Author response to reviews of

Manuscript AMPPS-22-0119

Reproducibility of published meta-analyses on clinical psychological interventions

Ruben Lopez-Nicolas et al.
submitted to *Advances in Methods and Practices in Psychological Science*

RC: ***Reviewer Comment*** AR: Author Response ☐ Manuscript text

Dear Dr. Flake,

We are very grateful to you for offering us the opportunity of sending a revised version of our paper. We are also very grateful to all the Reviewers, as they have made interesting suggestions and good recommendations to improve the quality of our paper. In the revised version of the paper we have tried to address all of your suggestions. Next, we provide point by point answers to the Reviewers' comments and specify how we have addressed them in the revised manuscript.

1. Reviewer #1

RC: The authors present their efforts to reproduce 217 published meta-analyses from Clinical Psychology. This is done systematically and thoughtfully.

Major points: "Each meta-analysis was labelled using the following reproducibility success scheme: (a) reproducible; (b) not process-reproducible; (c) numerical error, and (d) decision error." This initially read to me like there were four classes that each meta-analysis can fall into. However, it's actually a hierarchy, where c and d are necessarily also labeled a.

AR: In this scheme '(a) reproducible' involves both, process-reproducible meta-analyses without error. We admit that it can be a bit confusing. Therefore, in the revised version we have restructured this scheme as follows (see p. 10):

> Each meta-analysis was labelled using the following two-level reproducibility success scheme. First, each meta-analysis was labelled as: (a) process-reproducible; and (b) not process-reproducible. In our study, not process-reproducible refers to situations where we were unable to access the primary data neither through direct extraction nor upon request. Second, those labelled as process-reproducible were labelled as: (a) reproducible; (b) numerical error; and (c) decision error...

RC: It appears that, of the 217 meta-analyses, 146 fell into class a and 71 fell into class b. Of the 146 in class a, 52 were labelled as a numerical/decision error. 25 of these 52 were relabeled after double-checking for things like rounding errors, but the scatterplots focus on the full 52. The degree of discrepancy between the original and reproduced values is larger for the I-squared statistic than it is for the summary statistic or its associated CIs.

Overall, I found this paper interesting. I think it's important topic, and the biggest takeaway is a call for meta-analysts to make their datasets and code available in a systematic way. As someone who has tried to reproduce published meta-analyses on many occasions, I could not agree more! I did find the reporting of the results to be a little confusing, however. I'm not sure how much of this can be avoided, since listing statistics is inherently hard to read, but there are several cases where the authors simply refer to a set of meta-analyses without enough context (e.g., line 39, p. 16). One thing that might help is the use of a Sankey plot instead of Figure 3. Perhaps also designating each set of meta-analyses with some title and then referring to them in a consistent way. Or even having Figure 6 appear earlier in the paper?

AR: This is a quite relevant point, thanks for the feedback and suggestions on it.

We were exploring different ways to improve the reporting of the results. First, we tried to merge Figure 3 and Figure 6 into a single Sankey plot to illustrate the whole workflow. We found that plot a bit confusing and cluttered though, mainly due to the number of different primary data sources (Figure 3c). Second, we tried Figure 3 as a barplot and plotting the results of the Figure 6 using a Sankey plot instead of a bar plot. We deemed this an improvement, and therefore are using this plot in the revised manuscript. Furthermore, we agree with you that the (old) Figure 6 is much more helpful if presented earlier in the manuscript, so we have moved this figure to the "Outcome reproducibility" section. Now it's Figure 4 (see p. 16 in the revised manuscript).

RC: The authors bring up an excellent point about the importance of when these meta-analyses were published and how that should affect data availability. Some degree of analysis of this would be nice

to see in the paper (e.g., average years since publication for the 146 reproducible vs. the 71 not). I understand if the authors are wary about adding additional analyses that try to explain when and why a meta-analysis had errors, but I don't think an exploration of publication date falls into that category.

AR: We mentioned this as limitation instead of carrying out temporal comparisons because –as it wasn't among our primary aims– we didn't consider our sampling design suitable for it. As an unrestricted random sample of meta-analyses published between 2000 and 2020, as expected, the publication year distribution is clearly left-skewed, so the information provided by the data at the bottom range of publication year is limited.

We agree that more information on this matter in the paper would be useful. Hence, we have extended Figure 2 including a second panel showing the publication year distribution (see p. 13).

Furthermore, we statistically compared the possibility of retrieving the data as a function of publication year. These exploratory analyses are reported in the Supplementary file, section "Data availability over years". This analysis has been added to the Method section (see p. 12):

> Finally, the association between publication year and the possibility of retrieving the data in one of the ways conducted in this project were explored by fitting binary logistic regression models with publication year as predictor and process-reproducibility as dependent variable. We quantified the strength of the association by calculating odds ratios and 95% confidence intervals based on the profile likelihood. These exploratory analyses were not pre-registered. Details and results are reported in Supplementary file.

See the Supplementary file, section "Data availability over years" for full details of the exploratory analyses.

We have also briefly mentioned the results of the exploratory analysis in the Discussion section and have added some temporal expectation when discussing data sharing (see p. 22):

> A more positive sign comes from the positive association between publication year and the possibility of retrieving the data that was observed in an exploratory analysis in the supplementary materials. This tentatively suggests that data-availability is improving over time. This observation could be related to the existence of well-established meta-analysis reporting guidelines. For instance, the first PRISMA guideline (Moher et al., 2019) encouraged meta-analyst to report results of primary studies (e.g. primary effect sizes and their confidence interval through a forest plot, as was a common scenario among the cases included in this project), and the latest PRISMA guideline (Page et al., 2021) which puts more emphasis on appropriate data sharing through data files ready for reuse. At the same time in only 5% of the cases where data were retrieved in our sample were we able to retrieve the data in a machine-readable data file that was ready for reuse (e.g., *csv*, *xlsx*). Most often the data had to be retrieved from files in document format (e.g., *docx*, *pdf*). This forces people who want to reuse the data to manually recode the data, which is an inefficient and error-prone task. Even after partial double coding was carried out, this procedure did not avoid some coding errors, which were only detected by double checking meta-analyses with discrepancies. In our experience, the data retrieval process can be difficult when results are presented in general tables, as it involves matching subsets of these primary data with different meta-analytic results, while is not always clear which studies were used in which meta-analysis reported in a paper. Furthermore, because the tables in manuscript are often generated manually in document file formats (e.g. Word), we observed examples where this introduced another source of error. The foregoing discussion raises a key point about how time consuming the appraisal of meta-analytic reproducibility currently is, and how efficiency would be improved by having open access to the underlying meta-analytic data in data file formats ready for reuse. The latest PRISMA guidelines, along with some initiatives that promote appropriate data sharing (e.g., Wilkinson et al., 2016) have the potential to generate significant improvements in the re-use of meta-analytic data in the years ahead. In this regard, our results provide a useful baseline for future assessments.

RC: It strikes me that one of the main outcomes of this project is how much time and effort it must have taken to complete. Is there anyway this could be reported in some way?

AR: We couldn't agree more! Unfortunately, we don't have proper measures of this outcome as the hours invested in this project were not systematically tracked. Still, as we agree this is relevant information, we have mentioned it in the same paragraph as above (p. 22):

> The foregoing discussion raises a key point about how time consuming the appraisal of meta-analytic reproducibility currently is, and how efficiency would be improved by having open access to the underlying meta-analytic data in data file formats ready for reuse. The latest PRISMA guidelines, along with some initiatives that promote appropriate data sharing (e.g., Wilkinson et al., 2016) have the potential to generate significant improvements in the re-use of meta-analytic data in the years ahead. In this regard, our results provide a useful baseline for future assessments.

RC: Minor points:

- Why only clinical psychology? How might this impact things? Having spent a lot of time in clinical psychology research in grad school, my personal experience is that these researchers are slightly less likely to adopt advances in methods (e.g., reproducible code).

AR: We preferred focusing on a specific research area or subdiscipline because our prior of possible differences between different research areas. We didn't have priors about direction of these differences. We thought it

would be more informative to characterize one subdiscipline rather than to examine several research fields in one study.

We choose clinical psychology for the two main reasons:

- It is one of the areas that produces the most meta-analyses in psychology.
- Their direct impact on applied practice

We have also mentioned this as a limitation of the generalizability of the results in the revised manuscript (see p. 25):

> Finally, we only examined meta-analyses in clinical psychology as this is one of the areas that produces the most meta-analyses in psychology and these meta-analyses have a direct impact on applied practice, but it is unknown to which extent our conclusions generalize to meta-analyses in other sub-disciplines in psychology.

RC: -Paragraph 1 on p. 22 has several typos.

AR: Thanks for checking. Fixed.

RC: -In the scatterplots, I find the use of + slightly confusing, as it could indicate 2d confidence limits. I suggest x or circles.

AR: Thanks for the suggestion. We have changed it to `x`.

RC: -What's the correlation between error in the summary statistic and error in the heterogeneity statistic?

AR: The Pearson's correlation between those error was .172.

We have reported this in the revised manuscript (see p. 18):

> As shown in Figures 5 and 6, the discrepancies found in the heterogeneity statistic I^2 are larger than those found in the summary effects and their confidence intervals. The Pearson's correlation between the summary effect and I^2 discrepancies was .172.

RC: -Would it be possible to use something other than I2? It has bounding issues that might be affecting results.

AR: We used I2 because it's the most reported heterogeneity statistic. Hence it allowed us to carry out a large number of comparisons between original and reproduced results.

RC: -How is meta-analysis defined? Just when a model has been fit? How many of the meta-analyses are on the same exact data rather than from the same paper?

AR: Yes, here meta-analysis refers to standard pairwise meta-analytic models. For clarification we have rephrased the following (see p. 8):

From these 100 articles, each independent pairwise meta-analytic model of aggregate data fitted on at least 10 primary studies was selected

The meta-analyses selected from the same paper were independent, that is based on different data (e.g. different outcomes, timepoints...) based on the criteria explained in the manuscript. However, we don't have information on possible primary data overlapping between publications, because it's beyond our research question. It is highly likely that there is some degree of overlap, but we consider it a different (and a really interesting and relevant one) research question.

Thanks again for taking your time providing very useful feedback. We appreciate how your suggestions have improved the manuscript.

2. Reviewer #2

RC: **In this manuscript, the authors assess the analytical reproducibility of a set of meta-analyses. They report obstacles in obtaining the data in a workable format and identify numerical and decision errors in a sizable portion of the meta-analyses.**

I think this is an interesting paper that provides compelling arguments that we need to improve meta-analytic reporting. I also commend the authors for preregistering their research plan and sharing their data and materials. One aspect of the paper I particularly liked was the effort to identify the causes of any identified discrepancies. The authors describe the results of this qualitative procedure on page 16 (lines 3-14). Here, I do have some questions about the procedure and the subsequent conclusions.

- P. 16, line 6: you describe that some discrepancies could be explained by rounding issues. Could you explain this in a bit more detail? Does this mean that there was incorrect rounding in the meta-analyses? And did you count that as correct because it was only a small discrepancy? If so, why? Similarly for the minor reporting errors you mention in line 7: it sounds as if minor errors are not considered issues for the reproducibility. This is a choice that needs to be justified.

AR: We agree that this could have been more clearly specified. In order to improve our explanation, we think that the clearest way of providing this information is by reporting each of the 15 cases. In this regard, we have added a table in the supplementary file which provides an explanation about each of these 15 cases (see Supplementary file, Qualitative assessment section).

We have added a footnote on p.15 linking to the supplementary file:

> Furthermore, 15 were labelled as approximately reproduced or reproduced with minor adjustment in a qualitative check because the discrepancy was probably explained by rounding issues, inverted signs for results (when effect sizes were reported in absolute values) and primary data, minor reporting errors, or minor adjustments in the analytical scheme[a].
>
> [a] Full details in Supplementary File at `https://osf.io/fjhpw`

RC: **- After the qualitative check, you relabel 15 meta-analyses as reproduced, because you were able to spot the reason for the discrepancy, right? I'd argue that having a good idea why something didn't reproduce, doesn't suddenly make the result reproducible. I think labelling these cases as "reproduced with issues" or "approximately reproduced" or something along those lines makes more sense, than simply calling them "reproduced".**

AR: We agree that these meta-analyses deserve a specific category. In this vein we have made the following change in the manuscript (see p. 15):

> Furthermore, 15 were labelled as approximately reproduced or reproduced with minor adjustment in a qualitative check because the discrepancy was probably explained by rounding issues, inverted signs for results (when effect sizes were reported in absolute values) and primary data, minor reporting errors, or minor adjustments in the analytical scheme.

Furthermore, we have remade (old) Figure 6 (currently Figure 4) and now it's a Sankey plot displaying the whole flow of the reproduction attempts. In this new figure, these 15 cases are labelled also following this

recommendation (see p. 17).

RC: - Abstract: I think it would be insightful to also report percentages of meta-analyses that could be reproduced / had errors in the abstract, instead of only the absolute numbers.

AR: Agreed. We have added percentages of process-reproducible meta-analyses and of meta-analyses that were still labelled as error after checking for coding errors and qualitative assessments. We just highlighted those percentages in the abstract to avoid misinterpretations of what the other values represent.

RC: - Figure 1: I don't really understand the arrows going up from systematic procedure to research question/aims and eligibility criteria. As a matter of fact, I think the whole top part of the figure is somewhat hard to understand. Isn't the whole process pretty linear? So: rationale - research question/aim - eligibility criteria - systematic procedure? And what do the circles with a vertical line through them between the different systematic procedural steps mean? Shouldn't these just be arrows? Finally, and this might be a matter of personal preference, but I would switch the columns with the previous literature and the required elements: on the left you would then have the necessary ingredients for the procedure, and on the right the examples of previous literature.

AR: The arrows going up from the systematic procedure was a mistake, well spotted, thanks! About the top part of the figure, we wanted to point out to the functional relationship between RQ/Aims and eligibility criteria. We agree that could be a bit confusing.

Based on your suggestions we have modified Figure 1 (see p. 6). We have made the following modifications:

- Now the whole flow is linear.
- Now all the steps are connected using the same symbol (arrows).
- We have switched the side columns as you recommended.

RC: - P. 7: why did you select 100 meta-analyses? Was this (purely) for feasibility reasons or did you also do some sort of power analysis? I think it would be good to add a short justification for this sample size to your paper. Similarly, please (briefly) justify the choice to only include meta-analyses with at least 10 primary studies.

AR: We tried to keep a reasonable trade-off between informativeness and feasibility. This was based on our own judgment. We agree that this should be reported, so we have added the following text (see p. 8):

> This sample size was based on our judgement of an acceptable trade-off between informativeness and feasibility.

On the other hand, we just included meta-analysis with at least 10 primary studies trying to focus on the most relevant meta-analyses of each paper. We also have included a mention of it in the manuscript (see p. 8):

> This criterion was established to focus on the main meta-analyses of each paper, based on the assumption that the search strategies would be designed to maximise the number of primary studies included that were related to the main aims of the paper.

RC: - P. 8, line 50: what is the difference between primary statistics and primary effects and their standard errors?

AR: This differentiation was made to distinguish between the effects sizes already computed and the data used to compute them. We have specified this in the manuscript as follows (see p. 8):

> In order to be able to reproduce meta-analyses of aggregate data, primary-level effects sizes and their associated standard errors are required. These are generally computed from statistics retrieved from the primary studies such as means, standard deviations or sample sizes. We attempted to retrieve the least processed data possible. First, we sought for the statistics used to compute primary effect sizes (e.g., means, sd); second, we sought for the primary effects sizes already computed and their standard errors...

RC: - P. 8, final sentence: shouldn't this be "confidence limits" instead of confidential limits?

AR: Thanks, fixed.

RC: - P. 10: what was the outcome of the reliability checks? How much consensus was there initially?

AR: We have added a section in the Supplementary file to report the inter-rater reliability. Consensus was high.

RC: - Table 1: I'm not sure how informative the percentages are next to the N, mainly because you often just refer to a single case. If you choose to keep the percentages in the table, please double check the numbers: by my calculations, there are 8 authors who sent an email with a specific reason (adding up your Ns = 8), but 1 of 8 is 12.5%, rounding to 13%, not 12%.

AR: Well spotted, thanks! This table was generated by knitr::kable(), using a data frame whose percentage values were computed by base::round(). It appears that base::round() follows the IEC 60559 standard and it uses round-to-even method. We think you pointed this out because the sum of the percentages doesn't add up to 100. We agree that this could be weird and that the percentages, given the N, aren't very informative. Hence, we have removed this column from the table (see p. 14).

RC: - Figure 4A: the labelling "Decision Error: Error / No Error" is confusing. If a data point is classified as "No Error", is it not an error at all? Or only not a decision error (but still potentially a numerical error)?

AR: This label only refers to decision errors. We agree that this labelling can be confusing. We have modified it to "Decision Error: Yes/No" (see p. 18).

RC: - Figure 4A: you could consider making the crosses smaller to increase readability; because of the high overlap of the crosses, it can be hard to spot the "diamonds" of the errors.

AR: Thanks for the suggestions. We have reduced the size of the crosses and colored the "Decision error" label to increase readability (see p. 18).

RC: - Figure 4B: classifying upper and lower bounds of a confidence interval as Decision Errors (y/n) is confusing. How can a bound of a confidence interval be a decision error? Please explain this. Furthermore, can it occur that only one of the two bounds of a single CI is a (decision) error? Or are they always classified as a pair? In either case, it is hard to read this plot; in the first case, you can't see that only one of the bounds was an error (because you don't know which other value belongs to the pair), and in the second case it gives an overly negative view of the number of errors in the confidence

intervals (one error in a confidence interval would then result in two diamonds in the plot).

AR: They are always classified as a pair. Each pair comes from the same case, and we are classifying cases. You're right that it could be misleading due to decision errors are displayed twice per case. We have opted for not classifying the confidence limits as decision errors, and to do so only for summary effects (which also come from the same cases, so the total is the same) (see p. 18).

RC: - Figure 6, panel 1: I would simply call the last category "decision error", not "numerical and decision error", this is already implied and by labelling it so explicitly it almost sounds as if there also exist decision errors that are not also numerical errors

AR: We have made changes in this Figure following suggestions from another reviewer. Now, it is Figure 4 (see p. 16). As you can see, now this category is called only "Decision error" following your recommendation.

RC: - Page 22, line 10; typo: "which is both inefficient and *an* error-prone task" - Page 22, line 14: typo: "out" should be "our" - Page 23, line 17: typo: "issued" should be "issues"

AR: Thanks for checking! Fixed.

RC: I'm looking forward to seeing this manuscript in print.

Signed, Michèle Nuijten

AR: Thanks again for your useful feedback, we really appreciate your help in improving the manuscript!

3. Reviewer #3

RC: **- I think the reporting of results in the abstract could be imrpoved. Currently it states that 52 reanlysis results showed a discrepancy larger than 5% from the original results - I think that sounds far more concerning that it actually is given that there were very few 'decision errors' and the authors believe that the reproducibiltiy issues had "little or no effect" on the original conclusions. So I think its worth stating that in the abstract to avoid any miscommunications. The point at which the authors 'qualitative assessment' and the input from the original authors is also not clear in the abstract.**

AR: We agree that this is something that deserves to be mentioned in the abstract. We have added two sentences to the abstract:

> Second, through a multi-stage workflow, we tried to reproduce the main results of each meta-analysis using these data. The original data were retrieved for 146 (67%, 146/217) meta-analyses. Of these, 52 showed a discrepancy larger than 5% in the main results in the first stage. For 10 meta-analyses this discrepancy was solved after fixing a coding error and for 15 of them it was considered approximately reproduced in a qualitative assessment. In the remaining 27 meta-analyses (18%, 27/146), different issues were identified in an in-depth review of the papers, such as reporting inconsistencies, lack of data, or transcription errors. Nevertheless, the numerical discrepancies were mostly minor, with little or no impact on the conclusions. Current practices of data sharing in meta-analyses hamper the reusability of meta-analytic data. The implementation of new tools would help to avoid certain errors in the meta-analysis reporting process.

RC: **- The use of "analytic reproduciblity", "process reproducibility", and "outcome reproducibility" seems inconsistent, overlapping, and overall pretty confusing. The title and abstract refer to "analytical reproducibility", but the introduction focuses on "process reproducibility" and "outcome reproducibility" (analytic reproducibility is not mentioned). Process reproducibility seems to be equated with data availability (or lack of), but I think it was originally intended to refer to all aspects of the analytic pipeline (original data + original code + original software environment, etc). This results in setences like "Based on the availability of primary data, either retrieved directly from the paper or upon request, 146 meta-analyses (67%, see Fig. 3a) were labelled as process reproducible" This seems odd to me - because some of the cases labelled as "process reproducible" because the data were available, were in fact not reproducible when the authors tried to repeat the analyses. To be fair, I don't think the original sources of these terms defined them particularly helpfully, but I think the onus is on the authors to be clear about what they mean, and how they relate to each other in the present manuscript. For what its worth, I'd recommend just using "analytic reproducibility" to refer to the entire process of trying to repeat the original analyses with the original data to see if you get the same outcome (this seems an accurate description of what the authors did). If this process fails, identify the reason (if possible) which could be an early stage failure (e.g., lack of access to data or code) or a later stage failure (e.g., typo in original manuscript). Adding the terms "process reproducibility" and "outcome reproducibility" into the mix seems unnecessary and confusing to me.**

AR: We see your point and agree that consistency in this matter should be improved. Indeed, as you rightly pointed out, in our study process-reproducibility is equated with data availability. As we see it, this is due to our design and the stage of the meta-analysis pipeline we focused on, not to concept itself. So, in our study it's just a nomenclature matter.

However, we believe this distinction between process/outcome reproducibility is useful for the organization

of the previous (and future) literature on meta-analysis reproducibility (as we did in Figure 1). For instance, this distinction allows us to classify studies that review the availability of required elements to reproduce a search strategy of a meta-analysis (e.g., Koffel & Rethlefsen, 2016) in a different category than those that verify the reproduction of the electronic search result (databases output) using those elements.

Regarding the organization of our own results, there would be no difference other than the label used. Therefore, we think it is a good idea to keep those labels for consistency with the terms and literature mentioned in the introduction.

We completely agree that the general consistency on this matter should be improved. To this end, we have clarified the synonymy between process-reproducibility and data-availability in our manuscript, and have made the following changes in the manuscript:

We have changed "analytical reproducibility" to "reproducibility" throughout the manuscript (also in the title and abstract) to avoid confusion and we have highlighted in the introduction (see p. 5) that:

> Figure 1 displays a summary of the basic meta-analysis pipeline through a flowchart, outlining the different stages and listing previous work that has explored different facets of reproducibility of these, as well as a summary of the required elements to be able to reproduce each stage. In this project we focus on the last stage, related to the statistical analysis and quantitative results of the synthesis.

We have also restructured the reproducibility success scheme (see p. 11):

> Each meta-analysis was labelled using the following two-level[a] reproducibility success scheme. First, each meta-analysis was labelled as: (a) process-reproducible; and (b) not process-reproducible. In our study, not process-reproducible refers to situations where we were unable to access the primary data neither through direct extraction nor upon request[b]. Second, those labelled as process-reproducible were labelled as: (a) reproducible; (b) numerical error; and (c) decision error.
>
> [a]This hierarchy is a minor deviation from the pre-registered protocol. It is essentially the same and the results are identical. It was introduced to improve clarity.
>
> [b]Process reproducibility, as described above, could imply a different situation if more conditions need to be met to proceed with the reproduction attempt. In our study, this is equivalent to data availability due to our design and the stage of the meta-analysis pipeline we focused on.

Furthermore, in both figures (Figure 3 and new Figure 4) that display results and mention process reproducibility, data availability was also mentioned in brackets to avoid confusion.

RC: - In the abstract: "From a random sample of 100 papers containing at least one meta-analysis on the effectiveness of interventions in psychology" - should say "clinical psychology" I think?

AR: Indeed. Thanks, fixed.

RC: - Throughout the manuscript there is often reference to stages (e.g., "following the first stage of re-analysis") but I don't recall the different stages being clearly stated anywhere. Perhaps a flow digram would be helpful? I see Figure 6, but this appears to just be about the outcome reproducibility part of the study, and comes too late in the manuscript. This figure also doesn't include the authors "qualitative assessment" - its a bit unclear where that fits into the overall timeline.

