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: 14 Mar 2019

Reviewer: Lidwine B. Mokkink

Reviewer's report:

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

Delphi studies are increasingly being performed. I think it is a strong and useful design in cases where e.g. methodology, guidelines, COS or taxonomies are developed.

A Delphi study is a qualitative research design, in which proposals are made, and panelists are asked to rate their agreement to these proposals, and providing arguments for their ratings. The ultimate goal of a Delphi study is not just to reach consensus on as many as possible proposals, but rather to find all valid arguments to understand which proposals are relevant, and how these proposals can be improved. Based on the proposals that reached consensus and arguments given, for example, a COS, a guideline etc can be developed. In my opinion, the focus in many Delphi studies - in general, but also in this paper - is way too much on only rating consensus, without using the arguments. Often, panelists are asked to reconsider their rating on a proposal, only by providing the distribution of ratings per proposal of the panel, and their own previous voting, without providing a summary of pro and contra arguments. However, as a panelist you need the summary (as you don't have the time to read the whole feedback report). The impact of providing a summary of arguments for each proposal in the subsequent rounds would have been (much more) interesting to have investigated. As this is not possible anymore, in the discussion it could be argued that a Delphi study is an qualitative methods, and researchers who perform a Delphi study should make much more use of the arguments.

In the introduction, the authors talk about 'statistical methods' to achieve consensus. The methods used are just calculations, distributions, and applying different thresholds. Where are the statistics? And where are the arguments provided by the panelists? These arguments are the essential part of a Delphi study.
I'm a bit confused about the aim to evaluate test-retest reliability. The goal of a Delphi study is that people will change their opinion, based on arguments given by other panelists. However, the assumption in a reliability study is that (1) the subjects (panelists in this case) are stable and are not changed, and (2) ratings are independently given (i.e. without knowledge of the previous rating). Are these assumptions met? Moreover, it is not clear which scores are considered as test and which as retest should be explained. In the section on statistical analyses at line 226 it is stated that between the two surveys reliability is calculated (however, in that case part of the people are likely really changed). At line 254 and further it is written that scales are converted into three point scales. Is the reliability between different scales in one and the same round evaluated? However, in this case it is likely that the score are not independently give, because people could easily go back to previous scores. It is also stated that 'reliability was analyzed in addition to the original scales'. As the authors explain, to calculate kappa's the same unit of measurement is needed. How were these original scales assessed? The bottom line is why is reliability relevant to investigate?

In the method section / study procedure it is written that patients were asked to evaluate 19 pre-defined treatment goals using three different rating scales. They were asked to fill out a pen-and-paper version of the survey, as they received a pre-paid self-addressed envelope. Were these 19 goals precisely the same goals and were the questions or proposals in each of the three blocks formulated exactly the same? If not, what were the differences of the questions? Is so, people were able to look up their ratings on the different scales: how did this influence the other ratings, and why did the authors choose to ask the same questions to all people. Why not, e.g. making three subgroups and see whether each subgroup would come to the same consensus?

Patients were asked to indicate what their preferred rating scale was (line 172). It now seems this question was asked in the beginning of the survey. It seems easier to answer - and thus more valid - to ask this question at the end, after they experienced all three different scales.

The three scales are from important on the left to not important on the right. In Delphi studies I participated in or performed, it is always the other way around. Why did the authors choose to do it this way, and would the other way around make a difference?

Sample size calculation: why is the study sampled on the secondary aim (kappa) and not on the primary aim (differences in reaching consensus).

Why was a minimal expected kappa of 0.5 used.
LINE 216, 217: 'since sample size would decrease with an increasing number of categories': really? Do you have a reference? You need filled cells, I would argue that you need larger sample sizes with more number of categories, since it is harder to distinguish between people when people can be measured more precisely.

Line 315 'a wide range of the reliability…' this seems a somewhat awkward sentence as the assumption is not that people are stable.

What is meant by 'covers a range from' in the sentence 'between the two extreme threshold…reaching consensus' (line 322-324)? At a threshold of (exactly) 60% 16 goals reached consensus, and when consensus was considered to be reached at 90% agreement zero goals reached consensus?

Why is the discussion section started with the second aim?

Please, discuss the lack of asking arguments of panelist in Delphi studies, and the lack of providing a summary of pro and contra arguments in Delphi studies (in general and in this study) in the discussion section.

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.

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.