PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

<table>
<thead>
<tr>
<th>TITLE (PROVISIONAL)</th>
<th>The American Orthopaedic Foot and Ankle Society (AOFAS) Ankle-Hindfoot Score; Study Protocol for the Translation and Validation of the Dutch Language Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHORS</td>
<td>Van Lieshout, Esther M.M.; De Boer, A.; Meuffels, Duncan; Den Hoed, P.; Van der Vlies, Cornelis; Tuinebreijer, Wim; Verhofstad, Michael</td>
</tr>
</tbody>
</table>

VERSION 1 - REVIEW

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Taucha Inrig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>St. Michael's Hospital, Canada</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>27-Jun-2016</td>
</tr>
</tbody>
</table>

GENERAL COMMENTS

Review:
The American Orthopaedic Foot and Ankle Society (AOFAS) Ankle-Hindfoot Score; Study Protocol for the Translation and Validation of the Dutch Language Version

Abstract:
BMJ Open Author’s Guidelines direct authors to conform to the following headings in their structured abstracts “Abstract: this should be structured with the following sections. Introduction; Methods and analysis; Ethics and dissemination. Registration details should be included as a final section, if appropriate”. The current protocol does not conform to these heading. Ethics and Dissemination are missing and the inclusion of Discussion, Strengths and Limitations of this study are not usual practices for a scientific abstract. There are also conflicting statements as to the future use of the instrument to be created. In the Introduction, the authors state they wish to use the AOFAS-DLV to compare surgical results across hospitals, but in the Discussion there is a suggestion that the AOFAS-DLV may be used to monitor individual patient’s results. Both are valid goals for an instrument, but there are different statistical thresholds to be considered if one intends use for individual patients (further references to this will be included in the statistical review below).

Study Design: Protocol authors hold the AOFAS in high regard as being the most widely used instrument for ankle outcomes. While it is true that the AOFAS has been widely used, the AOFAS’ own position statement of 2011 (Pinser E, Daniels TR. AOFAS Position Statement Regarding the Future of the AOFAS Clinical Rating Systems. Foot and Ankle International 32 (9), 841-842.) does not. Specifically, Pinser and Daniels state that the AOFAS measures have not been found to be valid or reliable (contrary to statements made by this protocol’s authors) and do not recommend further use of the AOFAS. This is a serious design flaw in the protocol that
needs to be addressed. Protocol authors also misrepresent the findings of SooHoo et. al., 2003 (who report poor construct validity); and Westphal et. al., 2004 (who report descriptive results using AOFAS and SF-36 with low to moderate correlations between the scales but validation of the tools against each other is not a goal of their paper).

As far as the cultural adaptation measures taken to create the Dutch Language AOFAS, the authors propose employing the best standards available and would be rated favourably by critical appraisal tools such as COSMIN if they conducted the study by their proposed protocol.

Protocol authors suggest that the physician portion of the AOFAS score may be completed by a research physician or a research assistant. This is of concern as there are no details on how the research assistant will be trained to perform this task. Is it appropriate for a non-physician (possibly a person with no medical training) to be performing these assessments? Will inter-rater reliability be considered if conducted in this manner to prove use of research assistant is appropriate? Further description of how the AOFAS-subjective items will be administered is also necessary. Will the physician read those questions to the patient/will the patient be handed the survey to fill out for themselves/will the physician infer the responses according to a conversation with the patient?

Research Ethics
Protocol authors appropriately describe REB/IRB approvals and consent procedures.

Outcome Measures
Protocol authors state that the construct validity of the AOFAS-DLV will be the primary outcome measure (page 10/31, lines 36-43). Measurement properties cannot be the outcome measure; the AOFAS-DLV itself is the outcome measure, whilst demonstration of the measurement properties are the goal of the proposed study. Construct validity concerns whether an instrument measures what we think it measures. Authors correctly define construct validity later on page 16, and state that the SF-36 and FFI will be used to triangulate construct validity, but do not state their hypothesized correlations between these instruments. It should also be noted that the SF-36 has previously yielded poor correlations in its use in construct validity and the AOFAS (see SooHoo, 2003). The protocol authors did well to chose another region-specific instrument (the FFI) which will likely yield more meaningful results for them.

Statistical Analysis:
Sample size for group 1 is appropriate for testing. Sample size for groups 2 and 3 are not explicitly stated and range from 50-100 minimum. I would encourage the authors to use COSMIN as their guide in this regard and aim for the recruitment of ≥100 participants.

Construct validity: authors do not state their hypothesized correlations between AOFAS and FFI /SF-36. Given that the AOFAS is traditionally reported as one score and the FFI and SF-36 are reported using sub-scale or summary scores, this warrants clarification. Again the literature on the SF-36 has demonstrated poor correlation with the AOFAS and region-specific measures should be more useful.

Reliability (test-retest/reproducibility): the authors would be encouraged to include a “global change question” for the retest portion of the study so that they have an external way to ‘prove’ participants did not change in the 2-3 weeks between survey administrations. Test-retest ICCs should only be performed on those who state their ankle condition has not changed. Additionally authors should state which ICC they intend to run (most are using ICC 2,1 in cases like this).

Reliability (Measurement Error): if the MIC has not been published please reference “who” considers 10-20 points as relevant. [See the Position Statement for discussion of varying interpretations of scores.]

Floor and Ceiling effects: It should be noted that the limited number of response categories in the AOFAS make it highly prone to ceiling effects, and they have been previously reported. See Guyton G.(2001) Theoretical Limitations of the AOFAS Scoring Systems: An Analysis Using the Monte Carlo Modelling. *Foot and Ankle International* 22 (10), 779-787.

Responsiveness: The introduction to this section needs editing for clarity (page 19, lines 16-25). Again, hypotheses are not documented.

Final Remarks: This protocol represents an ambitious attempt to fill many of the psychometric gaps for the measurement properties of the AOFAS. It is important to note that the AOFAS has issued a position statement suggesting the AOFAS Clinical Rating Systems are not recommended for clinical use. The authors need to defend why their proposed study is meaningful in the face of the AOFAS’ position regarding its use. It would also be helpful for the authors to include references of the published psychometric properties of several foot/ankle instruments including the AOFAS and the FFI (e.g. see Naal FD, Impellizzeri FM, and Rippstein PF. (2010)) which are the most frequently used outcome instruments in studies on total ankle arthroplasty? Clin Orthop Relat Res 3(10), 815-826.

---

**REVIEWER**

Wolfgang Schneider, MD
Herz Jesu Krankenhaus
Orthopedic Dept
Vienna
Austria

**REVIEW RETURNED**
29-Jun-2016
GENERAL COMMENTS

The main shortcoming of this study is the mode of data acquisition for the AOFAS score: despite this scoring system consists of subjective factors and objective factors (function, alignment), this scoring system never was intended to be used on a basis of self-administered use for pain (maybe even other parameters like walking distance, activity limitations or problems on uneven surfaces). This scoring system always was intended to be completed by the observer during personal interview and examination. So if the authors split the AOFAS score into a self-administered part (to be answered by the patient alone) and an "objective" part (to be completed by the observer), they may create a new scoring system with potential bias in comparison to the original intention. This principle of data acquisition should be reassessed.

On the other hand the study design with exact description of cross cultural adaptation and calculation of construct validity, reliability / internal consistency, reproducibility, floor and ceiling effects and responsiveness is convincing and will help to reassess the AOFAS scoring system.

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer #1:

1. The main shortcoming of this study is the mode of data acquisition for the AOFAS score: despite this scoring system consists of subjective factors and objective factors (function, alignment), this scoring system never was intended to be used on a basis of self-administered use for pain (maybe even other parameters like walking distance, activity limitations or problems on uneven surfaces). This scoring system always was intended to be completed by the observer during personal interview and examination.

So if the authors split the AOFAS score into a self-administered part (to be answered by the patient alone) and an "objective" part (to be completed by the observer), they may create a new scoring system with potential bias in comparison to the original intention. This principle of data acquisition should be reassessed.

Response: We concur with the reviewer that a physician completes the AOFAS instrument. That is how it is done in this study. A research physician completes the AOFAS using an interview for the subjective part (so based on answers provided by the patients) and physical examination for the objective part. This has been changed consistently throughout the manuscript (Study design, 2nd sentence; Study population, 2nd paragraph; Methods and Analysis, 2nd sentence; Study procedures, final paragraph of Step 1 and 2nd and 5th paragraph of Step 2).

We have no plans to evaluate the parts separately, therefore reassessment of the principle of data acquisition is not necessary. The sentence “The objective portion of the scale has not been evaluated for reliability” has been removed in order to prevent confusion.

2. On the other hand the study design with exact description of cross cultural adaptation and calculation of construct validity, reliability / internal consistency, reproducibility, floor and ceiling effects and responsiveness is convincing and will help to reassess the AOFAS scoring system.

Response: We thank the reviewer for this positive valuation.

Response to Reviewer #2:

1. Abstract: BMJ Open Author’s Guidelines direct authors to conform to the following headings in their structured abstracts “Abstract: this should be structured with the following sections. Introduction;
Methods and analysis; Ethics and dissemination. Registration details should be included as a final section, if appropriate. The current protocol does not conform to these heading. Ethics and Dissemination are missing and the inclusion of Discussion, Strengths and Limitations of this study are not usual practices for a scientific abstract.

Response: The headings of the manuscript have been made consistent with the guidelines (Abstract, removed Discussion, Strengths and Limitations; Abstract, added Ethics and dissemination; Replaced Background with Introduction; Moved Author contribution, Funding statement and Competing interest statement to the end of the manuscript). The results for the ankle and hindfoot injury subgroups are planned to be published separately; this has also been added to the Ethics and Dissemination section at the end of the manuscript.

2. Abstract: There are also conflicting statements as to the future use of the instrument to be created. In the Introduction, the authors state they wish to use the AOFAS-DLV to compare surgical results across hospitals, but in the Discussion there is a suggestion that the AOFAS-DLV may be used to monitor individual patient’s results. Both are valid goals for an instrument, but there are different statistical thresholds to be considered if one intends use for individual patients (further references to this will be included in the statistical review below).

Response: The results will be used to compare results of groups of patients, e.g. across hospitals or between patient groups. This has been specified more clearly (Abstract, subsection Introduction; Discussion, 2nd paragraph).

3. Study Design: Protocol authors hold the AOFAS in high regard as being the most widely used instrument for ankle outcomes. While it is true that the AOFAS has been widely used, the AOFAS' own position statement of 2011 (Pinsker E, Daniels TR. AOFAS Position Statement Regarding the Future of the AOFAS Clinical Rating Systems. Foot and Ankle International 32 (9), 841-842) does not. Specifically, Pinsker and Daniels state that the AOFAS measures have not been found to be valid or reliable (contrary to statements made by this protocol’s authors) and do not recommend further use of the AOFAS. This is a serious design flaw in the protocol that needs to be addressed.

Response: We acknowledge the criticism raised against the AOFAS Clinical Rating Systems. Although linguistic problems of the questionnaire we affect all study populations, inadequate measurement properties were mostly found in studies with mixed populations. Therefor we decided that re-evaluation of the AOFAS Ankle-Hindfoot Score would only make sense if done in homogenous populations. Also, the inclusion of items specific for hindfoot problems (which are not included in other lower extremity instruments) was appealing to us. Both the criticism and the rationale for using the AOFAS has been added to the Introduction (3rd and 4th paragraph).

4. Study Design: Protocol authors also misrepresent the findings of SooHoo et. al., 2003 (who report poor construct validity); and Westphal et. al., 2004 (who report descriptive results using AOFAS and SF-36 with low to moderate correlations between the scales but validation of the tools against each other is not a goal of their paper).

Response: The publication of SooHoo et al. refers to a 2006-paper where they report increased responsiveness of foot and ankle specific outcomes tools (including AOFAS) compared to the SF-36. The 2003-paper reporting poor construct validity has been added (Introduction, 2nd paragraph). The reference from Westphal regarding confirmation of validity has been removed, and information on the correlation between AOFAS and SF-36 by Westphal has been added (Introduction, 2nd paragraph).

