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:** 14 July 2010

**Reviewer:** András Keszei

**Reviewer's report:**

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.

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

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.

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?

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

**Minor Essential Revisions:**

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

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.

3. Results section, 5th paragraph: Similarly, Kappas were not calculated due to multi-categorical nominal responses.

4. Results section 5th paragraph last line: taken instead of taking.
5. Conclusions section: typo, “we recommend making decisions in advance”
6. Competing interests section: typo, “except”
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.
8. Table 2 footnote: taken instead of taking

**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:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests