Reviewer's report

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

Version: 4  Date: 16 May 2012

Reviewer: Felix Angst

Reviewer's report:

Positive criticism
Overall, the authors have done a good job in responding to the comments of both reviewers. The article has been markedly improved.

However, there are still some, partly important points to correct and improve.

General principle: Make your corrections in the text of the paper, not only in your response: You do not have to do the corrections for me but for the future (eventually inexperienced) reader of your paper!

Major compulsory revisions

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 to R1:
Has been corrected very well. Table 1 is very informative. Still to do / to satisfy the legend (“n/% unless indicated differently”): Indicate that you reported mean (sd) of the scores.
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 to R1:
Table 4 is well done now. Still to do:
1. explain/interpret the negative correlation to the HADS.
2. The use of the HADS total score is unusual and not described in the HADS manual – state and explain why you did that.
3. Make all your corrections in the text of the paper.

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 Jacknife 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.

Comment to R1:
Thank you for partly addressing this issue. You stated an interesting point, that of the problem if the observed changes are small. Again: Discuss that in the paper. However:

1. This problem affects every responsiveness testing and responsiveness parameter, also that of Guyatt.

2. It is irrelevant because comparison of responsiveness is inter-individual, not inter-individual. Thus the problem remains the same for every instrument/scale and responsiveness parameter: Every responsiveness parameter has the score change in the numerator only the denominator is different (ES/SRM/Guyatt).

3. The responsiveness parameter of Guyatt is OK but has a disadvantage: It is not very often used because the standard deviation of a “stable” period is rarely available. Therefore, your results cannot be compared to those of many other studies using SRM or ES.

4. Table X is very informative, thank you. It shows that the LEFS is most responsive. This is consistent to your results obtained by Guyatt’s parameter. I would report both, Guyatt and SRM in Table 5.


4. You cite the fundamental textbook 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 to R1: OK.

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.

Comment to R1: Yes, you did that but:
1. at the wrong location: do it in the context of (cross-sectional) validity where you
used your 2x8 hypotheses not in that of responsiveness. And:
2. Again, discuss the advantages/disadvantages of the COSMIN method. You
didn’t do that yet.

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.

Comment to R1: OK.

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.

Comment to R1: OK, but do/explain that in the paper.

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.
Comment to R1: Partly OK. You could treat that in the correction to 5.1.

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.
Comment to R1: OK. It is well treated in the discussion, limitations.

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.
Comment to R1: OK, but do/explain that in the paper.

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 to R1: OK, but do/explain that in the paper.

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 to R1: Very informative, very well done.

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 to R1: OK. Did the developer also approve your final LEFS version?

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 to R1: OK, well done.

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.

Comment to R1:
OK. Explain how you checked symmetry! Comparison Spearman/Pearson: very good!

Minor essential revisions

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 to R1: OK – do not underestimate yourself!

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 to R1:
OK but what I meant is not a criticism to you but to the developer of the WOMAC, namely Bellamy N. The WOMAC is not available for free, its use has to be paid very costly. And Bellamy prohibits the use of the WOMAC to validate other instruments. I feel this is an issue that has to be mentioned by 1-2 sentences.

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

Author response:
We have stated this more explicitly.

Comment to R1:
OK, but I needed long time to find that in the limitations on p. 16. It should be
stated in the selection of patients on p.6.

13.-15. ........

Comment to R1: All comments are well responded.

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 to R1:
Partly OK. Please explain how you selected the partial samples. Randomly? By consecutive admission? Explain in the text why you do not expect any bias.

17.-19. ...........

Comment to R1: All comments are well responded.