5. Study Design: As far as the cultural adaptation measures taken to create the Dutch Language AOFAS, the authors propose employing the best standards available and would be rated favourably by critical appraisal tools such as COSMIN if they conducted the study by their proposed protocol.

Response: We thank the reviewer for this positive comment.
6. Study Design: Protocol authors suggest that the physician portion of the AOFAS score may be completed by a research physician or a research assistant. This is of concern as there are no details on how the research assistant will be trained to perform this task. Is it appropriate for a non-physician (possibly a person with no medical training) to be performing these assessments? Will inter-rater reliability be considered if conducted in this manner to prove use of research assistant is appropriate? Further description of how the AOFAS-subjective items will be administered is also necessary. Will the physician read those questions to the patient/will the patient be handed the survey to fill out for themselves/will the physician infer the responses according to a conversation with the patient? 

Response: The research physician is a PhD candidate with medical training and 1-year clinical experience in trauma care. The research assistant has a BSc in Medicine. An experienced trauma surgeon has given both assessors elaborate training on the administration and physical examination of the AOFAS Ankle-Hindfoot Score. Both the objective and subjective part of the AOFAS will be completed by the research physician. This is now specified in more detail in the text (Methods and analyst, 1st paragraph).

The research physician or research assistant will complete both the objective and subjective part of the AOFAS (see response to comment #1 of Reviewer #1).

7. Research Ethics; Protocol authors appropriately describe REB/IRB approvals and consent procedures. 
Response: No response needed.

8. Outcome Measures; Authors correctly define construct validity later on page 16, and state that the SF-36 and FFI will be used to triangulate construct validity, but do not state their hypothesized correlations between these instruments. It should also be noted that the SF-36 has previously yielded poor correlations in its use in construct validity and the AOFAS (see SooHoo, 2003). The protocol authors did well to choose another region-specific instrument (the FFI) which will likely yield more meaningful results for them.

Response: The hypothesized correlations have been added to the text (Construct validity, 1st paragraph).

9. Outcome Measures; Protocol authors state that the construct validity of the AOFAS-DLV will be the primary outcome measure (page 10/31, lines 36-43). Measurement properties cannot be the outcome measure; the AOFAS-DLV itself is the outcome measure, whilst demonstration of the measurement properties are the goal of the proposed study. Construct validity concerns whether an instrument measures what we think it measures.

Response: Reference to which parameter will serve as primary or secondary outcome measure has been removed (Abstract, methods section; Outcome measures, 1st paragraph).

10. Statistical Analysis: Sample size for group 1 is appropriate for testing. Sample size for groups 2 and 3 are not explicitly stated and range from 50-100 minimum. I would encourage the authors to use COSMIN as their guide in this regard and aim for the recruitment of ≥100 participants.

Response: The sample sizes mentioned are the minimum requirements according to COSMIN. As the reviewer states, we will recruit over 100 patients for both the ankle and the hindfoot subgroup. This is made more explicit in the text (Sample size calculation, 2nd paragraph).

11. Statistical Analysis - Construct validity: authors do not state their hypothesized correlations between AOFAS and FFI/SF-36. Given that the AOFAS is traditionally reported as one score and the FFI and SF-36 are reported using sub-scale or summary scores, this warrants clarification. Again the literature on the SF-36 has demonstrated poor correlation with the AOFAS and region-specific measures should be more useful.

Response: Hypotheses have been added as requested (see also response to comment #8 of this Reviewer). Since the SF-36 is a generic instrument (and more prone to effects of presence of
additional conditions and comorbidities), one would expect that correlation between AOFAS and SF-36 is less strong than between AOFAS and another region-specific PROM. That was the rationale for including both the SF-36 as generic instrument and the FFI as region-specific instrument. For evaluating construct validity, we aimed at expecting a range of correlations rather than finding only strong correlations.


Response: Future use of the AOFAS instrument will be at a group level. This has been added to the Methods section (Reliability/Internal consistency, 2nd paragraph. The reference has also been added as requested (ref. No. 22).

13. Statistical Analysis - Reliability (test-retest/reproducibility): the authors would be encouraged to include a “global change question” for the retest portion of the study so that they have an external way to ‘prove’ participants did not change in the 2-3 weeks between survey administrations. Test-retest ICCs should only be performed on those who state their ankle condition has not changed.

Response: We apologize for the omission not to include this in the manuscript. Indeed, we ask patients at t=2 for test-retest evaluation to confirm that they did not change since the administration at t=1 for test-retest. In case change is reported, they will be excluded from this part of the analysis. This has been added to the text (Study procedures, last paragraph of Step 2).

14. Statistical Analysis - Reliability (test-retest/reproducibility): Additionally authors should state which ICC they intend to run (most are using ICC 2,1 in cases like this).

Response: The ICC is indeed the ICC2,1. As stated in the manuscript, we plan a 2-way random effect model with absolute agreement. The ICC(2,1) is now mentioned in the text (Reproducibility, 2nd paragraph).

15. Statistical Analysis - Reliability (Measurement Error): if the MIC has not been published please reference “who” considers 10-20 points as relevant. [See the Position Statement for discussion of varying interpretations of scores].

Response: Based on this reviewer’s comment, we decided to drop this analysis (Reproducibility, last paragraph). As long as the MIC is not known, it is not possible to confirm if the SDC is smaller than the MIC. Any other surrogate for the MIC will be invalid.

16. Statistical Analysis - Floor and Ceiling effects: It should be noted that the limited number of response categories in the AOFAS make it highly prone to ceiling effects, and they have been previously reported. See Guyton G.(2001) Theoretical Limitations of the AOFAS Scoring Systems: An Analysis Using the Monte Carlo Modelling. Foot and Ankle International 22 (10), 779-787.

Response: Limitation of the AOFAS due to proneness to ceiling effects and the reference of Guyton have been added as requested (Introduction, 4th paragraph).

17. Statistical Analysis - Responsiveness: The introduction to this section needs editing for clarity (page 19, lines 16-25). Again, hypotheses are not documented.

Response: Hypotheses have been added (Responsiveness, 1st paragraph).

18. Final Remarks: This protocol represents an ambitious attempt to fill many of the psychometric gaps for the measurement properties of the AOFAS. It is important to note that the AOFAS has issued a position statement suggesting the AOFAS Clinical Rating Systems are not recommended for clinical
use. The authors need to defend why their proposed study is meaningful in the face of the AOFAS’ position regarding its use.

Response: The rationale has been added to the Introduction (3rd and 4th paragraph; see also comment #3 of this Reviewer).

19. Final Remarks: It would also be helpful for the authors to include references of the published psychometric properties of several foot/ankle instruments including the AOFAS and the FFI (e.g. see Naal FD, Impellizzeri FM, and Rippstein PF. (2010)) which are the most frequently used outcome instruments in studies on total ankle arthroplasty? Clin Orthop Relat Res 3(10), 815-826).

Response: Psychometric information and references have been added for the FFI (Study procedures, 3rd paragraph of Step 2. We preferred to restrict to data on traumatic foot injuries.

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Taucha Inrig</th>
</tr>
</thead>
<tbody>
<tr>
<td>St. Michael's Hospital, Canada</td>
<td></td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>08-Sep-2016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERAL COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The authors have made excellent progress on their protocol and have attempted to address many of the statistical issues.</td>
</tr>
<tr>
<td><strong>Background:</strong></td>
</tr>
<tr>
<td>- The authors still claim the AOFAS is valid in its original version, though this is in contrast with the AOFAS position statement of 2011 (The AOFAS states there is &quot;limited evidence&quot; supporting the AOFAS validity).</td>
</tr>
<tr>
<td>- The authors claim that previous studies of validation have been hampered by 'mixed populations'. Some validations have been completed in homogeneous populations (e.g. end stage ankle arthritis alone) so the claim that a homogeneous population has not been published is not true.</td>
</tr>
<tr>
<td>- The addition of the limitations of the AOFAS was a good addition to the background and reflects the AOFAS' position statement.</td>
</tr>
<tr>
<td><strong>Study Procedures:</strong></td>
</tr>
<tr>
<td>- Which version of the FFI will be used? (Assuming the Dutch Version, but it should be stated and referenced. Similarly the version of the SF-36 should be documented. It appears the FFI (Dutch Version) is performing very well and is probably superior to the AOFAS (see Marijke M Kuyvenhoven).</td>
</tr>
<tr>
<td>- It should be noted in your description of the FFI that &quot;strong correlations with the SF-36&quot; have been noted in some languages (not in all languages).</td>
</tr>
<tr>
<td>- Why are the Dutch norms not being used in evaluating SF-36 results (they appear to be published/available)?</td>
</tr>
<tr>
<td>- Specifically, how will 'stability' be assessed for the test-retest reliability? A description of the change question and where their response is recorded would be desirable (occurring twice in the paper - once above the sample size calculation, once on the top of page 20 in the track changes document).</td>
</tr>
<tr>
<td>- Please move the description of Effect Size above your hypotheses.</td>
</tr>
<tr>
<td><strong>General Readability</strong></td>
</tr>
<tr>
<td>- Consider reading the manuscript in its final form (not in track changes) there are grammatical errors that I think are just getting</td>
</tr>
</tbody>
</table>
missed in the track changes version (e.g. there are times when the plural "patients" is appropriate, but the "s" has been deleted, perhaps in error.

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Wolfgang Schneider</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orthop. Dept.</td>
</tr>
<tr>
<td></td>
<td>Herz Jesu Hospital</td>
</tr>
<tr>
<td></td>
<td>Vienna</td>
</tr>
<tr>
<td></td>
<td>Austria</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>06-Sep-2016</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS | The authors addressed all questions regarding shortcomings of the first version of the manuscript satisfyingly. |

**VERSION 2 – AUTHOR RESPONSE**

Response to Reviewer #1:

There were no comments from this reviewer.

Response to Reviewer #2:

1. Background: The authors still claim the AOFAS is valid in its original version, though this is in contrast with the AOFAS position statement of 2011 (The AOFAS states there is "limited evidence" supporting the AOFAS validity).
   Response: The strength of evidence has been reduced (Introduction, start of 3rd paragraph).

2. Background: The authors claim that previous studies of validation have been hampered by 'mixed populations'. Some validations have been completed in homogeneous populations (e.g. end stage ankle arthritis alone) so the claim that a homogeneous population has not been published is not true.
   Response: We agree with the reviewer and changed the statement to "some of these studies contained mixed populations". (Introduction, end of 3rd paragraph).

3. Background: The addition of the limitations of the AOFAS was a good addition to the background and reflects the AOFAS' position statement.
   Response: no response needed.

4. Study Procedures: Which version of the FFI will be used? (Assuming the Dutch Version, but it should be stated and referenced. Similarly the version of the SF-36 should be documented. It appears the FFI (Dutch Version) is performing very well and is probably superior to the AOFAS (see Marijke M Kuyvenhoven).
   Response: We indeed used the Dutch Language Versions (DLV), as mentioned in the Study Procedures (1st paragraph after translation process). For clarity, the DLV versions are now referenced specifically in this paragraph, as well as in the paragraphs describing the FFI and SF-36, respectively.

5.
Study Procedures: It should be noted in your description of the FFI that “strong correlations with the SF-36” have been noted in some languages (not in all languages).
Response: the language issue has been added (Study Procedures, FFI section).

6. Study Procedures: Why are the Dutch norms not being used in evaluating SF-36 results (they appear to be published/available)?
Response: The reviewer is correct that Dutch norms have been reported. The reason to use the US norms are twofold. First, the sample size of the Dutch norms is much, much lower than used for determining the US norms, making the US norm calculation more reliable. Second, most published studies have used the US norms; therefore it made more sense for us to also use those norms, for ease of comparison with other published studies. In a couple of previous studies, we used both the US and Dutch norms and compared the data. Results showed that on a study population level the means and median values were similar, but variance was larger using the Dutch norms than when using the US norm. This reassured us that the using the US norms had no undesirable effect on the results of our study. No changes were made to the text.

7. Study Procedures: Specifically, how will ‘stability’ be assessed for the test-retest reliability? A description of the change question and where their response is recorded would be desirable (occurring twice in the paper - once above the sample size calculation, once on the top of page 20 in the track changes document).
Response: Details have been added at both locations as requested (Study Procedures, last paragraph; and Statistical analysis/Reproducibility, 1st paragraph).

8. Study Procedures: Please move the description of Effect Size above your hypotheses.
Response: The description of Effect Size has been moved as requested (Study Procedures/Responsiveness, 2nd paragraph).

9. General Readability: Consider reading the manuscript in its final form (not in track changes) there are grammatical errors that I think are just getting missed in the track changes version (e.g. there are times when the plural “patients” is appropriate, but the “s” has been deleted, perhaps in error.
Response: The manuscript has been read in its final form and typos were corrected where needed. We apologize for the typos.

VERSION 3 – REVIEW

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Taucha Inrig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>St. Michael's Hospital, Canada</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>14-Oct-2016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERAL COMMENTS</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X Accept, so long as the issues below are dealt with prior to print</td>
</tr>
<tr>
<td></td>
<td>Minor Revision</td>
</tr>
<tr>
<td></td>
<td>Major Revision</td>
</tr>
<tr>
<td></td>
<td>Reject</td>
</tr>
</tbody>
</table>

Would you be willing to review a revision of this manuscript? N/A Yes No
Comments

Please leave your comments for the authors below bottom page 3/63:
Construct Validity is not an outcome measure, the sentence is probably meant to read “Descriptive statistics (including floor and ceiling effects), Construct validity, internal consistency (test-retest reliability, smallest detectable change), and Responsiveness will be assessed for the AOFAS-DLV. currently the statistical tests are mixed up under the wrong subheadings.

Page 6/63 near top:
sentence should read “Westphal et al., showed correlations between SF-36 and the AOFAS Ankle-Hindfoot Score were strong regarding…..” (i.e. drop the word “strong” which appears in the sentence twice”

Page 6/63 top:
Authors still claim all studies include mixed populations despite their response to the reviewer to the contrary. Please fix this line to read (as they said they would do in their response to the reviewer) “SOME of these studies have included mixed populations.” also Delete the following sentence “Whether or not the AOFAS Ankle Hindfoot score would be reliable and valid in homogenous populations……has not been published” There are published homogenous populations for arthritis and hindfoot disorders….and the authors are citing them

Page 6/63 bottom:
Delete sentence “Lack of evaluation of measurement properties…..in homogeneous populations for the same reasons as above.

Page 7/63 top:
Construct Validity is not an outcome measure, the sentence is probably meant to read “by assessing descriptive statistics (including floor and ceiling effects), Construct validity, internal consistency (test-retest reliability, smallest detectable change), and Responsiveness will be assessed for the AOFAS-DLV. Currently the statistical tests are mixed up under the wrong subheadings.

Page 8/63 line 45:
Delete “ones” in the sentence “the patient will be contacted by telephone ones”

Page 10/63 top:
exclusion of multiple trauma patients’ please define the functional limitations that would preclude their inclusion.

Page 14/63 middle:
The Dutch Language FFI needs to be referenced…currently American-English version is cited and measurement properties for the Korean and Chinese Versions are cited….yet they propose to use the Dutch Language version…..it needs to be referenced and its measurement properties should be cited here. Note their response
to the reviewer said this had been done…but it is still missing

Page 14/63 bottom:
SF-36, only citing humeral references. Lower extremity (and better yet ankle) uses of the SF-36 should be cited here. Also Dutch references and WHY Dutch norms are not being used should be inserted here Note their response to the reviewer said this had been done…but it is still missing, and they gave a very good defense of their reason for using the American norms in their response to me, it should be a part of their text…including their previously published works that discuss the similar mean/median values but increased variance. Excellent work which should be introduced to this section!

Page 15/63 bottom:
the actual wording of the ‘change’ question including its anchors needs to be inserted here in order for their study to be reproducible.

Page 18/63 line 45, again, we need to know how stability will be assessed.

VERSION 3 – AUTHOR RESPONSE

Response to Reviewer #1:

There were no comments from this reviewer.

Response to Reviewer #2:

1. bottom page 3/63: Construct Validity is not an outcome measure, the sentence is probably meant to read “Descriptive statistics (including floor and ceiling effects), Construct validity, internal consistency (test-retest reliability, smallest detectable change), and Responsiveness will be assessed for the AOFAS-DLV. Currently the statistical tests are mixed up under the wrong subheadings.
Response: Sentence has been changed, and statistical tests are mentioned under the headings as proposed by the COSMIN group.

2. Page 6/63 near top: sentence should read “Westphal et al., showed correlations between SF-36 and the AOFAS Ankle-Hindfoot Score were strong regarding…..” (i.e. drop the word “strong” which appears in the sentence twice”.
Response: The correction was made.

3. Page 6/63 top: Authors still claim all studies include mixed populations despite their response to the reviewer to the contrary. Please fix this line to read (as they said they would do in their response to the reviewer) “SOME of these studies have included mixed populations.”
also Delete the following sentence “Whether or not the AOFAS Ankle Hindfoot score would be reliable and valid in homogenous populations…..has not been published” There are published homogenous populations for arthritis and hindfoot disorders…..and the authors are citing them.
Response: Changed as requested.

4. Page 6/63 bottom: Delete sentence “Lack of evaluation of measurement properties…..in
homogeneous populations for the same reasons as above.
Response: The sentence has been deleted.

5. Page 7/63 top:
Construct Validity is not an outcome measure, the sentence is probably meant to read “by assessing
descriptive statistics (including floor and ceiling effects), Construct validity, internal consistency (test-
retest reliability, smallest detectable change), and Responsiveness will be assessed for the AOFAS-
DLV. Currently the statistical tests are mixed up under the wrong subheadings.
Response: The wording has been changed in analogy to the wording in the Abstract.

6. Page 8/63 line 45: Delete “ones” in the sentence “the patient will be contacted by telephone ones”.
Response: “Ones” has been deleted.

7. Page 10/63 top: exclusion of multiple trauma patients’ please define the functional limitations that
would preclude their inclusion.
Response: The exclusion criterion has been reworded. We mean to exclude patients who still
experience functional limitations due to other injuries at the time of enrolment, as that may have an
effect on the outcome scores.

8. Page 14/63 middle: The Dutch Language FFI needs to be referenced…currently American-English
version is cited and measurement properties for the Korean and Chinese Versions are cited….yet
they propose to use the Dutch Language version….it needs to be referenced and its measurement
properties should be cited here. Note their response to the reviewer said this had been done…but it is
still missing.
Response: Reference to the Dutch version has been added.

9. Page 14/63 bottom: SF-36, only citing humeral references. Lower extremity (and better yet ankle)
uses of the SF-36 should be cited here. Also Dutch references and WHY Dutch norms are not being
used should be inserted here. Note their response to the reviewer said this had been done…but it is
still missing, and they gave a very good defence of their reason for using the American norms in their
response to me, it should be a part of their text…including their previously published works that
discuss the similar mean/median values but increased variance. Excellent work which should be
introduced to this section!.
Response: The reason for not using the Dutch norms has been added. Also, references of the
comparative study and studies of lower extremity were added.

10. Page 15/63 bottom: the actual wording of the ‘change’ question including its anchors needs to be
inserted here in order for their study to be reproducible.
Response: We are not sure what needs to be changed, as the question with its answers is already
given: "How would you judge the condition of your ankle, compared with the last time you completed
this questionnaire? The item scored ‘better’, ‘no change’, or ‘worse’. The wording was changed a bit.

11. Page 18/63 line 45, again, we need to know how stability will be assessed.
Response: Stability between the two measurements is assessed using an anchor question (see also
item 10 above). Patients replying “no change” to the question are considered stable. This has been added to the text twice.