Reviewer’s report

Title: Comparison of different rating scales for the use in Delphi studies: Different scales lead to different consensus and show different test-retest reliability

Version: 0 Date: 19 Jul 2019

Reviewer: Reviewer 2

Reviewer's report:

PEER REVIEWER ASSESSMENTS:

OBJECTIVE - Full research articles: is there a clear objective that addresses a testable research question(s) (brief or other article types: is there a clear objective)?

Yes - there is a clear objective

DESIGN - Is the current approach (including controls and analysis protocols) appropriate for the objective?

Yes - the approach is appropriate

EXECUTION - Are the experiments and analyses performed with technical rigor to allow confidence in the results?

Yes - experiments and analyses were performed appropriately

STATISTICS - Is the use of statistics in the manuscript appropriate?

Yes - appropriate statistical analyses have been used in the study

INTERPRETATION - Is the current interpretation/discussion of the results reasonable and not overstated?

Yes - the author's interpretation is reasonable
OVERALL MANUSCRIPT POTENTIAL - Is the current version of this work technically sound? If not, can revisions be made to make the work technically sound?

Maybe - with major revisions

PEER REVIEWER COMMENTS:

GENERAL COMMENTS: This is an interesting paper on the effects of using different rating scales on achieving consensus in Delphi studies. In addition, the authors evaluated the test-retest reliability of three different rating scales. The design of the study is adequate and the results are important. The presentation of methods and results can be improved in some places, mainly for clarification.

REQUESTED REVISIONS:

Major comments

* In general, I recommend to separate out the two parts of the study more clearly throughout the paper: (1) effects of using different rating scales on consensus and (2) test-retest reliability of different rating scales. This would improve the readability of the paper.

* In the current study, the 9-point rating scale was divided into three categories as is usually done, where categories 9, 8, 7 = "important"; 6, 5, 4 = "important, but not critical"; 3, 2, 1 = "not important". Did the authors also consider using another classification, e.g. only 9,8=important and 1,2=not important? What would be the impact of such an alternative classification on the consensus and on the differences between the rating scales?

* The sample size calculation was based on assessing test-retest reliability, but the main aim of the study was to examine the effect of different rating scales on consensus. I think it would have been better to estimate the required sample size for detecting a relevant difference in percentage consensus among the three ratings scales, or detecting a relevant difference in percentage main goals between the three scales.
The results section is a bit difficult to follow. This section would clearly benefit from a better distinction between (1) results concerning the effect of different rating scales on consensus, versus (2) results of test-retest reliability. With regard to the first aim, I recommend to make a clear distinction between (1a) the effect of different ratings scales on the percentage main goals; and (1b) the effect of different rating scales on percentage consensus (using different cut-offs for consensus). I recommend to present the results in the order described above (1a, 1b, 2). Following this line of reasoning, the sentence in line 281 starting with "For a threshold of 75%...." can be removed, it is not required at this point. I also recommend to use the term test-retest agreement, rather than just agreement, to make it easier for readers to understand that you refer to test-retest reliability and not to consensus. Finally, it should be made more clear whether results refer to individual domains or to the average of the 19 domains. For example, in line 307, I do not understand where the Kappa’s of 0.63 to 0.78 refer to.

The first paragraph of the discussion should provide a short summary of the main findings of the study, regarding (1) results concerning the effect of different rating scales on consensus, versus (2) results of test-retest reliability, in this order. The current paragraph is not a comprehensive summary of results and starts with the reliability results, while the main aim of the paper was to examine the effect of different rating scales on consensus.

A clear conclusion on which rating scale has the best test-retest reliability is lacking. I think that the 3-point rating scale has the best reliability in your study. If the authors agree, this should be stated as a conclusion.

I agree with the authors that the decision of which ratings scale should be used, needs a critical discussion. I would like to add that this choice is a matter of validity (accuracy) rather than reliability.

Section 4.2 would also benefit from a more clear distinction between the two aims of the study, see comments above.

No feedback was provided to participants in the second round of this study, in contrast to usual Delphi studies aiming for consensus. I was wondering what the participants were told with regard to the aims of the study? Did they knew it was a reliability study? It might be helpful for readers to provide this information in the methods section.

The conclusion paragraph is not a comprehensive summary of the conclusions from the study. Again, it would be helpful to provide two conclusions on (1) the effect of different rating scales on consensus, and (2) the test-retest reliability of the three rating scales. The fact that the authors "found no clear evidence to emphasize a rating scale solely based on test-retest reliability properties" was a fact from the beginning, this is not a conclusion from this study.
In the very last sentence of the conclusion the authors refer to the preference of the 3-point scale, this can perhaps be better discussed earlier in the discussion, e.g. in section 4.4 (see also below).

In Table 2 I would like to see percentages main goals and absolute percentages consensus, not just symbols. I also think it would be helpful to switch the upper and lower part of the table.

Table 3 is unclear and seems to contain some errors (or I have misinterpreted the Table): I do not understand the first three rows of the table: I do not understand what is meant by "counts of ratings"; "prevalence" and "agreement". What does the weighted Kappa refers to, the average of the 19 domains? Is the weighted Kappa for the nine-point scale really 0.78 with a confidence interval of 0.78-0.78? This does not seem to be in line with the range presented underneath this result. Also, the difference between the weighted Kappa (0.78) and the unweighted Kappa (0.41) seems much larger for the nine-point scale than for the other scales, is this really correct? A decimal seems to be missing for the unweighted Kappa of the transformed nine-point scale?

Minor comments

In the introduction, there are a couple of terms and phrases that need clarification:

- What is meant by "global goals"?
- What is mean by "such indication criteria" and how does this relate to the global goals? If I understand the paper correctly you are seeking consensus on treatment goals, not on treatment indications or did I miss something?
- What do you mean by "indication criteria were adapted with respect to the modifiability of predictors that determine the outcome"? Is this relevant for achieving consensus on treatment goals or can it be left out to avoid confusion?

In the methods section, there is some overlap between sections 2.2 and 2.3, repetitive information can be removed

Line 211: What does Kappa0, KappaL, and KappaU stands for?

Discussion, line 326: I agree with the authors that pre-registration and pre-defined analysis plans for Delphi studies are relevant, but they don't solve the problem of the effects of different ratings scales on consensus, nor impact reliability of rating scales.
Section 4.4 refers to implications for further research, but does/could also discuss implications for future Delphi studies (e.g. discuss the preference for the 3-point scale).

The discussion is rather long. For example, the first paragraph of section 4.4 can be shortened, the last paragraph of section 4.4 is not relevant for this paper and can be removed. Section 4.5 is also not relevant for this discussion and can also be removed.

The lay-out of Table 1 can be improved.

The title and legends of the last figure can be improved: I recommend to include test-retest agreement, and include a legend about the transformation of the five- and nine-point scales.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable
Declaration of competing interests

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

This reviewer has been recruited by a partner organization, Research Square. Reviewers with declared or apparent competing interests are not utilized for these reviews. This reviewer has agreed to publication of their comments online under a Creative Commons Attribution License attributed to Research Square and was paid a small honorarium for completing the review within a specified timeframe. Honoraria for reviews such as this are paid regardless of the reviewer recommendation.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal