Author's response to reviews

Title: Measurement properties of the Western Ontario Shoulder Instability Index in Dutch patients with shoulder instability

Authors:

Just A. van der Linde (javdlinde@gmail.com)
W. Jaap Willems (w.j.willems@xs4all.nl)
Derk A. van Kampen (derkvankampen@gmail.com)
Loes W.A.H. van Beers (L.W.A.H.vanBeers@olvg.nl)
Derek F.P. van Deurzen (D.F.P.vanDeurzen@olvg.nl)
Caroline B. Terwee (cb.terwee@vumc.nl)

Version: 5 Date: 10 April 2014

Author's response to reviews: see over
Dear Editor,

I would like to re-submit our manuscript ‘Measurement properties of the Western Ontario Shoulder Instability Index in Dutch patients with shoulder instability’.

Enclosed we present a point-by-point response to our reviewers comments, including the adjustments to our manuscript. To improve the style of written English, our manuscript was also edited by Edanz Editing service, as you have suggested.

We hope that our manuscript will now meet your standards for publication in the BMC Muskuloskeletal.

Please let me know if any other adjustments are needed.

Sincerely,

on behalf of the co-authors,

Just van der Linde

REVIEWER 1.

Version: 4Date:22 December 2013
Reviewer: Björn Salomonsson

Reviewer's report:
This is an interesting study on shoulder instability scores. However to my surprise I received 2 almost identical papers for review the same week from different journals, where only the main focus on WOSI or OSIS scores was exchanged.

There are several good reasons for presenting two separate works on two separate scores, but in this case I will start with the common questions for both of the works and then go for the specific questions of each study separately.

To conclude my impression that this is 2 parts of one study:
It is the same authors.
It is the same ethical decision number.
It is identical patients description.
It is identical methods.
It mentions the other part of the study in the discussion text.

We decided to submit both manuscripts to different journals for the following reasons.
We wanted to present comprehensive papers. Validation of a PROM includes many data (hypotheses regarding correlations, ICC values, floor- and ceiling effects, etc), which should be compared to the original article as well as to consecutive validation studies (i.e. 6 articles for the WOSI validation alone).

To merge these data into one manuscript would, in our opinion, be at the expense of its transparency.

To enable rapid validation, we aimed to submit the first manuscript as soon as possible (i.e. September 20th 2013 for the OSIS), without awaiting the completion of the WOSI manuscript. To avoid confusion we have submitted the second manuscript to a different journal, after completion 3 months later (i.e. December 6th 2013 for the WOSI).

Unfortunately no gold standard or other validated PROM for shoulder instability is available to assess construct validity. Therefore, we chose to validate both the WOSI and OSIS simultaneously, and to use the preliminary results of both to test validity of the other. We added a comment about this issue in the discussion section on page 20.

I apologize for the confusion it has caused, and I understand your concerns. We nevertheless still believe that 2 manuscripts are justified.

Regarding both studies I have the same major points:

Score validation and other statistical analyses may not be considered as common knowledge in the large population of orthopaedics that would be interested as readers of these manuscripts. You introduce rather complex analyses with very little explanation on why and how it should be used, the references are not that much easier to the average reader like me.
If the less common analyses could be explained in a more practical way to the readers I believe the manuscripts would gain a lot in usefulness, as well as spread the knowledge regarding the score validation process.
Thank you, we have revised the measurement properties section.

A) Major Compulsory Revisions
 Mixing the paper version score with a digital version must be validated!
This mix of formats is done both for WOSI and OSIS.

To my knowledge these different formats first must be validated to each other to be used as interchangeable units? Just like any other translation in language must be validated, the translation in formats also need validation just as much.
There is quite some evidence in the literature that mode of administration has very little effect on scores and on the measurement properties of the instruments (see for example some references below). Therefore, we don’t think mixing the versions caused a problem and we don’t think there is a need for additional validation.


How was the different cohorts distributed? First paper format then change to digital in the second half, or paper format at first visit and digital in re-test situation? Are any of the other scores also mixing paper and digital versions?

Added to Results Patients section: ‘Sixteen patients completed their first questionnaire on paper followed by a web-based retest, 4 patients completed both questionnaires on paper.’

In description of the WOSI digital version you write that it was a 0-10 scale instead of a 0 mm-100 mm VAS scale, and that is also a rather special change of the format of the WOSI score. Were there no intermediate values or just the 0-10 as whole steps?

This is a valid point of discussion. We added to Discussion section: ‘Another point of discussion is the 10 mm interval on the VAS that was used in our WOSI version, compared to the 100mm interval on the original version. As our paper version was printed copy from the digital version, this was consistent throughout our studies.’

Would it be possible to validate the paper forms to the digital format of the scores as well within these studies, then that would be very valuable to the validity of the study.

As mentioned above.

B) Minor Essential Revisions

Floor and Ceiling effects, do you measure in the score or in the patients!

In both texts you discuss the floor and ceiling of the patients (in discussion it is obvious) and say that your SDC limits for F/C: “measurement error cannot be detected in these patients” and “detection of improvement and deteriorations that are distinct from measurement errors at follow-up”.

That indicates to me that you measure F/C in the patients, and not in the score validation?

Floor and ceiling effects represent the amount of patients that have the highest or lowest possible score. This reflects the performance of the instrument in this population (so a combination of score and patients). We tried to improve the explanation of floor and ceiling
To me measurement errors of the score would be e.g. by the ruler measuring mm on the VAS scale in WOSI. I think that in validation of the score I would choose to have F/C of the absolute result of the score. But the clinical F/C by the way you have measured is very valuable in the clinical study, and should ideally be done in every study to check the actual cohort as well as responsiveness?

However, you exemplify the SDC as measuring for change at follow-up, but you did not present any change/follow-up, and no Effect size or SRM for responsiveness. Is your calculation of the SDC comparable to an actual change at follow up to discuss detection of improvement or deterioration. To me this is confusing because it is two different thing in one analysis, the score itself, and the SDC in the patients, and difficult to understand how it should be interpreted. The SDC is the measurement error of a scale. It can be used to interpret change scores in longitudinal studies, assessing change over time. Only changes larger than the SDC can be considered real changes (and not measurement error). In this study, however, we did not intend to measure change over time (the follow-up period was short to measure reliability), and therefore did not apply the SDC, we only calculated it. We tried to explain this better on page 12.

C) Minor Essential Revisions
SDC
Please explain in more detail the rational for using SDC as Floor and Ceiling effects.
We postulated that for patients who don’t score the maximum or minimum score, but do score close to one of the extremes (within the SDC range), a ‘real change’ (beyond measurement error) can not be detected.

E.g. a fictional score ranges from 0-10 points and has a SDC of 2 points. Patients that score 9 points at baseline, would not be able to show a real improvement (>2 points). These patients are thus at their ceiling.
We calculated these percentages of the SDC-range as a supplement, not as a substitute for the F/C effects.
We clarified this in Methods section ‘Floor and Ceiling effects’ final sentences, on page 14.

Your SDC results in about 20 % of the upper and lower scores as F/C?
Do you really intend to propose that a score of 80 % is clinically the same as 95% ? As the SDC was calculated 23% (expressed within the WOSI score 0-100%), patients indeed have to increase at least 23% to ensure a ‘real change’.

No, 80% and 95% are not the same, however only in patients who score less than 77% at baseline, a ‘real change’ can be observed at follow up. Both in patients who score e.g. 80% or 95% might be regarded as having a ceiling effect (a ‘real change’ of 23% is not possible).

I agree that it might be possible for two individuals to have every other measure
equal and still score 80 % or 95 % in WOSI (or similar in OSIS). But it is difficult to see that if it is the means in two groups, then the difference would be substantial. I think this could be explained how you use the SDC to the readers to better understand the analysis as individual measures or in groups.

It should be noted that the SDC refers to the measurement error in one change score in one individual patient. When measuring change in a group of patients (as in a study) the measurement error of the mean change score is much lower (in fact SDC / √n). We added this to the discussion section on page 19.

Is the SDC homogenous through the whole scale of the scores?
We know that e.g. Effect sizes vary in different parts of a score/scale. And it may also be more difficult to earn the very last point in a score at the extremities of the scale, than to have change in the central part. How does the SDC behave regarding F/C effects?
When the SDC is calculated from the ICC it is assumed that the SDC is constant over the scale. However, it is likely that the measurement error is larger at the extremes of a scale.

You also discuss change measurements with regard to SDC?
Perhaps you could be more specific on that you have not had any change to measure in the scores (no follow-up after treatment) and (if I understand correctly) that this change is only a calculation/estimation from one measure at one time, without any change seen in the patients.
And how this is useful instead of a measured change.

See our reply to previous comments

D) Discretionary Revisions
CFA
In the manuscripts you conclude that the CFA was unreliable och not useful if I understand correctly?
Please inform on how it could have been useful if it would have a better fit?
The expected 4-factor model did not fit well. This means that the classification of the items into 4 subscales cannot be justified by our results. We added some comments on the WOSI structure in the discussion, see page 18.

I suggest that the analysis description could be shortened and just confirm that it was not very useful, but that you inform on the conclusions that this led to (that score changes are individual and that the total WOSI score is recommended instead of domains).
This section was shortened to:
‘The expected 4-factor model did not fit well: CFI 0.869, TLI 0.850 and RMSEA 0.104. Subsequently, a 3-factor, 2-factor and 1-factor model were tested in exploratory factor analyses (Table 3). Best interpretable results were found with only 1 factor, however confirmative testing of this 1-factor model in CFA showed worse fit (CFI 0.800, TLI 0.778 and RMSEA 0.127) compared to the original 4-factor model’ page 15-16.
We adjusted the discussion on page 18.

The full CFA analysis description and tables could be presented in a digital supplement instead of in the manuscript, with the rather large tables as well.

Ok.

E) Major Compulsory Revisions
Not using the full information of the data you really possess?
In both the manuscripts you say that the WOSI/OSIS score would need to be compared further but that the other score is not yet fully validated and published. This may be true but it is a bit unfair to the readers when you publish this in parallel manuscripts, and you do propose this as limitations to your studies and that would like to propose which of the scores that have the superior measurement properties?
It would not be impossible to do that in at least one of the manuscripts?
The reason to submit both validation studies separately was argued above.

We agree that it may seem be a bit unfair to validate the WOSI against the OSIS in one manuscript and the OSIS against the WOSI in another manuscript. Unfortunately there is no gold standard or other validated PROM for shoulder instability that could be used for construct validity. Therefore, we chose this method but also included other instruments. We added a comment about this issue in the discussion section on page 20.

And if you have a possibility to measure The effect size or SRM it would be advantageous as well.
It does seem likely that it would come available (from the same study cohort?): J Orthop Surg Res. 2013 Nov 14;8(1):40. [Epub ahead of print]
Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs).
van Kampen DA, Willems WJ, van Beers LW, Castelein RM, Scholtes VA, Terwee CB.
Correct, one of our future studies focuses on minimal important change (MIC), reflecting clinical improvement as a result of therapy. We plan to combine the MIC for both the WOSI and OSIS in one manuscript. We will also address whether there is a difference between the MIC when assessed in a homogeneous shoulder instability population versus, assessed in a heterogeneous patient population with other shoulder complaints as well (stated in the final sentence of the Discussion section). Data collection for this study is still running.

Other points: Discretionary Revisions
WOSI is Canadian, not from USA to my knowledge (PROM/WOSI descriptions) adjusted in text.

Chonbachs alfa is related to the number of questions as well, the more questions the higher correlation, that might be mentioned?
Correct, this explains the high Cronbach’s alpha of the total WOSI score. This was added to the discussion on page 19.
WOSI manuscript: Minor Essential Revisions

Patients part: Copy-Paste error from other manuscript, OSIS should be WOSI. adjusted in text.

Results part:
Reliability: Values does not match, WOSI retest score should be 959? As in table 4 adjusted in text.

SEM and SDC (abstract and results part) of SEM in WOSI is % not points I think, that would be easier to understand in the text. adjusted in text.

