Author's response to reviews

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

Authors:

Ronald N Wessel (r.wessel@antoniusziekenhuis.nl)
Nienke Wolterbeek (n.wolterbeek@antoniusziekenhuis.nl)
Anouk JM Fermont (a.fermont@antoniusziekenhuis.nl)
Henk van Mameren (henk.vanmameren@maastrichtuniversity.nl)
Heleen Sonneveld (h.sonneveld@meandermc.nl)
Sharon H Griffin (shgriffin@live.ca)
Rob A de Bie (RA.deBie@EPID.unimaas.nl)

Version: 2 Date: 22 April 2013

Author's response to reviews: see over
Dear editor, dear reviewers,

Thank you for your time and effort for reading and commenting on our paper. We revised the paper according to the comments.

**Associate Editor comments:**
The purpose of the present study was to cross-culturally adapt and evaluate reliability and validity of WORC for use in The Netherlands. The two reviewers provide important critics that need to be addresses before the manuscript can be considered for publication.
- Particularly the validation aspect of the paper is not conducted according to accepted standards.
  The validation process is explained in more detail and references are added.
- The description of the translations process may be shortened.
  The translation process has been shortened and is now limited to the method section.
- Details are missing regarding agreement.
  Details regarding agreement are added to the paper.
- The paper is evaluated in patients with rotator cuff disease, how many had tears considered for surgery?
  26 out of 30 patients with rotator cuff lesions had rotator cuff repairs. This number is not included in the text because responsiveness is not discussed in this article.

**Reviewer:** Ole Marius Ekeberg
Positive criticism:
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.
Description and references of the RAND-36 and Constant Score have been added. The Constant Score is a clinical score which is widely used by (Dutch) orthopaedic surgeons in international literature.

The questionnaires were filled out by the patients in the presence of the physical therapist. This information has been added as well.

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?)?

Thank you for pointing this out, background information has been added

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.

We added this to the methods section. No power calculations were performed for this study. However, based on the general recommendations for comparing measurement properties at least 50 patients should be included (Altman, 1999).

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?

Thank you for this remark, we have added information and a reference for the SEM.

Standard Error of Measurement (SEM) which estimates the reliability of the WORC was calculated as SEMagreement (de Vet et al., 2006). The larger the SEM the lower the reliability and precision of the instrument. The smallest detectable change (SDC), based on the measurement error, was defined as 1.96 * √2 * SEM (de Vet et al., 2006). References for the RAND-36 and Constant Score are added to the paper. In these articles information about the psychometric qualities of these instruments can be found.

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.

You are correct, this was not thoroughly described. Construct validation of the Dutch WORC was based on the original article of Kirkley et al., (2003) and the hypothesis described in this article were adopted for the current article.
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?).

We have added the definition and a reference. Floor and ceiling effects were considered to be present if more than 15% of the patients achieved respectively the lowest or highest possible score (McHorney and Tarlov, 1995).

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.

There were no missing items due to the setup of the study: “The questionnaires were filled out by the patients in a separate room in the presence of the physical therapist”.

We think that some corrections were necessary because of general problems and some to a different cultural background. These issues are 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.

We have rephrased the conclusion so that it better fits the results.

**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.

   The interest for the WORC was indeed fed by a need to evaluate rotator cuff repair. The text has been adjusted slightly.

   b. A Dutch QoL for RCD is lacking, the WORC index was published 10 years ago.

   You are correct, it has been a long time. But good ideas cannot be barred…

   c. There is no reason to list all available shoulder questionnaires, a reference of a review should be sufficient.

   This has been corrected.

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).

   When the difference between tests on this scale was more than 2 points, the patient was excluded for the test-retest analysis, regardless if this was caused by natural variation or by an actual change in condition.
3. Table 5. P values should be reported as <0.001 and not 0.000
We have changed the values to your suggestion.

Reviewer 2: Øystein Skare
Major Compulsory Revisions
1. Abstract: Results: Results of agreement (SEM) statistics is missing
We added this to the abstract.

2. Page 3 Methods: A description of the other instruments (RAND 36 and the Constant Score) with references is missing, and should have been placed here. There is no information about the measurement properties of the WORC (original and other language versions) or the other instruments, which should have been given in this section.
Description and references of the RAND-36 and Constant Score have been added.

The translation process has been shortened and is now limited to the method section.

4. Page 5: Reliability:
A) Define/explain what reliability, test-retest reliability, and agreement is.
We have extended the descriptions of these parameters and also added references to the paper.

B) What is the acceptable treshold for ICC? Remember references.
We added the thresholds and a reference to the method section. In general, 0.70 is recommended as a minimum standard for reliability; a correlation less than 0.5 is described as weak, whereas a correlation greater than 0.8 is described as strong (Nunnally and Bernstein, 1994).

C) Give information about how the SEM was obtained (formula), as the values is likely to differ according to the method used. The information is important for others to be able to compare results and to understand the strength of the findings (influence of potential errors).
We have added the formula and a reference of the SEM.

5) Page 5: Validity:
A) Explain/define validity and the different aspects of it in relevance to your study.
B) See the recommandation from the COSMIN group by Mokkink et al. 2010 considering the reporting of measurement properties of health related quality of life instruments. Analysis of correlation with other instruments is not a sufficient proof of validity in 2013.
Pearson Correlations Coefficients (r) were used to calculate the construct validity by comparing WORC with a general quality of life questionnaire (RAND-36) and a commonly used clinical shoulder score (Constant Score). Construct validation of the Dutch WORC was based on the original article of Kirkley et al. (2003) and the hypotheses described in this article were adopted for the current article. The a priori hypotheses are added to the article.

6) Page 5: What was the limit (%) for the floor and ceiling effects?
We have added the definition and a reference. Floor and ceiling effects were considered to be present if more than 15% of the patients achieved respectively the lowest or highest possible score (McHorney and Tarlov, 1995).
7) Page 5: Where is the statistical chapter explaining the analysis in details?
The statistical section has been made recognizable and the statistics section has been expanded.

8) Methods: Power calculations prior to the study or references to recommendations for power
(in this type of studies) is lacking.
No power calculations were performed for this study. However, based on the general
recommendations for comparing measurement properties at least 50 patients should be
included (Altman, 1999). This is added to the methods section.

9) Page 6: Results. Far too much details about the translation process.
The translation process has been shortened and is now limited to the method section.

10) Page 8: What is the limit for correlations (ranged) + (references)?
The limits for correlation and a reference are added to the statistics section. In general, 0.70 is
recommended as a minimum standard for reliability; a correlation less than 0.5 is described as
weak, whereas a correlation greater than 0.8 is described as strong (Nunnally and Bernstein,
1994).

11) The discussion should be re-written. Start with the findings. The presentation of MAP
Institute does not belong here. Discuss the findings more critically. What are the strengths and
the limitations in this study? How may they have an influence on your results and
conclusions. Where there any important methodological limitations present in the studies you
compared your results with?
Following your advise, we have rewritten the discussion.

12. Page 9: Conclusion. Should be re-written. In the first sentence, a result (no floor and
ceiling effects) is given, and does not belong here. The validity is not evaluated according to
recommendations in the recent methodological literature.
We have rewritten the conclusion of the paper.

Minor Essential Revisions
1. Background: Last sentence in the first passage may be misleading. You have stated that
quality of life measurement tool specifically suitable to investigate patients after rotator cuff
repair are lacking, but WORC is available and found to suitable by others.
You are correct, we meant to say that there are no Dutch QoL questionnaires available. The
sentence has been adjusted.

2. Page 3 Purpose of the study: Agreement should be mentioned.
We added ‘agreement’ to the purpose of the study.

tendinosis?
The in/exclusion criteria have been described in more detail. The role of the biceps(origine) in
rotator cuff disease is still a matter of debate, and is therefore not an exclusion. Anterior labral
tears are excluded by instability complaints.

4. Page 8 and 9: Discussion: The first sentence; See remarks made above (1).
We have rewritten the discussion.
**Discretionary Revisions**

1. Background: Second passage; Consider to use evaluator bias instead of surgeon bias.  
   We have adjusted this accordingly.

2. Page 3 Background; Consider to use the more common name Oxford Shoulder Score(s),  
   and to include the Oxford Instability Shoulder Score (also named Shoulder Instability  
   Questionnaire)  
   As advised by reviewer 1 we have shortened this passage.

3. Page 3 line 8 and 9 "We choose to translate and validate the WORC, as it was developed by  
   and for people with rotator cuff disease..." Consider to drop "by" as it was not developed by  
   the patients...  
   Thank you for pointing this out. We removed this.