Author’s response to reviews

Title: The iPhone Measure App level function as a measuring device for the weight bearing lunge test in adults. A reliability study

Authors:

Helen Banwell (helen.banwell@unisa.edu.au)

Hayley Uden (hayley.uden@unisa.edu.au)

Nicole Marshall (nicole.marshall@mymail.unisa.edu.au)

Carlie Altmann (carlie.altmann@mymail.unisa.edu.au)

Cylie Williams (cylie.williams@monash.edu)

Version: 1 Date: 16 May 2019

Author’s response to reviews:

“The iPhone Measure App Level function as a measuring device for the weight bearing lunge test in adults. A reliability study”

The authors would like to thank the reviewers for their constructive feedback and suggestions. We hope you find the changes below suitable and agree that they have strengthened the article.

Please be aware that line numbers have now been added to the manuscript (inadvertently not added to the original submission). We will refer reviewers to the (new) line numbers as required below. Apologies in advance for any confusion caused.

Reviewer #1

Congratulations on a neat and concise research project which will be a useful addition to the clinical tool bag. The content of this work is great, and I have no real criticism to the project, there are however a couple of suggestions that you may wish to consider.

1. You state a reason for your sample size and later go on to state about power, I would put this altogether in one paragraph to reduce any confusion.
Response 1: The sample size calculation and power consideration are within the same sentence (Line 216) as below:
“A minimum sample size of 18 was calculated to provide 80% power of detecting a correlation of 0.6 with a two-tailed alpha = 0.05 for the intra-rater reliability analysis.”
The confusion maybe due to the first line of the subsequent paragraph discussing the outcome of having 21 participants etc. Unfortunately, the second paragraph falls into the Result section, whereas the sample size calculation etc., by convention, needs to remain in ‘Methods’. Hopefully this doesn’t cause too much confusion for readers.

2. Page 8 you refer to Figure 2 and 3 in the text and I think this should be 1&2.
Response 2: Thank you for identifying inconsistency, it has been rectified (Line 180).

Reviewer #2
Thank you for the opportunity to review this paper investigating the reliability and validity of an iPhone App for measuring ankle joint dorsiflexion in a weight bearing lunge test.
Overall the paper supports the use of this widely available App in clinical practice as an alternative to a digital inclinometer.
I have the following comments related to the manuscript:

Abstract
3. The methods section does not contain information about the number of trials done, methods for calculating reliability etc please update. Please review your abstract for consistency of terminology and letter capitalization.
Response 3: The abstract has been updated to include the number of measures and the methods of calculation for reliability and concurrent validity as below (please see manuscript for specific changes as highlighted).
From line 37: “To determine intra-rater reliability, inter-rater reliability and concurrent validity, 168 measures were conducted on 21 participants, preconditioned prior to assessment, where ankle dorsiflexion in the knee extended and knee flexed positions of the weight bearing lunge test were measured by two experienced raters using an iPhone level function (of the Measure application) and a digital inclinometer in a randomised order. Concurrent validity was also determined via comparison of measures of the two devices at two known surface angles (0 and 15 degrees) in multiple planes. Reliability and validity were determined with intraclass correlation coefficients, concurrent validity was explored with the Bland Altman plot and an intraclass correlation coefficient. The Standard Error of the Mean and the minimal detectable change were also explored”
We have also reviewed the terminology and capitalisation of words and hope we’ve addressed any inconsistencies appropriately.

Background
4. Your second statement needs references.
Response 4: The following two references have been added to the statement (Line 90). Thank you for the suggestion.

Second paragraph:
5. You start the paragraph by identifying digital inclinometers are used in clinical practice to measure ankle joint range of dorsiflexion and are precise and repeatable. Then make the argument that these are expensive and clients/carers may wish to take the measurement at home. There is no mention of relevance of this to clinicians? Please update.
Response 5: The indication related to cost was aimed at clinicians, however, this was not clear in the
sentence. It has therefore been updated (Line 98) as “However, digital inclinometers may be considered costly for the average clinician and are not often accessible by clients/carers who may wish to assess range of motion changes at home.

6. I note the senior author has published a similar study using the Tiltmeter App and comparing this to a digital inclinometer. It would be helpful if there was a clear explanation of why one App may be different from another and therefore why an additional reliability study is required. Response 6: The authors believe this is introduced in the second paragraph of the Background (line 100) as below, however, have added the final sentence to clarify why this study is required.

“Specifically, the Tiltmeter App and the iHandy App (available on smart phones/tablets) have been shown as reliable measures of ankle joint dorsiflexion [13, 14]. These have the additional benefits of being cheap, easily accessible and quick to administer [13]. Unfortunately, with rapidly changing technology, these Apps become outdated and unsupported, as demonstrated with the recent discontinuation of the Tiltmeter App for iPhone users. With Apple’s recent software upgrade (operating systems IOS 7 and above) a new Measure App which includes a ‘level’ function has been introduced. This level function, if reliable, would potentially be a suitable alternative to the discontinued Tiltmeter App with the additional bonus of being included in the Apple App suite (that is, it is standardly installed/upgraded with each software upgrade). Furthermore, in Australia, iPhone users account for 45% of the smart phone market share (8.6 million users) [15], meaning the Measure App is freely accessible to a large population of smartphone users. To be confident in its use in the clinical setting, however, determination of the psychometric properties is required.”

Methodology
There are some aspects of the methodology that require further clarification.

7. In relation to the measurement of the ankle joint lunge test, previous studies have identified the likelihood of subtalar joint pronation affecting the measurement and used a vertical line placed perpendicular to the wall to limit the confounding effect of this on the measurement. This is done in addition to ensuring the knee drops over the second toe (as this could occur without the foot being perpendicular to the wall). Was this used in the present study or was a different method used to control foot alignment with the wall? Response 7: Whilst we agree that some studies to control for subtalar joint pronation, our study did not specifically do so beyond maintaining the knee over the second toe. Our reasoning was two-fold; 1) to obtain functional measures given it is taken at ‘end range of available motion’, and 2) to be consistent with much of the literature [e.g. Bennell et al., 1998; Chishom et al., 2012; Dickson et al., 2012; Jones et al., 2005; Konor et al., 2012 etc.,]. To clarify this for readers, reference to Bennell et al., 1998 (as the first paper to describe reliability of the weight bearing lunge) has been added to the protocol description as below (Line 162):

“The WBL test protocol used during testing was consistent with Bennell et al. [16] as follows: …”

8. I am unclear from the methodology as to whether an average of multiple trials by the assessors was used or, if only one measurement of each lunge test by each as assessor was used to determine the ankle joint motion and subsequent reliability. This needs to be specified as I am unsure from what has been provided whether the correct ICC model for interrater reliability has been used. Response 8: Additional information has been provided within the methods, data analysis and results to provide clarity on model chosen and which data was used within the model. For example, the methods section has been updated to include the point “A single measure was taken at each time point, in each position by each of the raters” (Line 173). Further clarification on the data analysis is reported in Line 198 “The intra-rater reliability between timepoints for equipment was determined using the raw data with the intraclass correlation coefficient (ICC) (Model 3,1),” and in Line 207 “The interrater reliability
determined with all raw data collected from two raters, for each position and each measurement tool using ICCs (Model 2,2) and 95% CI’s, Standard Error of the Mean (SEM) and the minimal detectable change (MDC).” The result section has also been updated accordingly.

9. In relation to the positioning of the measuring device "one cm superior to the posterior calcaneal tuberosity"- the referenced paper by Hoch et al does not measure from this point but uses distance from the wall to calculated the range of ankle joint motion. The other reference is in relation to a similar reliability paper and does not support the method as being one to measure ankle joint ROM. Please update the references and provide an alternative supporting using this technique.
Response 9: The references have been updated to Dickson et al, 2012; Williams et al., 2013 – as below) (Line 183) to reflect the measure protocol correctly. Thank you for identifying the irregularity.

10. Who deemed the sample size big enough to prevent recall bias- is this the judgment of the authors? In which case this should be clearly stated.
Response 10: The text has been updated to identify this was the judgement of the authors (Line 189):
“To minimise recall, participants were measured behind a partition that allowed the practitioner to visualise the person from their knees down only. The author group considered the sample size large enough to ensure raters were unable to remember the result; and the time space between retesting participants (minimum of 30 minutes) was appropriate to not cause fatigue to the target muscle group.”

11. The number of participants seems quite small. The sample size has been calculated for α = 0.05, β = 0.20 for detecting a correlation of 0.6. I assume this means the target ICC was 0.6? or was this the lowest acceptable ICC? Was the power calculation based on two observations per participant?
Response 11: The reviewer is correct; our minimum target ICC was 0.6 or greater. On review, we noted the sample size calculation and repeated measures methodology used was not referenced. This has now been provided and amended within the manuscript (line 216).

12. Was this an adequate sample size to determine measurement validity? More information needs to be provided as to how this number was calculated.
Response 12: We have provided a reference for appropriate sample size with a target ICC, based on two raters and repeated measures as per reference of Portney and Watkin, 2009. We are confident our sample size is adequate.

13. As a target ICC for a clinical measure 0.6 would be quite low and not what would be considered a particularly clinically useful measurement. Please identify the target ICC and the lowest acceptable ICC and justify accordingly or revise the power calculation.
Response 13: We disagree that 0.6 is quite low. Portney and Watkins (2009) described <0.5 as poor reliability and 0.5-0.75 to be moderate reliability – consistent with many frequently used podiatric measures (referenced below). As this has now been referenced, we have not made any changes within the manuscript.
14. References supporting the interpretation of the ICCs also need to be provided.
Response 14: These have now been supplied (e.g. Lines 204, 206, 215, 217). Thank you for alerting us to the oversight.

15. The SEM for the each rater for the intra-rater reliability has been calculated. This has not been calculated for inter-rater reliability from what I can see in your results table nor is it referenced (ie method used to calculate) in your statistical analysis section-please update this and calculate for all reliability measures.
Response 15: The SEM was calculated from the SD of the measures collected from a single rater in the single foot/leg position. To the best of our knowledge, SEM is unable to be calculated for the interrater reliability due to the values collected from two raters. We have consulted with our statistician who confirmed our interpretation. If the reviewer is aware of alternative SEM calculations that will allow for handling the two SDs we would be grateful for the guidance. We have, however, separated the outcomes of the calculations into two tables – which we hope reduces confusion (Table 1 and 2), and added the following to the manuscript at Line 202 “The SEM was calculated with the raw data with the following formula: \( \text{SEM} = \text{SD} \sqrt{1-r} \) where \( r \) was the ICC for intra-rater reliability [20].”

16. It would be worthwhile calculating the minimal important change (MIC) to identify the magnitude of change required for this to be a true change in ROM rather than error. Alternatively, 95% limits of agreement would identify the difference required for a true change. This would provide more clinically useful data. Conducting \( t \) tests between tests and retest data or rater 1 and rater 2 for mean ankle joint dorsiflexion measurements would also identify statistically significant differences between measurements or raters and identify is there was any effect from multiple measurements.
Response 16: It is our understanding, supported by the views of our statistician, that identifying differences with a \( t \)-test between rater 1 or 2 is not appropriate for reliability statistics (it can be calculated but it adds nothing to what we have already determined). We have added further information to the method and results section to reassure readers of no differences between timepoints (Lines 195 & 230 respectively). Additionally, we have provided the minimal detectable change and how it was calculated. We have not included the minimal important change/difference (MIC or MID). The MIC (or MID) is generally used within a longitudinal manner, particularly where there is a change in health status (e.g. change in ankle ROM over time). It also requires an anchor or additional measure to compare (de Vet H.C.W, Terwee C.V., Journal of Clinical Epidemiology, 2010). This was not the aim of the article, nor were any intervention put in play where a MIC would be expected. Where as the minimal detectable change or MDC is distribution based and the smallest change detectable by the instrument beyond error. This has been added to the methodology and results.

17. I am a bit confused about the statistical analysis for the concurrent validity. It seems an ICC was performed and a Bland Altman plot. Limits of agreement are reported in the results but not included in the statistical analysis-please update. How were the Bland Altman plot calculated? Did the calculation need to be adjusted for the small sample size? I am not sure why all the measures where included in one plot given you found worse reliability in the straight leg lunge versus the bent knee? It would seem to be a clearer method to do this as two separate plots for the straight leg and bent knee data.
Response 17: Additional information is provided and referenced for the Bland Altman plot and what included data was used to determine the limits of agreement and 95%CI for limits of agreement (Lines 210 – 213). This has been calculated and interpreted as per the referenced and standardly accepted manner for concurrent validity where two instruments are being tested rather than the test itself. We recognise that the ICC measure for the straight leg lunge was less than that of the bended knee, however, all measures were good to excellent. All data were analysed with Stata 15 and this is already provided within the manuscript. We hope this is satisfactory.
18. In addition there is no mention of whether these data were normally distributed? Were data assessed for these? Please update.
Response 18: This has now been included within the data analysis (Line 196), thank you.

Results
19. You currently have two tables labelled Table 1, please update your reliability outcomes table to Table 2 and change text accordingly.
Response 19: This has been updated (Line 442). Thank you for highlighting the concern.

20. Your reliability results table is a bit confusing. I assume the 95% confidence interval in the last column is for the ICC- this really needs to go in the column next to the ICC value with the SEM in the final column. In relation to the inter-rater reliability data I am not sure what is being reported here-it states it is for rater one versus rater two in straight or bent knee position but gives only one mean value for each knee position when there should be two? (one for each rater). Please revise this.
Response 20: The columns have been reordered and the additional data have been added. We originally just reported intrarater for rater 1 OR 2 due to the similarity, but we understand that this may have added to the confusion. We have now provided the additional data for each rater in each position with each tool (Table 2, Line 442 – 444).

21. The methodology for the concurrent validity (line 46) should be in the methodology. This aspect of the study is not very well described, please review.
Response 21: Unfortunately, due to the line numbers no longer being relevant we are unable to determine the exact concern in Line 46, however, assume it might be the sentence below:
“The concurrent validity was initially determined between the digital inclinometer and level function on static hard flat and angled (15 degrees) surfaces. This ICC was 1.0 (Limits of Agreement -1.0 to 0.61), indicating excellent reliability.” If this is the case, we agree and have altered it to reflect the results only (Line 239) “Initial concurrent validity, determined between the digital inclinometer and level function on static hard flat and angled (15 degrees) surfaces, was ICC of 1.0 (Limits of Agreement -1.0 to 0.61), indicating excellent reliability.” We have also moved this to just prior to the concurrent validity results determined by the rater measures. We hope this change, and the further information supplied within the methodology section is suitable.

22. Why were 21 participants recruited if your power calc required 18?
Response 22: The power calculation determined a minimum of 18 participants were required but we were happy to include all participants that enrolled into the study, met the criteria and were willing to give written informed consent.

23. Your mean ankle joint dorsiflexion seems very low given your population. Is this finding consistent with other studies using this method? In a healthy cohort of adults straight leg lunge means are typically closer to 40 degrees. Please ensure this is included in your discussion.
Response 23: We respectfully disagree. While lower than some means published (Bennell et al 1998 or Konnor et al 2012), it is also within the range of others (Burns & Crosbie 2005, Williams et al, 2013). We have however provided this within the discussion with the following text (Line 296):
“Lastly, the mean values of weight bearing lunge were lower than other reported ranges [3, 4], however, comparable to other published values in normative populations [5, 6]. It is unknown what impact that may have on measured reliability.”

24. In relation to the inter-rater reliability data I am not sure what is being reported here-it states it
is for rater one versus rater two in straight or bent knee position but gives only one mean value for each knee position when there should be two? (one for each rater). Please revise this.
Response 24: We have amended the table to separate the intra and inter-rater reliability results to provide clarity and the missing results (Lines 442). We hope that gives clarity for the readers.

25. The results for your 95%LOA for your concurrent validity are different in your results here to the results in your abstract, please clarify.
Response 25: We apologise, this was a transcription error from the statistical program. This has been amended.

Discussion
26. The first paragraph of the discussion would work better in the introduction- it is justifying why the study was done. This paragraph in the discussion should provide a summary about what was novel about your research and overview of what you found.
Response 26: We agree. We have removed the majority of the first paragraph and replaced it with the following text (Lines 251 – 254):
“To the best of our knowledge this is the first use of the new iPhone level function within the Measure App to review reliability in ankle joint range of motion measures. The outcomes of the study suggest the tool is comparable to digital inclinometers and can be used to measure the weight bearing lunge test in healthy adult populations with confidence.”

27. The second paragraph of the discussion gives a broad overview of reliability results stating all but two scores for reliability were above 0.9. Is this relating to ICCs? There needs to be more precise differentiation between the results of intra-rater, inter-rater reliability results.
Response 27: We have clarified this sentence to be more of a summary point than a repeat of the results. It now reads as follows, Line 260: “The results from this present study determined intra and inter-rater reliability of all measures were deemed good or excellent. Concurrent validity analysis was also determined as an acceptable comparability of the measurement tools, with a low bias and a mean difference close to zero.”

28. Additional statistical analysis relating to MIC/95%LOAs as mentioned above would more adequately assess reliability of the App and allow for a more detailed discussion of what these results mean for clinical practice. You make a statement in this paragraph suggesting the iPhone App can be confidently included in clinical practice to measure ankle joint dorsiflexion. Given you only had healthy people you need to be cautious generalizing your results to clinical populations where more extreme measurement may occur. Please revise this.
Response 28: We have revised the sentence to read (Line 263) “Within a healthy adult population, this method of measurement, along with the use of the level function within the iPhone Measure App, can be confidently introduced into clinical practice for quantifying ankle dorsiflexion range of motion.”

29. There needs to be a discussion of the concurrent validity results the limits of agreement and what these mean.
Response 29: We have included the following sentence regarding the concurrent validity (Line 261). “Validity of the level function was also determined as an acceptable comparison to the digital inclinometer, with a low bias and a mean difference close to zero.”

30. There also needs to be discussion about the mean ankle joint lunge measures recorded in this study and how these compare to other studies in healthy populations and why your results may differ.
Response 30: As per response 23, we have addressed this in Lines 295 to 297. We hope this is to your
Reviewer #3: Thank you for submitting this article. I have a number of comments for attention;

31. In the abstract, background, an additional sentence could be included between the first and second sentences related to the (putative) link between ankle joint range of motion and pathology. E.g. 'Ankle joint range of motion is frequently assessed by health care clinicians who manage lower limb pathologies due to a perceived...'
Response 31: The following has been added to the first sentence of the abstract (Line 30) “Ankle joint range of motion is a frequently assessed measure used by health care clinicians who manage lower limb pathologies to identify ankle equinus and/or other joint motion concerns that may negatively impact on function.”.

32. Line 3, background. '...known negative impact'. Again, the putative, complex nature of the link could perhaps be acknowledged.
Response 32: We have altered the sentence to be less definitive, it now reads “A reduced range of ankle joint motion (i.e. ankle equinus) has been shown to have a negative impact on lower limb function and economy of gait in healthy and pathological populations [1-6].” (Line 85). We hope this is suitable.

33. Line 12, background. 'Reassessment of measures used to determine success'. Maybe, but more likely to be pain and/or functional limitation - patient-oriented outcomes. Discriminating between patient-oriented and clinician focused outcomes would help balance the article.
Response 33: Whilst we agree that pain and function improvements are the primary goal for most interventions, identifying the change in joint ROM has its place within assessment practices and is often used to quantify success (particularly of prescribed exercise plans etc.). From our perspective, it is outside of the scope of the article for patient-oriented outcomes to be included within a reliability study and no change to the manuscript has been made.

34. Line 28. Use of term 'precision'. In the abstract the terms reliability and validity are used. Stick to a consistent technical language.
Response 34: The terms ‘precise and repeatable’ have been replaced with ‘reliable and valid’ accordingly (Line 96).

35. Line 22, Methods. Terminology - leg straight and knee bent; knee flexed and extended would be better, as is used later on.
Response 35: Agreed. The manuscript has been updated to consistently use the terminology of knee extended or knee flexed.

36. Line 24, Methods. Define concurrent validity either here or earlier. Generally, provide referenced technical definitions for key terms. In the discussion there may be room for some discussion of this class of validity and other forms that may warrant investigating in the context of this type of study.
Response 36: As per the similar considerations raised from reviewer 2, additional methods, results, discussion and references have been provided on concurrent validity and the following “layperson’s” explainer added to the Introduction at the first mention of the term (Line 117) “The secondary aim was to determine the concurrent validity of the two tools (i.e. how well does the level function measure when compared to the digital inclinometer).”

37. Use of two raters with post-graduate research training who use the WBL routinely during...
clinical practice. Implications for external validity - does this provide a representative insight to the potential that the measurement technique has for the wider body of clinicians in practice?

Response 37: The option of using a novice rater was considered by the author group, however, as the main aim here was to determine the reliability and validity of the tool, it was determined that using those with 8 years’ experience and post-graduate research training may better represent the “average” clinician. We have discussed this as a limitation in line 288 as “Experienced raters conducted all measures. Alternative studies on reliability have included a novice Rater to compare, therefore care should be taken in considering how these results may apply to the learner user.”

Response 38: Given the senior author (CMW) on this paper has been involved in a similar study and the raters were involved in the protocol development the final protocol was adopted without specific refinement. To the best of our collective memory there were not any changes made nor issues identified to report.

39. I can see that there were 21 subjects, and 2 clinicians. I can also see that there was a preconditioning step, whereby subjects held the WBL positions for 30 seconds, 3 times. I can also see that there were a total of 168 measurements. However, I can't seem to find the number of repeat measures conducted by each rater on each subject, how this links to intra- and inter-rater reliability assessment. More information - a simple flow chart - would help clarify this.
Response 39: Due to similar concerns raised from Reviewer 2, the methodology section has been modified accordingly. We hope this extra detail clarifies the process for the reader.

40. Blinding raters to the subjects is a useful technique for increasing the quality of measurements - internal validity. Statistical analysis is appropriate. In the discussion there is acknowledgement of the limitation incurred by using experienced raters and not including a novice for comparison. This is a very important point, and impacts significantly on external validity, compromising the conclusions that can be drawn. It may be more appropriate to state that this is an initial exploratory proof of concept study to determine whether clinically acceptable levels of intra- and inter-rater reliability can be achieved by experienced clinicians prior to wider evaluation of how this reliability transfers to the wider population of clinicians.
Response 40: Reviewer 3 is correct that this study has ‘given the green light’ to further studies within paediatric and known pathology groups using novice, experienced and non-health professional raters. This has been discussed in the final paragraph of the discussion as below:
Line 300 “Future research in the use of this technology for measurement should include understanding the reliability in children and in pathological populations, where there is (potential for) a smaller surface area for device placement. There is also the potential to consider including family/carers in future assessment of this and alternative measuring Apps to determine appropriability of non-health professional’s ability to determine success where interventions have been prescribed to improve ankle flexibility.”
However, these further projects were conceived and proposed independently after the completion of this current study – which was designed as a stand-alone investigation. Therefore, it would be remiss of the author group to retrospectively suggest otherwise.

41. 2nd page of discussion, final paragraph line 53. 'Suggested' would be a better word than 'promoted'.
Response 41: Actioned (line 291). Thank you.
Final sentence of discussion. 'Appropriability' is an unusual word. I couldn't find a definition in the Oxford English dictionary, but did find a definition through a general web search. In relation to economics: 'The environmental factors that govern and innovators ability to capture profits generated by an innovation'. I think appropriateness would be better, and think this is your intended usage. Response 42: Actioned (line 304) – as an aside note this came as quite the surprise to the lead author who has used that term without question for many, many years. Thank you for the clarification.