Discretionary Revisions

Table 4:
The difference in test/re-test values for the 3 last domains seems very strange compared to the total score and the change? And the ICC does not seem to be affected as well.
Were there really those differences in the domains without worsening the ICC? Revised

Table 5:
See C/F and SDC above.
Your C/F limits is half of the score scale, it must be explained better how this is interpreted. Adjusted as mentioned earlier.

References OK

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Not suitable for publication unless extensively edited

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests

REVIEWER 2.

Reviewer's report

Version:4 Date:27 February 2014

Reviewer: Stefan Buchmann
Reviewer's report:

Overall Purpose:
The presented study handles an interesting topic with a well performed methodology. Unfortunately line numbers are missing which makes the review rather difficult.

Abstract:
No correction or comments.

Introduction:
Well written introduction with a clear concept.

Materials and Methods:
Page 8: About what kind of instability do you talk (ant.inf. traumatic? Multidirectional? Also subluxations) – Please add definition for inclusion criteria!
Added:
‘Inclusion criteria were age 16 years or older and any form of glenohumeral instability (anterior, posterior, multidirectional) as diagnosed by one of our doctors.’

Page 8: ”’Exclusion criteria…” – please define glenoid or proximal humeral fracture. Is a bony bankart already a glenoid fracture – what about large Hill Sachs (humeral fracture?).
Revised:
‘Exclusion criteria were an inability to master the Dutch language and a large glenoid fracture or proximal humeral fracture, such as a displaced fracture of the greater tuberosity. Hill-Sachs and bony Bankart lesions were included.’

Page 10: Why didn’t you also use the self assessment Rowe score? Please also add an explanation in “Discussion”.
The self assessment Rowe score is published once (in German by Kupsch et al PMID 17160397). The authors have decided to use internationally adopted PROMS that are used more frequently in literature.

Discussion, revised:
‘The original Rowe and Constant scores are not PROMs but observer-based measurement instruments, and the Constant score is considered not applicable to shoulder instability [50,51]. We used only PROMs for the Dutch validation.’

Page 11: (OSS):
Please shorten paragraph : “In the original scoring system ….” like according to the OSIS the scoring system was revised to 0 to 48 points.
This part was removed, scoring ranges were added to table 2.

Page 14:
Did you ask the patient if within the time interval from first questionnaire anything happened (redislocation, subluxation) – this might change the whole scoring.
Especially when in about 10% the last dislocation was less than a month ago there might be a few incidents in this time interval.
That is a very interesting point. We have added to results section:
‘Table 2 ..... As measured with the OSIS, OSS, SST, and DASH evaluations, there was no significant change in shoulder function at baseline and retesting.’

Results:
Well written.

Page 17: All patients suffered anterior dislocations – so this was an inclusion criteria?
This was not an inclusion criteria, however no other directional instability was observed in either one of the cohorts.

Discussion:
Page 22: Please add numbers for references (Kirkley, Hatta, Salomonsson, Hoffstaetter, …)
Done.

Page 22, last line: we decided to use the OSIS….! There is no argumentation why. Please add.
Revised, line 430:
‘Because the SST and OSS are validated in Dutch, and because preliminary results of the Dutch OSIS validation are good, we decided to use these instruments instead of the UCLA shoulder rating scale and ASES. It should be noted, however, that the WOSI is validated against the OSIS, and the OSIS is validated against the WOSI. Unfortunately, there is no gold standard or other validated PROM for shoulder instability that could be used to assess construct validity. Therefore, we chose this method but also included other instruments. The high correlation between WOSI and OSIS (0.82) means that the two questionnaires are measuring the same construct, but it does not guarantee that both instruments are valid.’

Page 24: The homogeneity of the group is to discuss – one positive point is only ant.inf. dislocations but also early instability osteoarthritis has to be discussed in the patients with longer than 2 years dislocation ago. For these DASH or OSS might be more suitable (pain vs instability).
‘Homogeneous’ was removed from the discussion.

Overall conclusion:
The presented study is well designed and the paper well written. It additionally adds knowledge to current literature so that I would consider the article for publication after the above mentioned revisions.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests