Response to Reviewers
(Gold in, gold out. Quality appraisal and risk of bias tools to assess non-intervention studies for systematic reviews in the behavioural sciences: A scoping review)

Recommender

The intended research would be important as it addresses Risk of Bias assessment in evidence synthesis for behavioral sciences. Without the RoB, the evidence synthesis results (and thus conclusions) have an increased likelihood of being misinformative.

As I have obtained only one review for this submission, I have carefully read it myself and provided several comments as annotations in the text (see the PDF attached).

I find the report well-written, and the methodology thoroughly developed.

Thank you for the positive feedback and for taking the time to review the manuscript in lieu of another reviewer. Your comments were very helpful!

I provide comments/suggestions as yellow highlights (with a comment) in the PDF of the protocol.

Thank you for this - we have addressed each of your PDF comments as tracked changes in the new version of the manuscript. There were some exceptions:

- We weren't sure if you were suggesting we place the reference set of articles that we used to develop the search strategy (available on the OSF) in the main body of the manuscript or in its own appendix in the report. We didn't feel that it was useful to list all nine articles in the main body of the manuscript itself and so we've created an appendix within the RR itself and listed the articles there instead, in addition to keeping the version already available on the OSF repository (https://osf.io/r4kp7)
- We've decided against adding a definition of quality and bias to the abstract for two reasons: 1) it would make the abstract significantly longer, and so instead we've incorporated clearer definitions into the introduction (as you suggest in your next point) so that we and the reader are on the same page faster, 2) the abstract is likely to change following Stage 2 of the RR once we have completed the review and so we can take the opportunity to revisit the content of the abstract then in a way that more accurately reflects the content of the overall paper.
- For the 'concept' part of our inclusion and exclusion criteria section, you suggested including tools assessing reporting and openness. Our main interests in this scoping review are quality assessment and risk of bias tools rather than tools that explicitly assess reporting quality or transparency/openness of individual studies. Therefore, including these tools inside our main concept of the review would be outside of the core scope as these tools wouldn't help us answer our objective. However, we will still be looking for *items* related to reporting quality or transparency/openness within the tools we find as we believe it is a necessary prerequisite to quality assessment and we are

interested in seeing how many tools acknowledge and include this as a subset of their overall items. Therefore we decided not to add this additional criteria inside the concept part of the PCC framework.

We've broken down your remaining comments and numbered them for ease of reference.

1. In general, I lack a clearer distinction between risk of bias, methodological quality, and reporting quality (transparency). While biases are nicely defined (and described) later in the text, it should be clearly stated at the beginning of the introduction that there is a distinction between the risk of bias (which usually refers to internal biases of the study), the biases of the overall evidence base, and reviewers biases and that this work is primarily focused on the internal (within) study biases.

We have added a new paragraph (the third paragraph of the introduction, page 3) with a clearer definition of 'risk of bias' vs 'quality assessment' (and 'critical appraisal') as it pertains to this paper, which we hope addresses your first point. We discuss transparency/reporting quality further in point 3 of your feedback comments.

Additionally, just before the 'Many Different Biases Can Influence Systematic Reviews' section, we have expanded on that final paragraph to lay out the expectations for the rest of the introduction and make it clear that we are summarising three aspects (reviewer, literature, and individual studies) but focusing for the rest of the paper on the individual studies (page 4-5).

2. Further, it should be clear (from the start) what is meant by 'quality'. Even later in the text, I did not find a clear definition of what quality is, and sometimes, as written in the text, it seems to be equal to the risk of bias (see my marks in the text). I would not agree with this - bias is a systematic deviation from the truth (which has actually never been explicitly mentioned in the text), and quality (while it can be linked to bias) can also link to e.g. sample size (which might affect CIs of the estimates), use of the best possible methods to measure the outcome (while the other methods are not necessarily biased, but maybe less precise).

This ties in with the first of the changes we made in response to your previous point 2, where we've added a clearer definition of the distinction between risk of bias and quality in the third paragraph of the introduction (page 3).

