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: 5 Date: 23 May 2012

Author's response to reviews: see over
Response to reviewer 2’s remarks:

We would like to thank Felix Angst again for his constructive comments. We really believe the manuscript’s quality improved yet again.

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

Reviewers comment 1:
Context
Original reviewer’s query: 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:

Original authors’ 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.

Secondary reviewer’s query: 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.

Authors’ response to comment 1:
We have adjusted the legend of Table 1.

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

Reviewers comment 2:
Context
Original reviewer’s query: 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!

Original authors’ 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).

**Secondary reviewer’s query:**
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.

**Authors’ response to comment 2:**
1. We have added a statement about the negative correlation between the LEFS and HADS in Table 4.
2. We now provide data on the HADS anxiety and HADS depression separately. We agree with the reviewer that this is more common.
3. All corrections were made in the paper.

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

**Reviewer’s comment 3:**

**Context**

**Original reviewer’s query:** 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.

**Secondary reviewer’s query:**
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 you results obtained by Guyatt’s parameter. I would report both, Guyatt and
SRM in Table 5.

5. The modified Jacknife test is also valid for Guyatt’s parameter and described in: Angst F,
Goldhahn J, Drerup S, Aeschlimann A, Schweyzer HK, Simmen BR. Responsiveness of six
outcome assessment instruments in total shoulder arthroplasty. Arthritis Rheum

Authors’ response to comment 3:
In line with the reviewer’s request, we have elaborated further on the small observed changes in
our discussion section (Page 15). Moreover:
1. we have added the SRM to the manuscript (see Table 5, the method section (page 10) and
result section (page 13), as we agree with the reviewer that the inter-instrumental is more
important than the inter-individual comparison.
2. therefore we also added the modified Jackknife testing to the manuscript (see the method
section (page 10) and result section (page 13).
3. the article provided by the reviewer unfortunately did not describe a method for performing a
modified Jackknife testing for the GRI-statistic. In our opinion such a test would not be
feasible as the GRI-statistic itself is build-up from data from two different groups; GPE=1-2
and GPE=3-5.

Reviewer’s comment 4:
Context
Original reviewer’s query: 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.

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

Secondary reviewer’s query:
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.
Authors’ response to comment 4:
We have added an additional paragraph to our limitation section of our discussion section (see page 16-17). In this paragraph we describe the limitation of using cut-off values in hypotheses testing. We don’t think a full discussion of the advantages and disadvantages of the COSMIN would be of interested to the readers of the paper.

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

Reviewer’s comment 5:
Context
Original reviewer’s query: The number of 16 hypotheses is arbitrary, why not less, why not more?

Original authors’ 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.

Secondary reviewer’s query:
OK, but do/explain that in the paper.

Authors’ response to comment 5:
We have added the word ‘arbitrarily’ to the method section, see page 9.

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

Reviewer’s comment 6:
Context
Original reviewer’s query: Hypothesis: correlation is low <0.50 (true/false): There is a difference if it is 0.49 or 0.01, Discuss!

Original authors’ 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.

Secondary reviewer’s query:
Partly OK. You could treat that in the correction to 5.1 (Added by author: Please see reviewer’s comment 4).

Authors’ response to comment 6:
Please see Reviewer’s comment 4. We have combined, as recommended by the reviewer, comment 4 and comment 6.

------------------------------------
Reviewer’s comment 7:
Context
Original reviewer’s query: Pat. with complaints 5 years or longer… Why 5 years: arbitrary?

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

Secondary reviewer’s query:
Comment to R1: OK, but do/explain that in the paper.

Authors’ response to comment 7:
We have added this to the legend of Table 1.

Reviewer’s comment 8:
Context
Original reviewer’s query: If 75% of the hypotheses were confirmed: Why 75%: arbitrary?

Original authors’ response: This was chosen conform the recommendations of Terwee et al (2007).

Secondary reviewer’s query:
OK, but do/explain that in the paper.

Authors’ response to comment 8:
The article of Terwee et al (2007) is cited after the statement that 75% of the hypotheses must be full-filled for the construct validity to be considered good.

Reviewer’s comment 9:
Context
Original reviewer’s query: 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!

Original 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.

Secondary reviewer’s query:
OK. Did the developer also approve your final LEFS version?

Authors’ response to comment 9:
Yes they did, we have added this to the manuscript (page 5).
Reviewer’s comment 10:
Context
Original reviewer’s query: 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).

Original authors’ 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.

Secondary reviewer’s query: OK. Explain how you checked symmetry!

Authors’ response to comment 10: Symmetry was assessed by the visual inspection of data distribution as depicted in histograms. We have added this to the method section (see page 8).

Reviewer’s comment 11:
Context
Original reviewer’s query: 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….).

Original authors’ 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).

Secondary reviewer’s query: 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.

Authors’ response to comment 11:
The reviewer makes a very good point and we too feel this is an issue that hinders clinical research on osteoarthritis. However, we never intended to use the WOMAC to validate the LEFS-dlv. Therefore we believe this statement would come out of the blue and will add little to the article itself. We did add license statements to the introduction.

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

**Reviewer’s comment 12:**

**Context**

**Original reviewer’s query:** Patients, inclusion/exclusion, p.6: State explicitly that you only included patients with osteoarthritis.

**Original authors’ response:** We have stated this more explicitly.

**Secondary reviewer’s query:**

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.

**Authors’ response to comment 12:**

We have stated this more clearly in the method section (see page 6).

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

**Reviewer’s comment 13:**

**Context**

**Original reviewer’s query:** Results, Patient selection (p. 11). 401 patients at baseline, only 120 at follow-up: How were they selected? Randomly? Is there possible bias? Explain!

**Original 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.

**Secondary reviewer’s query:**

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.

**Authors’ response to comment 13:**

We have described the utilized procedure into more detail in our manuscript. Please see page 6.

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