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: 1 Date: 24 Sep 2019

Reviewer: Lidwine B. Mokkink

Reviewer's report:

peer review - BMRM - D 19 - 00053 R1

Thanks for the opportunity to review the manuscript again. The authors improved the paper, but I do have some issues left.

I'm still confused about the data collection used in this study. In my understanding, the first round of the Delphi study is conducted twice. And only this data is used in this study. However, to be able to answer both aims (i.e. (1) to explore the influence of rating scales and different consensus criteria on the selection of treatment goal(s) and (2) to investigate the test-retest reliability of the rating of these treatment goal(s) on different scales used in Delphi studies), two different designs are required. For the first aim, the Delphi study should have been finished, so at least two subsequent rounds should have been performed, as only then you decide on the final treatment goals. In a Delphi study, the subsequent round should be influenced and build on by the results from the previous round, so they (should) deviate in content from each other. For the second aim, the exact same round should be repeated after e.g. one or two weeks. To me, only the second design has been conducted, so the first aim cannot be answered, as the second (and even third) Round seems not to be incorporated in this study. But the authors may mean something else by 'questionnaire', and 'survey'.

I still have my doubts about the relevance of a reliability study. Results of a Delphi study are based on group consensus, so an agreement type of parameter which tells you about the 'precision' of a rating given by an individual, seems not relevant at all. When I think of the quality of a Delphi study, I would like to know in the first place about the whether the final result (in this study a list with important treatment goals) is robust. I challenge the authors to explain in the introduction of the paper the relevance of knowing the test-retest reliability and agreement of different (even 'adjusted') ratings scales.
Some of the people completed the second assessment only 2 days after the first assessment. As the same 19 questions were asked three times in each survey, the chance of remembering your scores is quite high in the second assessment. This will have influenced the results of the reliability results. Can the authors discuss this issue?

The authors choose to collapse the 5- and 9-point scales, and consider the option 'somewhat important' on the 5-point scale as a secondary goal. First of all, why are these scales collapsed, are they also collapsed in the reliability analyses? Did this influence the conclusions the authors draw? What happened when the option 'somewhat important' of the 5-point scale was considered as a main goal? What is the patients' opinion about this decision?

The authors state (page 30 line 16-21): "In the context of treatment goals for TKA, it is clinically necessary to distinguish between a treatment goal that must be achieved with a direct clinical implication, in comparison to questioning the relative importance of treatment goals such as the five-point and nine-point scale. Therefore, in our setting to develop a set of global treatment goals, we preferred the three-point scale because further translations/transformations of importance into a clinical context of "main goals" were not required". Why did they investigate other rating scales in the context of this Delphi study? Shouldn't you always choose the rating scale in the context of the aim of the Delphi. How can another study use the reliability results which aims to understand the importance of goals? Isn't it all about (content) validity?

Statistical analyses page 25 line 28: 'we calculated reliability of the test-retest of each scale'. Only overall and mean results were given. However, I can I use this information? And what is more relevant to look at weighted or unweighted kappa's? Why are both presented? I think the authors should make a choice (based on the goal of both kappa's, not on the results) and only present those. (Probably weighted kappa's, and it takes into account the closeness of deviations).

Why are the correlations calculated between the rating scales? If it is to assess validity, hypotheses about expected correlations should be specified (and an aim should be added).

What do the authors mean by: 'results should be queried in a context-based dichotomous manner' - page 31 line 21?

Page 31 line 38/39: 'However, the prevalence effect… kappa statistics'. This is because kappa is a reliability parameter (similar as the ICC). Such measures are dependent on the homogeneity of the sample.
Page 32 line 48/49: 'in an online survey there is usually no possibility for discussion'. I strongly disagree. If the researchers conducting the Delphi study did not take the opportunity to ask for arguments - which is indeed often the case - it is a poorly conducted Delphi study. However, if these arguments are asked, thoroughly read, and taken into account in different ways in the second round (e.g. by using pro and contra arguments, by responding on deviate opinions, and rate such issues within the panel) a discussion is indeed possible. However, it takes more of the researchers, and the panelists. In my opinion, in a high quality Delphi study the focus is on the arguments and not on the simple ratings of consensus, without trying to have a discussion. Asking and listening to the arguments can avoid a face-to-face meeting, which is much more expensive and time-consuming.

Page 34 line 24: I agree that the selection of the rating scales and consensus threshold should be based on the specific context and (in my opinion) purpose of the study, but I do not understand how this studies adds to this conclusion.

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

No

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

Unable to assess

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

Unable to assess

**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.

Not relevant to this manuscript

**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.

I declare that i have no competing interests

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