Author’s response to reviews

Title: Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) Checklist

Authors:

Lidwine B Mokkink (w.mokkink@vumc.nl)
Caroline B Terwee (cb.terwee@vumc.nl)
Elizabeth Gibbons (Elizabeth.Gibbons@dphpc.ox.ac.uk)
Paul W Stratford (stratfor@mcmaster.ca)
Jordi Alonso (jalonso@imim.es)
Donald L Patrick (donald@u.washington.edu)
Dirk L Knol (d.knol@vumc.nl)
Lex M Bouter (lm.bouter@dienst.vu.nl)
Henrica C De Vet (hcw.devet@vumc.nl)

Version: 2 Date: 20 August 2010

Author’s response to reviews: see over
Dear BioMed Central Editorial Team

Thank you very much for the opportunity to revise and resubmit our manuscript on the inter-rater reliability study of the COSMIN checklist. Below you will find our detailed reaction to the comments of the reviewers (in bold). In the manuscript we indicated our changes highlighted in yellow.

Kind regards,

Wieneke Mokkink

**Reviewer 1: Jan Kottner**

Major Compulsory Revisions

1. Background, third paragraph: I completely agree that users of the COSMIN checklist should independently score the items identically when judging the same article. Therefore the aim of this study should rather be determining interrater agreement than reliability. Both terms are usually used interchangeably but as pointed out for instance by De Vet et al. (2006) “agreement” and “reliability” are conceptually distinct. Reliability refers to the ability of scores or ratings to distinguish between objects (articles in your case) or subjects (De Vet et al. 2006, Kottner et al. 2010). Regarding your research it would be reasonable to focus on the question of agreement (Were the scores identical?) between the repeated ratings.

Thank you for this suggestion. We absolutely agree. We included agreement in the aim of the study and tried to make a clearer distinction in the text (and title) of the manuscript between agreement (calculated as percentage agreement) and reliability (calculated as Kappa).
2. Methods, first paragraph: Again, I think you were rather interested in agreement than reliability when using the COSMIN checklist by many different raters, or both?

See my response on the first remark.

3. Methods, Statistical Analysis, third paragraph: You excluded the 22 items with the response option “NA” from the reliability calculation because you regard them as (multi)nominal. Firstly, albeit limited in general it is possible to calculate kappa coefficients for multinominal classifications which is frequently done. Secondly, from a practical point of view it would be important to know, how reliable users of COSMIN judge “NA” vs. “yes/no”. This decision is somehow comparable to Step 1 of COSMIN. In fact, all the “NA” items can be regarded as dichotomous-ordinal (DO) scales in the sense of Cicchetti et al. (2006), because of the distinction between absence and ordered presence of a trait in one scale. Specific weighting schemes or dichotomization might be applied in these situations.

The calculation of kappa coefficients for multi-nominal classifications is not possible because of our complex design. In the article by Cichetti the two raters evaluated all subjects. In our one-way design, only some of the raters rated the same article, and not all items of the COSMIN checklist were completed in each article. Moreover, some articles were rated by 2 or 3 raters, and others by 4, 5 or even 6 raters.

You suggest dichotomizing the nominal items. We would not primarily be interested in NA vs other responses (i.e. yes, ?, no), but also in yes vs others, no vs others, and - if applicable - ? vs others, to see for which response option raters have any problems. As a consequence the Table would become too confusing, and therefore, we ask to abandon this suggestion.

5. Methods, Statistical Analysis, fourth paragraph: Please clarify what exactly is considered as “poor” or “moderate” .... Values of kappa coefficients per se are neither poor nor good. The arbitrary limits are often linked either to reliability or agreement (please clarify) but more important is the practical relevance for the COSMIN application. What (relative) error seems to be acceptable? When using such limits I would also recommend focusing rather on the lower limits of the confidence intervals.

We used these limits proposed by Fleiss (1981). He uses the term ‘moderate to good’, for the whole range (0.40 to 0.75).
Although we use the limits in the text to describe the results, in the Table we present all results, so if people want to use different limits, they can apply those on the results. Moreover, this study was performed to see which items showed lower reliability and agreement than others. For these items either the wording in the checklist or the instructions in the manual needed to be improved. We did not have the intention beforehand to delete any items from the COSMIN checklist, based on arbitrarily chosen limits. For the same reason we did not calculate CIs.

6. Methods, Statistical Analysis, fifths paragraph: Numerous authors have discussed the so called “base rate problem” and how low or high prevalence or skewed data distribution affects reliability estimates. However, the point of Vach (2005) and many other authors is that the dependence of kappa and other reliability coefficients on the composition of the sample is in fact a desired property. Therefore it is illogical to calculate percentage agreement BECAUSE of low reliability coefficients. Low kappa values in your study indicate the inability of the respective item scores to clearly distinguish between the rated papers. This is not a drawback of kappa or of your study; it just indicates that many ratings were similar (what can be interpreted as a desired result). On the other hand you should calculate percentages of agreement, because this was your study aim (see above).

Thank you for your clarification. We deleted our comment that we calculated agreement BECAUSE of low Kappa values. We now explain that agreement and reliability are two different parameters with different interpretations. We explain the interpretation of both measures on page 6-7 and put equal emphasis on agreement and reliability.

If there was a clear conceptual distinction between agreement and reliability you could also omit the problem of “skewed distributions”. The problem here is that 75% is clearly arbitrary again, why not 80% or 70%. Additionally, even if the data is “extremely skewed” kappa might be 1.0 (Example: 99 times raters agree that a certain trait is present, one time raters agree that this trait is absent (n = 100)). This should also considered in the COSMIN checklist manual (page 11).

Thank you, on page 7 we now describe what the percentage agreement reflects, i.e. ‘This measure indicates how often raters who rated the same items on the same articles choose the same answer category.’ However, we are interested in
both the percentage agreement and the kappa. We made a clearer distinction in the text between agreement and reliability, but we kept our remarks about the influence of skewed data on the Kappa. However, we rephrased this paragraph in the discussion on page 10.

We thank the reviewer for critical reading of the COSMIN manual and suggestions on how to improve the text.

7. Methods, Statistical Analysis, fifths paragraph: Is there any rationale to label > 80% as appropriate. In scoring practice there seems to be hardly any difference between 79% and 81%. Additionally, the varying numbers per item influence the precision of the calculated proportions limiting their comparability.

We chose it arbitrarily. We acknowledge this now in the method section on page seven. Although we used these limits in the text to describe the results, in the Table we present all results so if people want to use different limits, they can apply those on the results.

8. Results, interrater-reliability: Please reconsider how to deal with the 22 NA items (see above). The presented rationale that kappa coefficients could not be calculated due to nominal response options is somehow weak and in a strict sense not true.

We agree that kappa is able to handle multi-nominal data. However, we have a very specific situation, because the number of raters who rated items per article is not equal. The intra class kappa is able to handle dichotomous data, but not multi nominal data. In the results section we added this explanation. See also our response on remark 3.

9. Discussion, Reasons for low kappas: The calculated interrater reliability coefficients simply reflect the measurement situation. There is no “contradictory finding”, because kappa values indicate reliability, not agreement.

We agree, and deleted this sentence.

10. Discussion, Strength and weakness of the study: Under this heading only weaknesses are listed. Consider to change the heading or list some strengths: you used samples that are representative to intended users (raters) and a wide range of papers.
Also the rater object crossings (no pairs, no ordering) enhances generalisability of your results and leads to conservative estimates.

**Thank you for your suggestions, we have now added these issues in the discussion section (p11).**

11. Discussion, Strength and weakness of the study, last paragraph: I strongly disagree with this conclusion. Kappa coefficients are easy to interpret and they are perfect measures of reliability. Please also reconsider the second sentence, because proportions of agreement are perfect agreement measures. There are two main reasons why the last statement is also problematic: (1) Proportions of agreement can’t lead to “artificial” high agreements, because they simply reflect the concept of agreement what seems to be the focus of your work. (2) It is true that Cohen (1960) introduced a so called “chance corrected agreement” but this rationale is debatable. Kraemer (2002) called this a “historical curiosity”. Today, we also do not “chance correct” sensitivity or specificity or other predictive values. Why for agreement? Furthermore, in your study you explicitly refer to Kraemer’s intraclass kappa as an ideal measure of reliability which she conceptualized as population parameter, not as (descriptive) sample estimate.

**Thank you for your remark. We agree and deleted the section.**

12. Conclusion: Based on the changes of the manuscript the conclusion should be reworded.

**We agree, and changed the conclusion (page 14) into ‘The items of the COSMIN checklist showed high percentage agreement and low kappa coefficients.’**

References:


Minor Essential Revisions

1. Background, third paragraph: In a strict sense one cannot determine interrater reliability/agreement of items because reliability/agreement is a property of scores or ratings, not items or instruments. To make this clear I would suggest inserting “interrater reliability of item scores” or the like.

**Thanks, we changed this as suggested.**

2. Methods, first paragraph: What was the rationale for your decisions: four ratings per article, one box per 20 articles? Did you perform any sample size considerations?

**It was an arbitrary choice. This is now added at page 4.**

3. Methods, Statistical Analysis, first paragraph: Interrater reliability of COSMIN item scores or ratings (see above).

**Thanks, we changed this as suggested.**

4. Methods, Statistical Analysis, second paragraph: Using the described weighting scheme is appropriate for ordinal scales, but is “yes”, “?”, “no” ordinal indeed? Is “?” more than “yes” and is “no” more than “?”? Furthermore, when looking at the respective items in the boxes there is another order: “yes”, “no”, “?” and when studying the COSMIN manual the ordered logic becomes not apparent.

**In the COSMIN checklist the order is different than we used in the analyses. We now describe this at page 6.**

5. Table 2: This table is very complex with much information. If you omit the “distribution issue” you could delete the italics. I would also recommend to indicate dichotomous and ordinal items (given that they are ordinal indeed), because the values of obtained reliability and agreement coefficients depend on the number of categories.

**We tried to increase the readability of the Table by omitting the italics indicating a skewed distribution of some of the data. Instead of this we added footnotes: for this distribution and also indicating which items were dichotomous and which items were nominal. We also deleted the variance components to get a more simple Table (both in Table 1 and 2).**

6. Results, last paragraph: Consider using CIs. The last statement about the small numbers is somehow weak. What does “with caution” mean? Despite low sample sizes the coefficients might be precise.

**We have not calculated CIs for three reasons: Firstly, we were interested in which items needed improvement, and the CIs would not give more information than the
point estimates. Secondly, the CIs for the intra class kappa are very difficult to calculate and don’t come automatically with the computer program. Thirdly, presentation of these CI would make Table 2 even more difficult to read. Therefore, we think that warning the readers for the small sample sizes in this way is the best solution.

7. Discussion, first sentence: I would recommend to say “… reliability of COSMIN item ratings…” (see above).

Thanks, we changed this as suggested.

8. Discussion, Other measurement properties …: I would recommend deleting this section, because it neither belongs to your question and the content of this discussion is not linked to the methods and results.

Thanks for this suggestion. We deleted this section.

Discretionary Revisions

Consider to change the term “inter-rater” into “interrater”.

Thanks for your suggestion. I’m not a native speaker, but I noticed that both spelling methods can be used. Within COSMIN we use ‘inter-rater’; therefore, we would like to keep it as it is.

Minor issues not for publication

I am not a native speaker and my English is clearly limited, but I feel that there are some linguistic errors:

1. Methods, first paragraph: Please insert “s” in the word part: “Consequently, only parts of the …”

Thanks, we changed this as suggested.

2. Methods, Statistical analysis, third paragraph: I think you should write: “Neither kappas nor ICCs are …”

Thanks, we changed this as suggested.

3. Results, Interrater reliability: I would suggest using kappa coefficients than kappas. Furthermore, the plural of kappa would be kappas, not kappa’s.

Thanks, we changed this as suggested.

4. Results, Interrater reliability: Please insert “s” in “components”

Thanks, we changed this as suggested.
5. Results, Interrater reliability: Please replace “taking” by “taken”

Thanks, we changed this as suggested.

Reviewer 2: András Keszei

In this manuscript Mokkink and colleagues describe the inter-rater reliability of the items of the COSMIN checklist, which was developed for the evaluation of the methodological quality of studies concerning the measurement properties of health-related patient-reported outcomes. Overall, the paper is well written, the study objective has been well defined, the methods are appropriate and limitations have been discussed.

Discretionary Revisions
1. Consider providing more information on the selected articles (e.g. distribution of levels of workload and setting), to allow the reader to judge to what extent they are representative of studies on measurement properties.

We agree that we could give more information to allow the reader to judge to what extent they are representative of studies on measurement properties. We decided to provide more information on the database, since we randomly selected the articles from this database. Therefore, we added in the method section ‘The bibliography includes evaluations of PROs with information about psychometric properties and operational characteristics, and applications where for example a PRO has been used in a trial as a primary or secondary endpoint. The online PROM database comprises records downloaded from several electronic databases using a comprehensive search strategy (details available on request).’.

2. Provide standard errors or confidence limits for the measures of reliability.

The results are basically used to indicate which items needed improvement. For this purpose the confidence intervals do not provide useful information. Moreover, the Table would become very large and even more complex. Therefore, we prefer not to give this information. (see also our response on the first reviewer, point 6 at page 6.)

3. The reliability is likely to be different among trained raters. It would be interesting to see the kappas among the subset of the raters who have the most experience with assessing measurement properties.
In our design we did not take such an analysis into account. Although we have information on how many years raters are involved, we can not perform subgroup analyses, due to the design. When distributing the articles over the raters, we did not account for the numbers of years experience. So articles can be rated by e.g. only experienced raters, or by two or more experienced raters and one less experienced rater or visa versa.

4. The authors discuss alternative designs for the study of reliability (Discussion section, Strength and weaknesses of the study subsection, 4th paragraph). Since reliability is assessed per COSMIN item, it is not completely obvious to me why a ‘many raters evaluating the same few articles’ could not be feasible. Would it be less efficient to select different sets of articles and raters for different measurement properties?

This is a matter of choice. With our design, we aimed to have many raters and many articles to achieve representativeness. By choosing only a few articles it is hardly possible to obtain a good representative sample of studies on measurement properties.

5. The authors might consider shortening the discussion on validity issues as it is not the main focus of the paper.

Thanks for this suggestion; we have now deleted these paragraphs.

Minor Essential Revisions:
1. Please spell out IRT before first use in Background, rather than in the Methods section.

Thanks, we changed this as suggested.

2. Statistical analysis, 2nd paragraph: It might be misleading to state that Kappas are not able to handle nominal items, since dichotomous items are also nominal. I would suggest saying that multi-categorical kappas are not recommended as a measure of reliability.

We agree that kappa’s are able to handle nominal items, except in the specific design we have used, i.e. in a one-way design calculating intra class kappa’s for multi nominal variables is not possible. We now added this reason. (see also our response on reviewer 1).
3. Results section, 5th paragraph: Similarly, Kappas were not calculated due to multi-categorical nominal responses.

We agree that kappa’s are able to handle multi categorical nominal items, except in the specific design we have used, i.e. a one-way design. (see also our response on reviewer 1).

4. Results section 5th paragraph last line: taken instead of taking.

Thanks, we changed this as suggested.

5. Conclusions section: typo, “we recommend making decisions in advance”

Thanks, we changed this as suggested.

6. Competing interests section: typo, “except”

Thanks, we changed this as suggested.

7. Table 2: In the subheadings indicate Box C, D, etc., as it is indicated in the first part of the table for Box A and B.

Thanks, we changed this as suggested.

8. Table 2 footnote: taken instead of taking

Thanks, we changed this as suggested.

Reviewer 3: Javier Rejas

- Major Compulsory Revisions

I have no major compulsory comments.

- Minor Essential Revisions

I have a comment for researcher which could improved easily some of the points raised by authors. The comment refers to the low inter-rater reliability showed in the research and the fact that many participants were enrolled with different level of training and experience in HR-PRO tools utilization and/or assessment. As apparently, authors knew the level of experience of participants, one interesting possibility could be calculating the reliability not in the whole sample but divided in groups of participants as per their level of experience. If reliability were high or appropriate as much as the level of experience increases, then, this could support one of the proposals of authors for future research that was to repeat the study with well-trained participants. In any case, this additional analysis would serve to test whether the level of experience/training in HR-PRO evaluation impact in the reliability of COSMIN tool.
In our design we did not take such an analysis into account. Although we have information on how many years raters are involved, we cannot perform subgroup analyses, due to the design. When distributing the articles over the raters, we did not account for the numbers of years experience. So articles can be rated by e.g. only experienced raters, or by two or more experienced raters and one less experienced rater or visa versa.

I also miss in the discussion section references to other similar tools in areas different than HR-PRO. For example, if similar tools exist to evaluate clinical trials, health economic evaluations, etc., and if yes to comment on how these other tools work from the inter-rater reliability point of view. Perhaps, the manuscript could benefit of adding an extra table including data on participants and descriptive information of the study rather than in a narrative form which is less impacting from the reader perspective.

Thanks for your suggestion. We found two studies on the reliability of checklists, and discussed these in the discussion section on page 13.

- Discretionary Revisions

It seems to me that some extra work could have been done in including more references in the manuscript. This comment is linked with previous minor comment about other similar tools than COSMIN but used to test inter-rater reliability in other areas of human health research.

Thank you, we added the references of the two reliability studies (see your previous remark).