Reviewer’s report

Title: Field assessment of balance in children and adolescents, reproducibility and validity of the Nintendo Wii board

Version: 1 Date: 13 February 2014

Reviewer: Arnold Huurnink

Reviewer’s report:

Review comments:

Discretionary Revisions
- Title: I suggest to remove ‘and adolescents’, as ‘children’ adequately refer to the sample used (10-14 years old). Furthermore, the term ‘adolescents’ has differential meaning in different countries.
- Please be aware that a sampling rate of 125Hz lead to unnecessary increase of the noise, due to aliasing of the noise in the higher frequencies domain, which cannot be removed by filtering techniques. Therefore I recommend to use a sampling rate of 1000Hz. I understand that for the present measurements this is not possible. Moreover, I think that 125Hz is sufficient for the present objective, but for future measurements it is an easy way to decrease noise levels.

Minor Essential Revisions
- Methods-Test procedure- paragraph 5: barefoot instead of bare-footed.
- Equipment and data:
  - paragraph 1: Please describe the type/model of AMTI force platform, while it is not necessary to describe the type of amplifier.
  - paragraph 2: ‘Data was stored on a disk for later analysis’ may be removed. This is obvious.
  - Paragraph 3: Again, it is not necessary to provide amplifier model or digitization card model. The type of AMTI force plate is far more important to mention.

Major Compulsory Revisions
General:
1) I think the paper is well written and will contribute to the existing literature concerning balance testing in children and the use of the NWB.
2) Nevertheless, I also think it’s current form needs substantial revision, as the authors interpret the concurrent correlation coefficient (CCC) as most important analyses for reproducibility and validity of the NWB. Although the analyses of CCC may be informative, the CCC values should be related to the clinical use of the tests, and without this consideration, the CCC values cannot infer proven reproducibility or validity. Therefore, the point estimate used to describe the ICCs (>0.75 = excellent), and the consequent interpretation and conclusions of the
paper based upon this value may be misleading. For instance, the Bland and Altman LOA clearly indicates substantial intra-subject variability, while conclusions are based on CCC values, and a discussion that states the most important findings as “Reproducibility of both the NWB and the AMTI platform proved excellent in a population of children and adolescents (CCC: 0.76-0.86). Concurrent validity between the platforms in the four different tests was found excellent for COPL in three of the tests and modest in the remaining tests begin only 0.01 point form categorization as excellent (CCC:0.74). Thus, if the variable of interest is COPL, the results from NWB are comparable to results from the AMTI....”.

The paper should be rewritten, with the methods of Bland and Altman as the primary finding to indicate reproducibility and agreement of NWB and AMTI. CCC findings should only be used as additional information, and the current point estimate should not be employed.

3) Another important issue, is the presence of substantial intra-subject variability (as to be expected), that obscure comparison between NWB and AMTI. This is currently not adequately addressed. Please explicitly describe how this important limitation is handled, and why it is still possible to provide conclusions about the validity of NWB.

Background:
Paragraph 2:
- Although this paragraph describes that children differ from adults in their postural stability, it does not say why this could lead to error of measurement of the NWB compared to an AMTI, neither why a field assessment is important to consider in this particular group. I think a stronger case can be built when the following arguments are implemented in the background: 1) The NWB possess higher noise levels than laboratory grade platforms, and it has been shown that these noise levels even increase with lower weights. Therefore, the NWB may be less accurate for children (Huurnink et al. 2013). This is even more important when COP speed values are low, as in bipedal stance. 2) The NWB does not measure horizontal forces. A relatively higher centre of mass location, and/or a differential postural control strategy may result in relatively higher horizontal forces, hence higher error of measurement.
- Please make explicit why reproducibility of postural stability is important to consider in field assessment, as many papers already analyzed this in a laboratory setting. Why may a field assessment result in different reproducibility? (this is mentioned in the discussion).
- Why may children differ in reproducibility compared to adults? In the discussion the authors mention some explanation why children may be less stable in their performance. I think this could be incorporated in the background to justify the focus on children. The current indirect argument used that children differ from adults in balance performance seems weak, please make more explicit.

Paragraph 3:
Why focus on COPL? COP speed (or often denoted as COP velocity) seems
more intuitively, can be directly compared between studies/tasks, irrespective of trial length, and is more commonly applied. Please remove ‘as Centre of Pressure path length (COPL) excursions, or’ and direct your paper to COP speed instead.

Paragraph 4:
- Please change COPL in COP speed.
- I think that two additional papers should be cited after the statement “… are both reproducible and comparable with sway measures obtained from laboratory force platforms”, as these papers substantially increase the body of knowledge on the topic.

http://dx.doi.org/10.1016/j.gaitpost.2013.07.010

Furthermore, reference 15 does not hold for the above statement, but demonstrates specific sensitivity for older adults, rather than reproducibility or comparison with laboratory force platforms. When reference 15 has to be within the background, please rephrase the statement, or make an additional appropriate statement. Although, I am not sure if it is really necessary to mention older adults in the context of the background.

Paragraph 5:
- The objective is described as: … reproducibility of the NWB and a laboratory force platform (AMTI)…

At this point, the background does not directly state why reproducibility of AMTI is of interest. 1) Probably, the application in a field assessment adds significantly to the literature, as the authors mention this has not yet been performed. 2) and/or maybe a sample of children has not been thoroughly investigated yet, and this is of interest, since children may not be as stable in their performance as adults (e.g., due to lack of concentration). Please describe why there is a need for reproducibility analyses of AMTI postural stability. Maybe it helps to consider the purpose of the paper as twofold: 1) reliability of postural stability of children in field assessment, 2) reproducibility and agreement of NWB with AMTI.

- The field assessment as mentioned in title and background is not mentioned. I think this should be added to the objective statement, as this adds to the existing literature. Maybe more importantly, otherwise, I don’t understand why the NWB was not placed upon a laboratory platform to assess the validity for children in a laboratory setting (see point below).

Methods:
Test procedure:

Paragraph 3:
- The authors mention that the unilateral stance is performed on the dominant, and non-dominant leg. Please describe what definition was used to classify the dominant leg.

- Furthermore, I think it is better to avoid the term ‘the dominant leg’, as leg dominancy seems task and difficulty dependent (Huurnink A, Fransz DP, Kingma I, Hupperets MD, van Dieën JH.2014. The effect of leg preference on postural stability in healthy athletes. J Biomech. 3;47(1):308-12), and the term is also used for ‘the best’ leg on a task. If the authors determined ‘dominancy’ through leg preference tasks, or a questionnaire, please use the term ‘the preferred leg’ and ‘non-preferred leg’.

Equipment and data:

Paragraph 2:

- The authors state the sampling rate of the NWB was 60Hz. Is this the actual sampling rate, or is it the retrieving rate? In other words, does the NWB actually provide new data points 60 times per second, or was the laptop software configured in such a way to ‘ask’ for Bluetooth signals 60 times per second?

Paragraph 4:

I understand that in the present study the COPL is analogous to COP velocity (more accurately COP speed). However, I strongly recommend to provide the results of the COP speed instead of the COPL, as this measure is far more often used, in my opinion more intuitively, and may be compared between tasks/studies far more easily (as it is irrespective of trial length). Subsequently, this phrase can be removed.

Statistical analysis

Paragraph 1:

I am not convinced for the choice to apply the median score as the outcome value, rather than the average value of three trials, for the following reasons:

1) it is common practice to average the outcome over 3 trials. This method has been tested through reliability analyses in many papers, in contrast (to my knowledge) to the median score, hence the number of trials that should be assessed for reliable measurement are not known. Furthermore, comparison with other studies becomes more complicated.

2) the median score has not yet (to my knowledge) demonstrated sensitivity for injuries, in contrast to the average score.

3) the median score was used to avoid the effect of outliers on the outcome, however, with only 3 trials, an outlier cannot be identified, as the ‘true’ score is unknown and a confident estimation of the true score cannot be made with only 3 trials. Therefore, the actual effect of deleting the worst and best score is unknown.

4) I would recommend to present the inter-trial and test-retest reliability (see below) in the paper. If the authors do so, than it makes more sense to consider the mean of 3 trials, since this corresponds to the value of inter-trial average
measure ICC2.3 (3 trials).

Paragraph 2:
- Why was the NWB not placed upon a laboratory platform, as this would delete intra-subject variance? Please address this issue thoroughly in the discussion.
- Otherwise, the design to measure two equipments and to perform a repeated measurement design for both equipments is sound (J. M. Bland and D. G. Altman. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22: 85–93) This allows the authors to estimate intra-subject-variability and compare this with the differences found between the two equipments. I think the methods of Bland and Altman should be performed as primary analyses for both intra-subject variability (reproducibility) and equipment comparisons (agreement).
- Although the SEM is commonly used, I think it is better to avoid this term, as the standard error of measurement imply a definite error for the test. However, it should be handled with care and is dependent of the population used, and which variances (inter-trial, inter-session, inter-occasion et cet) is considered. Instead, Bland and Altman (2003) suggest to use the ‘repeatability coefficient’ (SD of difference * 1,96) for repeated measurements, and the limit of agreement (±SD of difference *1,96) for agreement between methods. This can be directly compared, (The MDC has exact similar values, but I recommend not to mix different concepts) and the interpretation discussed.

