and the last measure. Due to listwise deletion, only 45 of 120 effect sizes remained on the first measure and 41 of 95 effect sizes for the last measure.

**#20**

pages 30, line 527: The dfs for the Q\_E-test are 32. With k = 45, this implies that the model must have contained 45 - 32 = 13 fixed effects (including the intercept). However, in Table 4, I only count 12 coefficients.

**Answer**

*We thank the reviewer for noting this error. The dfs should indeed be 33. This is now adjusted in the revised manuscript.*

***#*21**

page 31, line 537: The dfs for the Q\_M-test are 12. Assuming that the intercept was not part of the coefficients tested, this implies that the model included 13 fixed effects. However, I only count 12 coefficients in Table 5.

**Answer:**

*We again thank the reviewer for noting this error. The df should be 11 and is adjusted in the revised manuscript.*

**#22**

page 31: Please report the results from the Q\_E-test here as well.

**Answer**:

*We added the results. On page ?? of the revised manuscript we now say:*

*QE* (29) = 214.69, p < .0001

**#23**

Tables 4 and 5: For a categorical predictor with more than 2 levels, please provide a test of the factor as a whole (i.e., an omnibus test of the coefficients corresponding to the factor). Also, the tables only show the results of tests comparing each level against the reference level, but there may be significant differences when comparing other levels against each other. Please examine/report this.

**Answer:**

*The Q\_M test is an omnibus test and is reported. The dummies are indeed only compared to the reference group. Moreover, we already included confidence intervals in the original version of our manuscript. These CIs indicate that all comparisons between these dummies will yield similar results (overlapping CIs). Indeed, the requested analyses confirmed this:*

*If we only look at the countries, QM(df = 2) = 0.3494, p-val = 0.8397, first measure, QM(df = 2) = 2.6394, p-val = 0.2672, last measure.*

*If we only look at the different needs scales, QM(df = 4) = 6.0702, p-val = 0.1940, first measure, QM(df = 4) = 0.4257, p-val = 0.9803, last measure.*

*Because these analyses provide the same information as the overlapping confidence intervals we decided not to incorporate them in the revised manuscript.*

**#24**

page 41, line 738: I don't understand what the authors mean by "difference index" or how this was coded. What "value" are the authors referring to when they write: "coded value on first measure minus coded value on last measure"? In fact, I have a hard time understanding this entire paragraph.

**Answer**

*We thank the reviewer for this comment. We wanted to explain that differences in findings between first and last measurement could not be attributed to differences in types of dependent variables. We now write:*

Importantly, we did observe that the confidence intervals of both the first and last measure did not overlap, suggesting that there is a difference in effect size between first and last measure. The question then is whether this difference is indeed caused by time of measurement or in part caused by the type of measurement used across the two different time points. This explanation can be addressed by inspecting whether the composition of measures is different across time points. On the first measure 0.84 was intrapersonal self-report, 0.02 was intrapersonal physiological, 0.01 was intrapersonal other, 0.08 was interpersonal anti-social, 0.03 was interpersonal pro-social, and 0.01 interpersonal other. On the last measure 0.79 was intrapersonal self-report, 0.04 was intrapersonal physiological, 0.02 was intrapersonal other, 0.05 was interpersonal anti-social, 0.08 was interpersonal pro-social, and 0.01 was interpersonal other. This shows that the different types of dependent variables are similarly distributed across time points (maximum discrepancy of 4.9 percentage points). Substantive differences in proportions of measures across time points are minimal and thus form an unlikely driving force for our findings.

Minor Issues:

**#25**

Maybe this term is well understood by the intended target audience, but I find the term "cross-cutting variable" less than clear. Why not just call them "other factors" or something along those lines?

**Answer**

*The term cross-cutting factor is a standard term in the Cyberball field. It refers to design in which the ostracism manipulation (inclusion vs ostracism) is orthogonally crossed with another manipulation (e.g., ingroup vs outgroup). Additionally, because we also include other moderator variables (i.e., time, structural, sampling), we use “cross-cutting” as a term to prevent confusion. Cross-cutting refers to the 52 studies that explicitly manipulated a factor in the experimental design. The other moderator variables (e.g, time, structural, sampling) were investigated for all 120 studies.*

**#26**

page 3, line 47: The "(4)" is superfluous (or also number the other moderator types).

**Answer**

*Adjusted*

**#27**

page 3, line 53: Write out "i.e." when used outside of parentheses.

**Answer**

*Adjusted (also checked rest of i.e. occurrences)*

**#28**

page 3, line 54: "an unique" should be "a unique" (the use of "a/an" is not based on the spelling of the first letter of the following word, but its pronunciation).

**Answer**

*Adjusted*

**#29**

page 7, line 150: "set-up" should be "set up" (set-up or setup is a noun).

**Answer**

*Adjusted*

**#30**

page 9, line 182: "extend" should be "extent" (the latter is the noun). And the more common phrasing would be "to a large extent".

**Answer**

*Adjusted*

**#31**

page 11, line 226: Write out the acronym (SPSP) the first time it is used.

**Answer**

*Adjusted*

**#32**

page 13, lines 291 and 293: Since you are giving examples here ("e.g.,"), the "etc." at the end is superfluous.

**Answer**

*Adjusted*

**#33**

page 14, line 301: Missing comma after "e.g.".

**Answer**

*Adjusted (checked all occurences of e.g.)*

**#34**

Table 1, table notes: I think the "whereas column wise" should be "whereas row wise".

**Answer**

*Adjusted*

**#35**

page 41, line 754: "conditional on that these measures are valid" is very odd phrasing.

**Answer**

*Deleted this sentence.*

**#36**

The Oxford comma is used inconsistently throughout the manuscript.

**Answer**

*We checked the manuscript for consistency and adjusted where needed.*

Appendix:  
  
**#37**

1) df\_w needs to be defined.

**Answer**

*Adjusted. Added that this is equal to conditions minus 1.*

**#38**

2) The top part of a fraction is called "numerator", not "nominator".

**Answer**

*Adjusted*

**#39**

3) Isn't the first term in the numerator the ostracism effect \*in the non-moderated/control condition\* (and the second term is the effect in the moderated condition)?

**Answer**

*We calculated it in the order we describe. It can also be done the other way around, which would lead to a change in interpretation but equal results.*

**#40**

4) In what sense does Delta-d "correspond" to partial eta-squared of the interaction? Numerically it cannot be the same (partial eta-squared must be between 0 and 1, while Delta-d as defined is not a proportion and may be larger than 1 and can be negative).

**Answer**

*When the resulting d is transformed into a squared correlation coefficient it gives the exact same value. This is highlighted in the Appendix and now reads*

When transformed to a squared correlation coefficient, this Δ*d* corresponds to the partial eta-squared of the interaction.

**#41**

5) Please add ^2 to s\_g and s\_d to make it clearer that these are variances.

**Answer**

*Done.*

**#42**

Final comment: In the spirit of open science, I appreciate the use of OSF and the authors' transparency in conducting this meta-analysis.

**Answer**

*Thank you. We also like to thank the reviewer for the thorough review and thus for making this a better manuscript.*

**Reviewer #2**:

**#1**

Overall this study looks competently executed and acceptable for publication. My only real concern is that authors could have done more to explore and account for the variability in their data. The meta-analysis demonstrates that the variability was considerable, but beyond establishing that moderators exist, the researchers appear to be not overly concerned with the question what is causing this variation. That leaves me slightly unsatisfied at the end: all this effort to conduct a meta-analysis, and the main thing we learn is that (a) the effect of rejection is strong (something we knew because it has been shown time and again), (b) the first sharp shock diminishes over time (new to me, but then I’m not an expert), and (c) the intensity of that shock depends… If authors were willing to stick their finger out a bit more and clarify just what this depends on, I’m sure I would find the study more valuable than it is now. I don’t care if their hypotheses were deposited beforehand: exploring is a scientists’ duty, as much as hypothesizing in advance (e.g., Tukey). But to be clear: this is just meant an encouragement; it’s very much up to to the authors to decide what course of action to pursue.

**Answer**

*We thank the reviewer for his/her kind words and regarding the manuscript as competent and acceptable for publication. We agree with Reviewer #2 that exploring the data is a valuable avenue for any study, including this meta-analysis. As a matter of fact, we were also puzzled by the heterogeneity in the data and we therefore conducted several exploratory analyses to understand this heterogeneity. The most important exploratory analysis that we conducted was the one in which we selected the most homogenous subset possible (i.e., only immediate fundamental need measures, 30 throws, 3 players), but still found high heterogeneity. Meta-regressions also failed to indicate any explanation for the heterogeneity. We agree that further exploration is definitely interesting, but also believe that we exhausted all possibilities that were available to us in the current dataset.*

Some other points that would help authors improve the paper up to a level that would match my expectations for PLOS One standard mainly concern the quality of the writing and the care about the argument being made. The introduction reveals that authors could have spent some more care writing (and perhaps thinking about) their subject. Suffice to say that it’s important to be precise. Some examples:  
  
**#2**

“Cyberball participants simply do not obtain a ball and thus need to infer that they are excluded” I think authors are trying to say something about implicit and explicit exclusion here. I also think they are trying to say something about acting together versus communicating with each other. But it’s not being said.

**Answer**

*This sentence was deleted, because the preceding sentence already contains the information.*

**#3**

The sentence “This focus on ostracism makes it an unique paradigm...” is clearly erroneous, because it is not the focus on ostracism that makes cyberball unique.

**Answer**

*The first paragraph in the Historical background section is changed into:*

Cyberball was introduced in 2000 as a means to study ostracism, that is: being excluded and ignored [1]. This focus of Cyberball on ostracism sets it apart from other paradigms that are tailored to study rejection, such as the future life rejection [2], the get-acquainted paradigm [3], and the autobiographical memory manipulation (i.e., remember a time when you were excluded [4]). The difference is that participants in Cyberball are not explicitly informed that they are excluded whereas in the other paradigms participants are provided a reason pertaining to why they are excluded.

**#4**

Further on, a sentence such as “research suggests that most people are ignored and excluded at least once a day” sits happily side by side with the sentence “research on school shootings has suggested a direct link between ostracism and revenge”. This could be spelled out more clearly. If everyone is a victim of exclusion, then obviously those who go on a shooting spree are, too. So is the point that ostracism is a frequently occurring post-hoc justification for this kind of behavior?

**Answer**

*We adjusted the sentence. It now reads:*

The social relevance is further evident in that ostracism not only affects the person who is ostracized (intrapersonal effects), but often also others (interpersonal effects). As a grim example, research on school shootings has suggested a direct link between ostracism and revenge. People who were ostracized may retaliate by murdering those responsible and sometimes even innocent bystanders [5].

***#5***

Further on authors write “This initial response is theorized to be socially painful, threatening [9] and easily detectable due to evolutionary over-sensitivity to cues of ostracism [12].” In a sentence such as this, please carefully distinguish phenomenon and hypothesis. There is abundant evidence for the first inference, but the evolutionary origins of this phenomenon can only be inferred indirectly from its existence and prevalence.

**Answer**

*We adjusted the sentence. It now reads:*

This initial response is theorized to be socially painful, threatening [9] and, following overdetection theory [12], should be easily detectable due to evolutionary over-sensitivity to cues of ostracism.

**#6**

It is stated that all selections and hypotheses were preregistered on OSF. But what is not spelled out is whether authors tried to learn something new from their data by exploring it?

**Answer**

*We explored several avenues. For example see reviewer #2, answer #1, but also answer below (reviewer #2, answer #7)*

**#7**

“Examples of interpersonal measures are donations to charity, helping behavior, money allocations in economic games, and aggression measures such as irritating sounds blasts or hot sauce allocation.” Please split the effects of positive and negative behaviors—they are qualitatively too distinct to be lumped together in this way. Later on I noted that K=10 for these studies (?). If small K was the reason for lumping things together please explain the criteria and total K in this section to help readers understand your decision making process.

