Reviewer's report

Title: The Conceptually Equivalent Dutch version of the Western Ontario Rotator Cuff Index (WORC)(c)

Version: 1 Date: 24 January 2013

Reviewer: Ole Marius Ekeberg

Reviewer's report:

In this paper, the authors report the translation of the WORC index into a conceptually equivalent Dutch version and aims to test reliability and validity of the translated version. The question is well defined. The translation will be important for Dutch-speaking users.

The translation process is thoroughly described and follow accepted standards of cross cultural translation.

1. The sample size in the reliability study is adequate and the time frame between test and retest is sufficient. An 11 point shoulder hindrance score was used to exclude patents that had changed between tests.

- Major Compulsory Revisions

1. Please describe the RAND 36 and the constant score and the scoring process of all questionnaires (what instructions were given to participants?). Is the Dutch version of constant score properly translated? The measurement properties of the comparator instrument should be adequate, otherwise it would be difficult to judge negative results is due to poor translation of the instrument under study or poor quality of the comparator instrument.

2. Although some background information is given, a more thorough description of the patients in the study would improve the paper. The reader needs to know how generalizable the results are. What were the inclusion criteria for the study (clinical tests?, radiology?)? Duration of pain? What was the setting (secondary care?)?

3. The sample size of 50 patients in the reliability study is considered adequate, but this should be an a priori decision and should be reported in the methods section.
4. The ICC (2.1) and Chronbach alpha are adequate statistics of reliability together with SEM. The reader should be informed on how SEM was calculated; different methods of calculation may affect reported results. Is data for the reliability of the RAND 36 and the constant score available?

5. The cross-cultural translation is thoroughly described. The preferred analysis for cross-cultural validity is confirmatory factor analysis. In my opinion, the criterion validity analysis is flawed. The authors themselves state that there is no consensus in outcome evaluation after RCT surgery, thus there are no gold standard. RAND 36 and constant score are no criterion for evaluation of the WORC score. Further, the constant score is heavily affected by clinical measurements and is conceptually different from a HRQoL score. This may be the reason for a moderate correlation between the WORC score and the Constant score (0.60 (table 5)). The paper miss discussing possible implications of this finding (both scores are used in this population, but do they measure the same? Would the results of a clinical study depend on the choice of instruments? Validation (at least construct validation) is a process of hypothesis testing. What were you’re a priori expectations? What is the rationale in correlating the total WORC score with the domain scores in RAND36? In general, the results from the validity study lacks interpretation. Comparing the correlations in the present study (WORC and RAND36) and the results of Turkish and Brazilian translation (WORC and SF-36) does not add any support to the validity of this translation.

I would recommend either removing the entire validation study from the paper, or testing hypotheses that will increase our knowledge and interpretation of the WORC index scores.

6. Floor and ceiling effects are important to detect. Please describe how this was analysed (percentages of highest and lowest scores? What were the cut-offs?).

7. The results of the translation process are well described and different problems in the translation process and corrections are reported. It was surprising that there were no missing items. Do you think the correction was necessary because of general problems with the WORC index or only related to different cultural background of the Dutch and North American population? This is an issue of content validity and could be discussed in the discussion section.

8. The conclusions are not supported by the data or the analysis. Suitability to measure change in longitudinal studies requires establishing validity of change scores. Floor and ceiling effects may affect responsiveness but you have not established longitudinal validity in this study. I do not agree that these data proves that the Dutch translation is a valid health related quality of life questionnaire. The conclusion needs to be rephrased.

Minor Essential Revisions

1. Comments regarding the introduction:
   a. The authors’ main interest seems to be patients with rotator cuff tears. There is a lack of consensus in the evaluation of shoulder disorders in general, and not only in patients operated for RCT. Patients included in this study have different diagnoses, and the introduction should therefore be harmonized with the study
group.
b. A Dutch QoL for RCD is lacking, the WORC index was published 10 years ago.
c. There is no reason to list all available shoulder questionnaires, a reference of a review should be sufficient.

2. An 11 point shoulder hindrance score was used to exclude patents that had changed between tests. This is a commonly used strategy and a strength in design; still, possible implications could be discussed in the discussion section (natural variation vs artificially low SEM).

3. Table 5. P values should be reported as <0.001 and not 0.000

- Discretionary Revisions
None

**Level of interest:** An article of limited interest

**Quality of written English:** Acceptable

**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