Reviewer’s report

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

Version: 1  Date: 10 July 2010

Reviewer: Jan Kottner

Reviewer’s report:

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.

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?

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.

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.

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).

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).

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.

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.

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.

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.

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.

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

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.

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?

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

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.

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.

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.

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

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.

Discretionary Revisions

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

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 …”

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

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

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

5. Results, Interrater reliability: Please replace “taking” by “taken”

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests.