Author's response to reviews

**Title:** The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study.

**Authors:**

Thomas J Hoogeboom (t.hoogeboom@maartenskliniek.nl)
Rob A de Bie (ra.debie@maastrichtuniversity.nl)
Alfons A den Broeder (a.denbroeder@maartenskliniek.nl)
Cornelia H.M. van den Ende (e.vandenende@maartenskliniek.nl)

**Version:** 4  **Date:** 25 April 2012

**Author's response to reviews:** see over
All authors would like to thank both reviewers for their useful and insightful comments on our manuscript. We are convinced the manuscript has improved tremendously. Thank you.

Comments Reviewer 1:

Major compulsory revisions:

Comment 1:
On page 9 authors state they chose to use the MDC 90 as measure on minimal detectable change and they refer to Donaghue and Stokes. Although the latter report that 90% MDC might be acceptable when making decisions on effectiveness of interventions, they do not add a solid foundation for this statement. I’m not convinced 90% is better and believe the authors aren’t either since in the result section the report on the MDC 90 and the 95% limits of agreement. Please elaborate why 90% CI would be appropriate when making decisions on effectiveness and formulate the result section accordingly or change the method section and report that both 90% MDC and 95% MDC will be reported in the results.

Author response:
We have added both the MDC90 and MDC95 to our manuscript.

----------------------------------------

Comment 2:
Authors chose to report on responsiveness as the capability to distinguish between improved and not improved patients on the basis of the ROC curves. In doing so it is difficult to interpret the difference between true change and measurement error as measured by the smallest detectable difference. I suggest authors add a distribution based method (responsiveness ratio).

Author response:
We have added the Guyatt Responsiveness Ratio to the manuscript.

-----------------------------------------

Comment 3:
Authors report that 401 patients returned the baseline questionnaire. They do not report on loss to follow up which seems to be 39% (table 1, follow up 246 participants). Please elucidate on the figures.

Author response:
We aimed to include 100 participants to assess the research question regarding test-rest [Pua et al 2009] and another 100 participants to assess our research question regarding responsiveness. To do so, we contacted 121 and 125 participants at random from the initial 401, with satisfying response ratios (>80%). Unfortunately, this has been erroneously reported in Table 1, in which we described the contacted sample and not the sample that responded. We checked all other figures of Table 1. Now, we have altered Table 1 by describing the test-retest and responsiveness sample independently instead of one follow-up group. Moreover, we have stated the figures more prominently in the results section.

-----------------------------------------

Comment 4:
I would expect responsiveness to be calculated on 246 subjects responding at follow up. I miss the point of reporting 246 participants at follow up (table 1) since 108 and 106 participants responded in the test retest reliability study and responsiveness study respectively. Please explain why responsiveness was calculated on 108 subjects where information on 246 patients seems to be available

Author response:
Apparently we were unclear in our data presentation. As mentioned in the previous response, we contacted 246 to include 100 participants for our test-retest analysis and 100 participants for our responsiveness analysis. So, individuals who received the test-retest questionnaire did not receive the responsiveness questionnaire.
Minor essential revisions:

Comment 5:
Page 7 concerning the SF-36; it is widely used, reliable, validated into Dutch and is easy to complete. Higher scores indicate better health. Please add references. Please add information on psychometric properties of the HADS. Page 7 Please add information on psychometric properties of the HADS, if available.

Author response:
We have added references to justify our statements concerning the psychometric properties of the SF-36 up. Moreover, we have added information on the validity of the HADS and also provided references.

Comment 6:
Authors suggest that the HOOS and KOOS include the full WOMAC physical functioning scale. In my opinion WOMAC asks for PF in the last four weeks whilst HOOS and KOOS use one week intervals. Furthermore, the questions on HOOS and KOOS are imbedded in a larger set of questions which might be of influence in patients responses. I'm not convinced authors can draw conclusion on the responsiveness of the WOMAC by extracting the WOMAC PF questions out of the KOOS HOOS questionnaire.

Author response:
We agree with the reviewer that the WOMAC-PF extracted from the HOOS-PF and KOOS-PF is not identical to the original WOMAC questionnaire. In our study, we initially aimed to study the responsiveness of the HOOS-PF and KOOS-PF, unfortunately due to the low number of responders we decided to go with a disease-specific rather than joint-specific measurement instrument. We have stated this clearer in our manuscript (page 10).

Comment 7:
Page 9 to evaluate …and other instruments (see Table 1) Please change table 1 in table 2.

Author response:
Done.

Discretionary revisions

Comment 8:
Page 4 line one of those new measures, the LEFS, showed promise as a competitive alternative to the WOMAC-PF. Please add references to this statement.

Author response:
We have cited the article of Pua et al. who recommends that the LEFS should be at least used together with the WOMAC.

Comment 9:
Table 3 does not seem to add a lot to the author's arguments, consider skipping table 3.

Author response:
We agree that Table 3 in its current form adds little to the manuscript. Following the recommendation of Reviewer 2, we have added more correlation data to this Table thereby providing more information regarding the measured construct.
Comments Reviewer 2:

Major compulsory revisions

Comment 1:
The major concern is about the lack of basic descriptive data – a basic need for every study report. By that, it is impossible to get an overview and a feeling for the validation results and to rate the quality of the results of the whole study. Please add means, standard deviations (or min, max, medians, … if the score distributions are not approximately symmetric) for all instruments and scores used (incl. frequencies of the global health change ratings=transition item) and for all time points of assessments. Consistently to that, the following items 2 and 3 are major criticism:

Author response:
We agree with the reviewer that the results of our analysis in this paper would be more clear in the context of descriptive data. Therefore we have reported descriptive data from the total, reliability and responsiveness sample in Table 1. Data from our global health change rating scale are more clearly presented in the results section, under the header Reliability and Responsiveness.

-----------------------------------------

Comment 2:
Cross-sectional validity: Report the/all correlation data (e.g. in a matrix) for all scores used in one Table (as partly done in Tables 3 and 4). Then, convergent and divergent constructs can be overviewed. Further, there are some ambiguities, for example: HADS: depression or anxiety?, SF-36: other scales?, HADS: negative correlation: more depression -> better function? Please explain!

Author response:
We have expanded Table 4 with additional correlation data. Regarding the SF-36, we presented only data on subscales Bodily Pain and Physical function as we formulated hypotheses for these scales and not for the other subscales (we did not want to unnecessarily burden the participants). For the HADS we used the total score. We have made this clearer in the table. And regarding the negative correlation, the correct interpretation is that more depression is associated with worse function (due to the inverted scales, HADS higher score is worse and LEFS higher score is better).

-----------------------------------------

Comment 3:
Longitudinal validity: Report the “classical” measures of responsiveness, the effect size and the standardized response mean for all scores used. This is the methodology used over decades and by many hundreds of responsiveness studies. Then, sort them by constructs (pain/function/mental health) and level. By that, longitudinal construct validity and (joint-/domain-) specificity can be quantified and overviewed. This is necessary in addition to the ROC method.

Then: Compare the responsiveness within the same constructs by the sensitive, modified Jackknife test; description and one example can be found in (1).

Author response:
We agree with the reviewer that the ROC method alone might not provide the complete picture for the reader to fully understand the questionnaires’ ability to measure change over time. However, we do not think that effect sizes and standardized response means are the most appropriate methods, as these figures are highly dependable on the change that occurred over time (i.e. ES will be greater in a RCT studying an effective treatment, than in an RCT studying an ineffective treatment). In Table X we depicted these effect size figures which are all very low, due to the fact only a small amount of patients reported significant improvements over time.
### Table X. Effect size (ES) and standardized response mean (SRM)

<table>
<thead>
<tr>
<th></th>
<th>Mean change</th>
<th>SD (baseline)</th>
<th>SD (change)</th>
<th>ES</th>
<th>SRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEFS</td>
<td>1.42</td>
<td>14.41</td>
<td>11.25</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>WOMAC-PF</td>
<td>0.19</td>
<td>14.84</td>
<td>8.86</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>SF-36-PF</td>
<td>0.00</td>
<td>22.73</td>
<td>13.68</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Abbreviations: LEFS = Lower Extremity Functional Scale, SF-36-PF = Short Form 36 - Physical Function scale, WOMAC = Western Ontario and McMasters University Osteoarthritis Index.

A method - we believe - more suitable in this particular study is the Guyatt Responsiveness Ratio, which takes the Global Perceived Effect (GPE) anchor into account. We have presented these data along side the ROC data. Please see Table 5 and the results section ‘Responsiveness’. Since we are unknown to the modified Jack Knife test we are unsure whether we can use this test to evaluate differences in Guyatt Responsiveness Ratio.

---

**Comment 4:**
You cite the fundamental text book for establishing and testing instruments, the “Health measurement scales” of Streiner and Norman by the old, obsolete 3rd version/edition (Ref 24). Cite and use the current 4th version of 2008 (2). For example, the concepts of items 2 and 3 above are well described in that book.

**Author response:**
We were unaware that a newer version was available. We have updated the reference.

---

**Comment 5:**
Validity testing: The use of the “hypotheses concept” method (Table 2). You used the method as described by the COSMIN group. I understand that because the authors are your Dutch colleagues. You can do that in addition to the method of items 2 and 3 above, but do also the “classical” methodology.

5.1. Discuss shortly the advantages/disadvantages of the COSMIN method and cite the reference (3). Some problems (only some of many examples) are inherent in the COSMIN methodology:

**Author response:**
We have referred to the article by Angst et al (2011) regarding the responsiveness analyses in our manuscript.

---

5.2. You did not establish 16 hypotheses but each 8 hypotheses for knee and hip. Two versions of the same hypothesis for hip and knee have another dependence to each other than hypothesis 1 to hypothesis 2 etc.

**Author response:**
Agreed, we have made this clearer in the method section.

---

5.3. The number of 16 hypotheses is arbitrary, why not less, why not more?

**Author response:**
We agree that the number of hypotheses we tested is arbitrary and that maybe more hypotheses should to be preferred. We think that we included the most relevant hypotheses.
5.4. Hypothesis: correlation is low <0.50 (true/false): There is a difference if it is 0.49 or 0.01, Discuss!

**Author response:**
This discussion is inherent to the use of cut-off values, but appears to be outside of the scope of this article. An alternative could be to use the lower bound or upper bound of the 95% Confidence Interval of an association. However, this method would not be valid in this study as our hypotheses were formulated *a-priori*.

5.5. Dependency of educational level / socio-demographic status. There is a huge number of epidemiologic literature that demonstrate that health and almost every outcome has a social gradient. Explain and discuss!

**Author response:**
Although we agree with the reviewer, we believe this discussion is outside of the scope of this article. In previous studies the LEFS was unaffected by the level of education. The fact that it was related in our study is, in our opinion, a point lost to the LEFS and should therefore be commented upon in the discussion section.

5.6. Pat. with complaints 5 years or longer... Why 5 years: arbitrary?

**Author response:**
Yes, arbitrary. 5 years is often used as a cut-off for complaint duration in osteoarthritis.

5.7. If 75% of the hypotheses were confirmed: Why 75%: arbitrary?

**Author response:**
This was chosen conform the recommendations of Terwee et al (2007).

Comment 6:
Construct validity of the LEFS for function. Test the uni-dimensionality for function by (for example) factor analysis.

**Author response:**
We have added factor analysis data that confirms the uni-dimensionality for the LEFS, KOOS and HOOS function scales.

Comment 7:
Translation: Stage/step VI of the process is lacking: “sending all versions and the protocols of steps I to V to the developer of the original questionnaire” (Ref 17: Beaton 2000). Explain!

**Author response:**
This step has been executed and the developers of the original questionnaire approved our methods and the translations. We have added this statement to the paper.

Comment 8:
A Cronbach’s alpha of 0.96 means that some items of the scale are redundant. Discuss! See also ref. 27 Bot et al.: a value in the range 0.70-0.90 would be good. See also in Streiner (2).
Author response:
There is a lot of debate on the proper cut-off for internal consistency. Since the Cronbach’s Alpha statistic is affected by both the magnitude of the sample size and the number of items in the questionnaire, the cut-off should not be considered static but dynamic (see for example Ponterotto and Ruckdeschel (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. Perceptual and Motor Skills, 105, 997-1014). They state that in a study such as ours, the Cronbach’s alpha should be at least 0.90 to be considered excellent. We have discussed this in the limitation section of the article.

-----------------------------------------

Comment 9:
Results. Score distribution of the LEFS (p. 11). Analyze and report the characteristics of the score distribution: normal (Gauss), symmetric? This belongs also to validity. In this context: If the scores will not be symmetrically distributed one should rather use Spearman correlations that Pearson correlations (Tables 3 and 4).

Author response:
The distribution of the LEFS is symmetrical (added to the results section, page 12). We have checked all distributions of the other questionnaires and calculated spearman correlations for the Quality of Life and Sport/Rec subscales of the KOOS/HOOS and for the HADS total score (see Table 4). This had no impact on the interpretation of the hypotheses.

Minor essential revisions

Comment 10:
Abstract: Match the aims with the title. You did much more than only discriminant validity.

Author response:
We have matched the aims in the abstract with our title by stating the psychometric properties more explicitly.

-----------------------------------------

Comment 11:
Introduction: As you described, the WOMAC is the most often used tool for the leg. State why you did not use the WOMAC (License problems, prohibited for validation studies….).

Author response:
We never planned to study the WOMAC as its inability to discriminate between pain and functioning has been established in numerous studies. However, this has not been studied yet in the HOOS and KOOS. We have rewritten the introduction section to make this clearer. Moreover, we have added a statement to the method section responsiveness why we decided to use the WOMAC-PF based on the HOOS AND KOOS (namely due to a power issue).

-----------------------------------------

Comment 12:
Patients, inclusion/exclusion, p.6: State explicitly that you only included patients with osteoarthritis.

Author response:
We have stated this more explicitly.

-----------------------------------------

Comment 13:
Measures. Report the “missing rules”, i.e., how many completed items are necessary to determine the scale for all instruments, especially for the LEFS.
Author response:
We have added the “missing rules” for the LEFS. Reporting them for each of the questionnaires would not benefit the paper’s readability in our opinion.

Comment 14:
Methods, discriminat validity, p. 9: Cite a reference for the Meng’s test to compare two correlations or list the formula. Moreover, is this comparison of correlations a established method? Give a reasoning and cite a reference for that.

Author response:
The Meng’s test is an established method to test two correlations and has been used before to test the discriminate validity of the LEFS and WOMAC (for example, see Pua et al 2009). We have added the reference to the paper.

Comment 15:
Methods, Validity (p. 9). Citing of the reference 27 on p. 9. This paper of Bot et al. is a review of shoulder instrument studies. This seems not be appropriate to explain the 75% proportion of positive hypotheses. Explain of look for that rather in (3)!

Author response:
We have replaced the erroneous reference with the reference proposed by the Reviewer and by the reference that we meant to cite namely Terwee et al Journal Clinical Epidemiology 2007.

Comment 16:
Results, Patient selection (p. 11). 401 patients at baseline, only 120 at follow-up: How were they selected? Randomly? Is there possible bias? Explain!

Author response:
Yes, we aimed to contact a random sample of 120 participants for the reliability analysis and another 120 participants for the responsiveness analysis; eventually we ended up contacting a total of 246 people. We don’t expect any bias.

Comment 17:
Reliability. You reported the MDC90% but the 95%-CI for the Bland-Altman plot. To compare, report also the MDC95%.

Author response:
We have added the MDC95% data to our manuscript.

Comment 18:
Reference 2. For the WOMAC, cite the WOMAC manual (4) not the preliminary study Ref 2.

Author response:
We have adjusted the reference.

Comment 19:
Discussion, p. 16. “Convergent validity should be assessed in future studies.” Yes, but you also did that by the correlation data of Tables 3 and 4 and by the responsiveness data that will given (item 3): the LEFS shows convergent validity to H/KOOS function and SF-36 PF. Adapt the discussion!
Author response:
This recommendation was based on the lack of performance measures in our study. We have changed the last sentence of this paragraph to make this clearer.