Author’s response to reviews

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

Authors:
Toni Lange (toni.lange@uniklinikum-dresden.de)
Christian Kopkow (Christian.Kopkow@hs-gesundheit.de)
Jörg Lützner (Joerg.Luetzner@uniklinikum-dresden.de)
Klaus-Peter Günther (Klaus-Peter.Guenther@uniklinikum-dresden.de)
Sascha Gravius (Sascha.Gravius@ukb.uni-bonn.de)
Hanns-Peter Scharf (Hanns-Peter.Scharf@umm.de)
Johannes Stöve (Johannes.Stoeve@st-marienkrankenhaus.de)
Richard Wagner (Richard.Wagner@fdk.info)
Jochen Schmitt (Jochen.Schmitt@uniklinikum-dresden.de)

Version: 3 Date: 20 Dec 2019

Author’s response to reviews:

Point-by-point response:

Manuscript# BMRM-D-19-00053R2

Dear Anna Melidoni,

Thank you for the third revision of the paper. We have taken the remaining concerns of our reviewer 2 very seriously. However, we feel that there has been a misunderstanding as of how our results should be interpreted. We have changed the corresponding parts in the manuscript to avoid this confusion.
However, we want to point out that while reviewer 2 has entirely changed a number of his overall positions throughout the review process, the concerns he names refer to a fraction of the overall manuscript only. Indeed, the concerns about labelling of the scales relate to sensitivity analyses and not the general design or interpretation of our results. The misunderstanding here may be that we indeed had not very clearly explained that our “labelling” of categories is merely just a labelling used for the mapping of scales. What matters is how scales are transformed into each other, rather than how the resulting categories are named. This transformation was, from the beginning, discussed in depth throughout the manuscript and we now have taken care of explaining it more thoroughly in the text and tables.

The scales that have been used in the original surveys along with their labelling are scales that are established across disciplines. Hence, there is, as reviewer 2 states indeed a “problem [] with the confounding of the number of response alternatives with the response labels”, but this problem is by design of these scales. Our aim is exactly to investigate this problem. For stable statistics, we used the mapping onto a three point scale for the reliability as a sensitivity analysis. Therefore, we reported the results for both the original and transformed scales.

As our mapping and re-labelling of scales is not part of the study design, but a sensitivity analysis, we disagree with reviewer 2 on that this strategy is interfering with neither the study nor our results.

We feel that our data and analyses provide the valuable finding that researchers should be aware of the choice of rating scale both when identifying treatment goals and their use within Delphi studies.

We kindly ask you to consider our point of view.

With kind regards,

Toni Lange on behalf of all co-authors

Reviewer #2 comments

Comment 1

For the most part, the authors have addressed the comments of the reviewers.
However, I have one major concern about the paper. The scales differ from each other in two regards: the number of points on the scale (that is fine), and the labeling of the points (that is a problem). I can see some congruence between "main goal" and "very important." But, I think there can be differences in interpretation between "secondary goal" and "a little important" or between "I do not expect this" and "no goal." Thus, it is difficult to know if the differences between scales is due to a different number of items or to different labels.

As mentioned, the major problem is the confounding of the number of response alternatives with the response labels.

Response 1

We feel that there has been a misunderstanding with respect to “labelling” of scales.

The concerns of reviewer 2 do not affect our investigation of the ‘effects of using different rating scales on consensus’. For this question, “very important” and “9,8,7” were mapped onto a top category called “main goal”, which reviewer 2 deems as not critical.

We agree with the reviewer that this study cannot explain the reason for the differences in reliability of the scales, be it number of categories or their original labels. The point of the manuscript is to illustrate that the reliability differs vastly and that the consensus depends on the chosen scale.

For stable statistics, we used the mapping onto a three point scale for the reliability as a sensitivity analysis. Therefore, we reported the results for both the original and transformed scales. We want to point out that our “labelling” of categories is merely just a labelling used for the mapping of scales.

What matters is how scales are transformed into each other, rather than how the resulting categories are named. This transformation was, from the beginning, discussed in depth throughout the manuscript and we now have taken care of explaining it more thoroughly in the text and tables.

We furthermore addressed the effect of transforming scales in the discussion and separated the terms “mapping” and “labelling” throughout the manuscript.

We very much hope that we have been able to address your concerns appropriately.