AR: We have remade and moved (old) Figure 6 to a new Figure 4 in response to comments by the other reviewers.

The current Figure 4 is a Sankey plot displaying the flow of the reproducibility attempts and it's presented earlier in the manuscript (see p. 16). We hope that this figure at this place will be more informative about the stages we discuss.

RC: - "The availability of necessary primary data was checked" - it would be helpful to precisely define "necessary" and "primary" here, especially in the context of meta-analysis. Its not clear to me whether primary means all the individual-unit data from the studies included in the meta-analyses or the summary effect sizes. And its not clear which data are considered necessary or unnecessary.

AR: We have modified the structure of this paragraph to begin by specifying the necessary data and at what level (see p. 8):

> **3.1. Retrieval of primary data**
> In order to be able to reproduce meta-analyses of aggregate data, primary-level[a] effects sizes and their associated standard errors are required. These are generally computed from statistics retrieved from the primary studies such as means, standard deviations or sample sizes. We attempted to retrieve the least processed data possible...
>
> [a] By primary-level data we mean aggregate data from included primary studies.

RC: - "confidential limits" - typo - "confidence"

AR: Thanks, fixed.

RC: - Including "data reusability" in the title and abstract is superfluous given that reusing the data is a component of analytic reproducibiltiy and the authors did not try to reuse the data in any other way (the way its written implied to me that there would be some additional assessment of reusability).

AR: We agree that including "data reusability" could be redundant and, above all, misinterpreted as though some kind of additional reusability analysis had been done. Hence, we have removed it from title and abstract.

RC: - Some parts of the methods section contain results (see e.g., bottom of page 8) and some parts of the results section contain commentary or opinion I would expect to be in the discussion section (e.g., "a remarkably low reponse rate").

AR: We have checked for this, and we have removed misplaced statements. Specifically, we have made the following changes:

- Remove results on data types retrieved in each meta-analysis from Method section (p. 8)
- Remove the rate of meta-analysis without analysis script code available from Method section (p. 9)
- Remove "remarkably" from Results section (p. 15)

\RC{- Throughtout, it would be helpful if the numerator and denominator is provided as well as percentages. The manuscript could be clearer about where the denominator comes from too. For example, "In 141 cases (97%)" the denominator is presumably 145, but I'm not sure how we got there from the original 217 meta-analyses and I shouldn't have to do that calculation in my head. Another example is this sentence: "Only in 3% of the cases the primary data was retrieved upon request" - is that 3% of all cases or 3% of the cases in which data was requested?}

The two examples you mentioned are no longer in the current version. The first one was removed in the revision of results reported in method section (see previous comment). The second one was removed because it was indeed confusing. This 3% is of all cases. The rate of data retrieved from request out of the total of cases in which data was retrieved is not very informative of anything.

Furthermore, we have reviewed the text and added the numerator and denominator in cases where percentages are presented without enough context:

p. 14:

> Although attempts were made to retrieve data for 78 meta-analyses from 25 different papers by emailing the corresponding authors, data was only retrieved for 7 meta-analyses, from 3 different papers (12%, 3/25, see Fig. 3c). For the remaining 71 from 22 different papers, a reply providing some reasons not to share was received in 32% (8/25, see Fig. 3c), whereas no reply was received for the remainder of the meta-analyses.

p.15:

> Data were retrieved from tables or forest plots in *pdf* or *docx* format—either in the document itself or in the supplementary materials—in 92% (134/146) of the cases.

p. 18:

> These examples of inconsistencies in original results or data were found in 4 cases (3%, 4/146). These appeared to be typos.

RC: - Authors were contacted when data were not available, but it seems they were not asked for analysis scripts that were not available, was there a reason for this?

AR: Good point. There were some reasons for this:

- Not necessarily all authors of included meta-analyses will actually have an analysis script to share, because many used point and click software.
- We expected script availability to be very low, and requesting it would have meant sending request for virtually every paper included in our re-analysis
- We designed an approach to reconstruct the original analysis scheme .