**Answer**

*These were indeed split into positive (pro-social) and negative (anti-social) behaviors initially and were indeed lumped together due to small K, hence, low power for detecting moderation effects. For the first measure, there were 14 interpersonal measures, of which 4 are positive and 10 negative. For the last measure, there were 14 interpersonal measures, of which 8 are positive and 6 negative. We added a sentence in the manuscript to clarify this. One page ?? of revised ms we now say:*

These were initially coded into pro- and anti-social, but were collated into the category interpersonal due to small *k* the first measure (4 and 10, respectively) and last measure (8 and 6, respectively).

**#8**

For various decisions to include or exclude studies or factors, please provide an indication of the number of studies affected by your decision. E.g., “continuous variables that were dichotomized into factorial levels were also collapsed due to the many problems dichotomization can cause”. How many studies were collapsed in this way? I’m trying to assess the impact of your coding decisions.

**Answer**

*This collapsing occurred a total of four times, for the studies from (i) Stock 2011, (ii) two studies from Boyes 2009, and (iii) Zadro 2006. We added this number in the manuscript.*

Some other minor points:

**#9**

“we used the metafor package”: include version.

**Answer**

*Version 1.9-5. Added in the manuscript.*

**#10**

I do not understand this sentence: “Model indicates that the first measure was indeed reflexive and the last measure reflective.”

**Answer**

*The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2*

“Subset All” contains all measures. Subset Fundamental contains only fundamental need measures. Subset Intrapersonal contains all intrapersonal measures. Subset Interpersonal contains all interpersonal measures. Subset Model contains those where first measures is immediate and last measure is delayed. See Supplement S3.

*On page 28 this was clarified under the heading Measures, where the subsets are named.*

**#11**

“meta-analyses” is plural

**Answer**

*Adjusted*

**#12**

“by a large extend”  
= to a large extent

**Answer**

*Adjusted*

**Reviewer #3**: This study is a system review and meta-analysis of cyberball studies for effect size of ostracism. The manuscript is well-written and provides many detailed information for readers. The statistical analysis is rigorous and well-thought. The primary and secondary hypotheses are clearly stated. The results and discussion are also clearly presented. I have following comments.

*We thank the reviewer for his kind words and stating that our analyses are rigorous and the manuscript is well-written.*

**#1**

1. First, I appreciate the authors’ efforts in providing detailed information about the data and implementation, which greatly improve the transparency and reproducibility of the research. More importantly, the information is very helpful for readers to have an objective view of this study.

**Answer**

*Thank you for your kind words.*

**#2**

2. I would suggest moving the “code procedure” sub-section in Method section to supplementary. Although the code procedure is very important and helpful for some readers, it is too technical for most of readers.

**Answer**

*Although we understand the concerns for the technicalities, the supplement is meant for additional information only, while we consider the coding a crucial aspect of our method. We had thorough discussions on whether it was possible to have directional coding in spite of the bidirectionality of the expected effects and we think a reader will want to know how we were able to make directional claims despite this variety of measures and predictions. Hence, we think it is vital to retain this in the main manuscript.*

**#3**

3. I suggest adding a figure for study inclusion criteria. Many system review and meta-analysis paper in PLoS ONE use a figure to demonstrate the procedure for selecting studies.

**Answer**

*The manuscript contains the PRISMA flowchart in the supplemental materials that addresses this point. We added the flowchart in the manuscript.*

**#4**

4. It’s better to present the information in Table 2 as a forest plot, while putting the table 2 in supplementary. A forest plot summarizes the information and gives readers a intuitive understanding.

**Answer**

*We agree that a forest plot gives an intuitive overview of the effects. However, we think that the forest plot across 120 effects will be too sizable. More importantly,* t*he American Psychological Association prescribes that meta-analyses are to report the data on which main analyses are performed in a table. We therefore think it is more informative to retain the current format.*