3. Finally, there is the reporting quality (or transparency) of primary studies. At least in ecology, CATs will cover some of these (or all) aspects, so I feel it would be important to highlight this early on in the introduction.

We have clarified the importance of transparency (or reporting quality) in facilitating assessment in the last paragraph of the Study Bias section in the introduction, right before the Aims. We also took this opportunity to justify our additional interest in open science practices and highlighted the importance of reporting quality in critical appraisals. This also helps to address point 7 of the other reviewer's comments.

4. I think it would also help to specify 'methodological quality' whenever the word 'quality' is used in relation to methodology (otherwise it can be e.g. quality of reporting).

Thank you for the suggestion. For clarity, we added “methodological” to each mention of quality that refers to study designs. Where we discuss reporting quality we refer to it explicitly as “reporting quality”.

5. On page 8, it is stated that particular attention will be given to items that assess transparency. I am unsure if this is something that relates to the RoB (other than enabling assessment of the RoB). I would like to see a better justification on including these in the RoB or methodological quality assessment (which is the main aim of this mapping effort).

Thank you for pointing this out. It seems like this refers to the following sentence from the ‘Aims’ section:

“Particular attention will be given to items which assess a study’s openness and reproducibility, such as open data and open access practices, as well as integrity checks like funding sources and ethics approval.”

It is correct that this was not sufficiently addressed in the introduction prior to the ‘Aims’ section. Now the sub-heading ‘Study bias’ includes information in the final paragraph that explains this in more detail (page 11).

6. On the other hand, if these are interesting from other perspectives (e.g. understanding the transparency of a certain evidence base), then I am unsure what recommendations will be given for those (once your research is finished). Would you recommend that they are included in CATs? I feel that RoBs should be as simple as possible, as then they are more likely to be used. So if we do add any extra elements to asses, this will likely reflect on the usage of the tool.

Thank you for this suggestion. We will extract information about the number of items in each tool as well as information about additional guidance and instructions to assess how long and time-consuming these tools are on average. Once we have our results and think about future recommendations we will use this information to consider how adding additional items, that are currently uncommon in these tools, could impact their usability. This will be reflected in the discussion.

7. In the section describing the literature bias, I see that it is concentrated mostly on the publication bias. However (at least for my field of study, ecology), here we need to consider other biases in the evidence base (e.g. geographic, taxonomic) which obviously will also be linked with the inferences made from the meta-analysis. I do not know if this applies to behavioral studies too.

Thank you, we have addressed your suggestion by clarifying what we mean by literature bias in the following two sentences:

"However, it does not protect the review outcomes from being influenced by pre-existing literature bias i.e. when the body of available evidence does not accurately reflect all the research done on a particular topic. Literature bias includes selective reporting of positive findings (reporting bias) and selecting articles with positive findings to be published in a journal (publication bias; McGauran et al., 2010)." (page 6)

We agree that the assessment of limitations in the evidence base, such as geographical bias, is relevant to consider and should be assessed as a separate core element of a meta-analysis or a systematic review. However, in behavioural sciences, this type of study or literature limitation does not lead to reduced confidence in study quality or reliability of results as long as authors conducting literature reviews account for this limitation and do not overgeneralise their findings. We believe this to lie beyond the scope of this paper and more in line with the decision-making that systematic review authors should make as part of defining their research questions.

8. The protocol is detailed, especially when combined with the additional information available on the OSF. I am however unsure if some of this info would be useful to have in the Registered Report itself.

Thank you. We were unclear on your comment, however - are you suggesting that we move more of the supplementary information into the report itself, or that some of the information in the RR is unnecessary and should be relegated to the supplementary information? We have separately addressed a direct comment on the annotated pdf that explicitly mentions moving supplementary information back into the RR, which we hope also accounts for this suggestion too.

Reviewer 1

Thanks for the kind invitation to review "Gold in, gold out. Quality appraisal and risk of bias tools to assess non-intervention studies for systematic reviews in the behavioural sciences: A scoping review".

I think this study would address a very important issue. In systematic reviews of RCTs, assessing primary studies with tools such as Cochrane's Risk of Bias is very common. However, outside this context, including such assessments is much less frequent. This may be due to a lack of tools that may be applicable or a lack of knowledge of the authors conducting the reviews. Therefore, I believe that conducting a review to highlight what tools currently exist, what aspects these tools assess, and for what contexts they could be used would be very valuable for authors. I also believe that it will serve to identify gaps that are not currently assessed by existing tools but are aspects that should be examined.

Overall, I think the protocol for the scoping review is very well designed. I commend the authors for their efforts to be transparent and to increase the reproducibility of their work. I also liked very much the comprehensive checklist of items to be evaluated for each tool. There are some points that I think the authors could reconsider, although I do not think it is mandatory to make

these changes and I would also be happy if the authors argue why it would be better not to make them.

Thank you for the positive comments, your enthusiasm for the project, and your thoughtful and thorough review. We've addressed your points in turn below, which we have broken down and numbered for ease of reference.

1. Regarding the introduction, I believe that the outline on the three types of biases is very illustrative. It explains the potential sources of bias that a meta-analysis may have and mentions the tools or guidelines that can be followed to address them. However, I notice that the 'Study bias' section is as extensive as 'Literature bias' and 'Researcher bias.' I was wondering if it might be appropriate to expand the 'Study bias' section by introducing some key concepts when evaluating the quality of primary studies. For example, I think it could be helpful to already introduce the concepts of construct, external, internal, and statistical validity, as well as provide some examples.

Thank you for this suggestion. Indeed it was an important element that was missing from the manuscript. This section has been restructured and expanded, as you can see in the tracked changes. We mention the different types of validity. In addition, we now added a paragraph on the importance of study rationale and about finding errors in reports. In line with this, we also added a new item to be extracted (item 19) to check whether the tools assess for study rationale and research questions.

2. It could also be mentioned that some tools cover certain aspects but not all—for example, the Cochrane Risk of Bias scale focuses primarily on internal validity (Hartling et al., 2009).

Thank you for pointing this out. We have added two citations to the end of the second paragraph of the Study Bias section to reflect this.

3. Other considerations regarding the 'study bias' section. When mentioning concepts like 'selection bias,' 'interviewer bias,' or 'citation bias,' it might be helpful to add some references to studies that discuss these biases in more detail.

We've expanded this section out with more references to relevant literature.

4. However, I don't quite understand why citation bias is placed in the analysis stage. I think it might correspond more with the discussion stage (where only studies that confirm certain results are cited). Also, I also don't fully understand placing "analytical flexibility" there. I think it could be more appropriate to describe it as "inadequate use of statistical tests", which might occur when there is analytical flexibility.

You're correct that 'citation bias' has been placed within the wrong set of parentheses here. We've now moved it to be within the 'writing and publication process' parentheses.

In this case we use ‘analytical flexibility’ to mean the degrees of freedom that researchers have with regards to choice i.e. the garden of forking paths. We consider that separate from the inadequate use of statistical tests (though not unrelated), so we’ve added in use of statistical tests as an additional point instead.

5. Regarding data extraction, the authors mentioned that have piloted this process. If I understand correctly from the shared files, this coding has been done from one tool (ROBINS-I). The procedure that is indicated to be followed is “items from each tool will be extracted by one reviewer and items marked with an asterisk will be independently validated by a second reviewer, who will focus solely on these tagged items”. I wonder whether it might not be more appropriate to assume double coding of all items for at least a percentage of the tools (e.g. 25%, although this might depend on the total number of tools encountered) to ensure that all items are being well understood (the pilot has only been done with one tool). Of course, I understand that this would be more costly, and that it may be a waste of time to perform a double coding if the inter-rater reliability is very high, but perhaps it could avoid errors. I leave it to the authors to decide.

We understand these concerns, and agree with them. Our original stance was that double-checking certain items will ensure reliability for all papers more than coding 25% of the articles would. We expect heterogeneous data that we believe we cannot completely train for. Moreover, committing to 25% (or any other reasonable percentage) could prove problematic in case of a large number of tools to assess, while not guaranteeing reliability of the data that has been extracted only once. We updated the extraction items and instructions now following the reviews and piloted the process again to find discrepancies in the asterisk-marked items, which are more subjective. We therefore decided to change the plan for the extraction process. We will have two people extracting all of the items, which will help to reduce the subjectivity as they will agree on the extracted data by consensus.

6. Regarding the data to be extracted from each tool, I think this should be explained more in detail in the main text. In the current version of the manuscript, I found it difficult to see that this information could be found in the ‘Data Extraction Instructions’ file. I think the main manuscript should at least mention the domains for which information is going to be extracted (even if ‘ad-hoc’ items are added later). For example, you could explain that you will extract metadata such as title, format or whether they offer support, as well as content information related to aspects such as ‘Open & Reproducible Scholarship’ or ‘Validity’. I would also reference explicitly that all the items can be found in the ‘Data Extraction Instructions’ document (https://osf.io/ewm7x).

Thank you for the suggestion, we have added a paragraph describing the domains and items and pointed to the instruction document explicitly, towards the end of the Data Extraction section (page 18).

7. Regarding the data extraction items, I have some concerns. First, about the section ‘Open & Reproducible Scholarship Content’, I wonder whether items such as ‘Open

access publication', 'Open source software used' or 'preprinting archiving' are really necessary. While these are of course desirable properties of a study, I don't see how these aspects relate to the quality of a primary study. I would be surprised if any quality assessment tool for primary studies assessed this aspect. Therefore, I think not assessing them could save time.

We understand your viewpoint here. We agree that it would be surprising if any traditional tools assessed this, especially given issues of open and reproducible scholarship are part of a newer paradigm shift that will have taken place after many of the tools were potentially created. We believe that open and reproducible scholarship should be part of the quality assessment, however, and by extracting data around it we would then be able to demonstrate more clearly that there is a gap here to be filled (as is our hunch!). We've added a section about this stance into the manuscript itself, within the final paragraph before the Aims (in which we also address point 3 of the recommender's comments).

8. Besides, I don't understand how 'Sample size estimator' relates to 'Open & Reproducible Scholarship Content'. Shouldn't this be an aspect of internal validity (underpowered studies have higher probability of Type I error and, because of the lower precision also Type II error)?

We understand the confusion here. What we really mean is an assessment of whether a study has transparently justified its sample size, rather than whether the sample size was appropriate or whether the study was - for example - underpowered. We've updated the data extraction item accordingly to say 'Sample size justification'. We hope that this is clearer.

9. Similarly, regarding the section 'Integrity Assessment', I am not sure whether 'Methodology assessment' or 'Analytical approach assessment' are in the right place. Wouldn't 'internal validity' be more appropriate as well?

I think you could definitely argue the case that they belong under Internal Validity, but we don't believe this necessarily means they *don't* belong within the Integrity Assessment category. We believe that the most important thing is that we're capturing the information, regardless of how we've categorised it. We can definitely see arguments for doing it either way, and for that reason we've decided to leave the categorisation as-is.

10. Also, although the current checklist of items is very comprehensive, it is likely that the tools you find will assess aspects that you have not yet considered. I see you mentioned that you will add them as 'ad hoc' items. Do I understand correctly that if you find this item in the middle of the data extraction process, it would be evaluated retrospectively for all other tools?

Correct, we have added a sentence in the final paragraph of the 'Data Extraction' section clarifying what we will do if we decide to add new items (page 18).

11. Lastly, another aspect that might be interesting to know is whether there are studies that assess the validity and/or inter-rater reliability of the tools. However, I understand that

looking for such studies may take considerable work and may be outside the scope of this review, but I think it is very important to know that these tools are adequate. Related to this, it might also be interesting to know if any training is specified as necessary to apply these tools.

Thank you for the suggestions! While we agree that both the validity/inter-rater reliability of the tools and the information about training are valuable and interesting, we have decided to only include the item that assess training in this review. The first item would not be feasible to assess at this time. We have added the item in the metadata domain which evaluates whether the tool suggests training before use, and have updated the instruction and rmarkdown files to reflect the change (Item 11).