We have mentioned those reasons in the revised manuscript (see p. 10):

> We designed this procedure to reconstruct the original analytical scheme when the original analysis script was not available instead of trying to request it from the original authors due to: (a) not necessarily all authors of included meta-analyses will actually have an analysis script to share, because many might have used point and click software, and (b) we expected analysis script availability to be very low, and requesting it would have meant sending request for virtually every paper included in our re-analysis.

RC: - "Primary-level data and aggregate-level data described above were coded by five members of the group." - this is a bit unclear, it could that the data for each article were independely extracted five times (I dont think that's the intention).

AR: Thanks for spotting it. We have rephrased this as follows:

> Data collection procedure was carried out by five of the authors...

RC: - It would be interesting to get some sense of how important/influential the assessed meta-analyses are in order to gauge the impact of any reproducibility issues - perhaps include some descriptive data about citation counts?

AR: Great suggestion, thanks! We have added information on citation count of the papers included. See p. 12 and (new) Figure 2(c):

> ...were cited 108.39 times on average (sd = 151.00; median = 57; interquartile range = 29-128; range = 3-1036)...

RC: - There are several sentences where I think its important to report the actual numbers rather than the authors verbal interpretation of the numbers e.g., "Data available in tables or forest plots shared in pdf or docx format –either in the document itself or in the supplementary materials– was found to be highly prevalent" and "This example of inconsistencies in original results was also found in other cases".

AR: We agree. We have added quantitative information on these cases:

p.15:

> Data were retrieved from tables or forest plots in *pdf* or *docx* format—either in the document itself or in the supplementary materials—in 92% (134/146) of the cases.

p. 18:

> These examples of inconsistencies in original results or data were found in 4 cases (3%, 4/146). These appeared to be typos

RC: - Figure 4 - the caption is cut off. Its difficult to make out error vs no error because of overlapping points, perhaps use colour here?

AR: Thank you for spotting it. With regards to the scatterplot, we have reduced the size of the crosses and added color to the Decision error label (see p. 18 and 19).

RC: - "In one of them, the original authors sent back a link to an OSF repository where the original data and analysis script were stored. According to the authors, this link was not reported in the paper by mistake. The script was run on these data and the results were successfully reproduced." - Do the dates in the OSF archive indicate that the files were uploaded before the paper was published?

AR: Interesting point that we missed checking. It seems that the OSF repository was created just before the paper was accepted (02/06/2019 created, 13/06/2019 accepted). It seems that the OSF repo was created upon request of the journal. We think that this deserves to be mentioned in the manuscript, so we have added the following text (see p.21, footnote 7):

> According to the repository timeline the project was created on 02/06/2019 and according to the journal's article history the paper was published on 13/06/2019. It seems that the repository was created as a journal requirement.

RC: - The authors responded "No" and "Not Applicable" to these prompts: "Does your paper rely on new or previously unpublished empirical data from your lab?", "If your paper presents new empirical work, does it include the following statement: "We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study"? It seems to me that both of these apply to the current manuscript (the nove empirical data is the results of the reproducibility assessments).

AR: We have modified in the resubmission form the answers to these prompts. Therefore, we have also included the "Reporting" subsection in the "Disclosures" section (see p. 7):

> **3.2. Reporting**
> Below we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Accordingly, we have included information on how we determine our sample size. We have included the following statements (see p. 8):

> Of this pool, 100 were randomly selected using a random number generator between 1 and the total number of meta-analyses identified. . . This sample size was based on our judgement of an acceptable trade-off between informativeness and feasibility. From these 100 articles, each independent pairwise meta-analytic model of aggregate data fitted on at least 10 primary studies was selected. In case no meta-analysis reported in a paper had at least 10 studies, the meta-analysis with the highest number of primary studies was selected, which was the case for 29 of the articles included in this report. This criterion was established to focus on the main meta-analyses of each paper, based on the assumption that the search strategies would be designed to maximise the number of primary studies included that were related to the main aims of the paper.

RC: - I may have missed it, but it would be good to see data on when the target articles were published - a histogram in a supplement would do the job.

AR: Agree. We have added two panels to Figure 2 and one of them is a histogram of the publication year (see p. 13).

Thanks again for taking your time providing very useful feedback. We're sure your suggestions have improved the manuscript.