Paragraph 3:
As said above, Bland and Altman methods do best quantify the agreement and reproducibility. Another concept is not necessary to provide. Concordant correlation coefficients is a concept that seems valid, but I wonder if this is helpful for the interpretation, or make things more complicated or even misleading (the chosen point estimates that denote 0.75-1.00 as excellent). Therefore, please remove the concordant correlation coefficients, or only consider them as additional information.

Instead, I would add reliability analyses for within-session (inter-trial: intra-class correlation average measure (3 trials), random effects ), and between sessions (test-retest: intra-class correlation absolute agreement single-measure, random effects). This would provide information about within subject variance between trials, and whether three trials are appropriate for children in a field assessment. Point estimates can be used to denote the variance components relative to between subject variance in terms of ‘acceptable >0.70; and desirable >0.80 (Doyle et al. 2007), as these are more neutral qualifications, and seem more appropriate for clinical use of the current tests. These ICC values may be compared with previous studies. This would strongly add to the significance of the paper.

Results:
The results can be presented more straightforward, but this is related to my earlier comments on the choices of analyses.
I suggest the sub-headings (and corresponding analyses):
1) reliability analyses: only presenting ICC values for within-session, and between session reliability for both NWB and AMTI.
2) repeatability of balance tests: presenting ‘repeatability coefficient’ (SD * 1.96) for NWB and AMTI, with corresponding Bland and Altman plots.
3) agreement between NWB and AMTI: presenting ‘95% LOA’ for NWB and AMTI, with corresponding Bland and Altman plots.

Discussion:
Paragraph 1:
The CCCs (even without considering the low lower bounds) are not sufficient to proof ‘excellent’ reproducibility and validity. For example, a concurrent validity of 0.74 to 0.86 corresponds with high 95% LOA (see Table 2). If these limit of agreements were due to measurement errors of NWB, the NWB should have been rejected as valid measurement tool. However, this LOA is mostly due to intra-subject variability, rather than to NWB errors, as the LOAs are comparable. So the present data justifies the conclusion stated: ‘Thus, if the variable of interest is COPL, the results from NWB are comparable to results from the AMTI…’, but the correct line of reasoning is not stated in the discussion. Therefore, please remove the CCC’s from the paper, en rewrite the discussion with appropriate interpretations of the results of Bland and Altman for reproducibility and agreement.
In my view the most important findings of the present study is that intra-subject variability is substantial (see MDC (or to be denoted as ‘repeatability coefficient’)), but that test-retest differences are comparable to the differences between NWB and AMTI (see 95% LOA).

Paragraph 2:
- I agree that point estimates are arbitrarily chosen. This paragraph indicates that these point estimates should be taken with care, but also suggest that the CCC (or ICC) values should be considered in the specific measurement setting. Therefore, Bland and Altman methods should be seen as the primary analyses for reproducibility and agreement, and these values should be related to clinical data. The discussion should be based on Bland and Altman results, and this paragraph can be removed.
- A comparison with Clark et al. should not be an objective. The validity of the NWB in children does not depend on the results of Clark et al. The 95% LOA may be related to intra-subject variability, between subject variance, and expected difference with injuries. For reproducibility results, many studies are present to compare with.
Paragraph 3:
- Please move parts to the background as a rationale why it is important to assess reproducibility analyses in children.
- If a comparison with adults has to be made, there is a broader range of literature available.

Paragraph 4:
- Indeed, intra-subject variability is substantial. Why is reproducibility then denoted as excellent?
- Probably compare this MDC value with other studies?
- Although the problem of the high intra-subject variability is well described, I think it is important for the readers to understand that the MDC (or repeatability coefficient) can be used for an individual outcome, but is considerably smaller when comparing groups (as to be divided by n participants). Probably, that’s why similar tests have been shown to be of clinical importance in previous work.

Paragraph 6:
- ‘… are small and the reproducibility results are excellent. Thus, it is anticipated to be of minor importance’. Do you maybe mean that this is small compared to within subject and between subject variance? Since I don’t understand why the reproducibility results will lead to the stated conclusion.

Paragraph 7:
- It is a bit confusing to discuss MDC, and SEM in separate paragraphs, do they not in essence have similar meaning? Why discussing both in results and discussion? Readability can be greatly enhanced by presenting/discussing only necessary estimates.
- Part of the results is explained with an assumption: ‘Thus, the performance of the participants will be more precise in the bilateral tests, since variation between the trials is smaller.’ This can be illustrated when the ICC within-session (ICC2.3) is calculated

Paragraph 8:
In view of the confidence intervals of the CCC values, I wonder if the CCC values really differed between age, and sex. Without the actual results presented (which I understand in view of the power), please remove this from the discussion. However, the considerations made may be postulated as limitations, and the directions for future research can stay within the manuscript.

Paragraph 10:
‘The strength of the current study is that the setting was a field test setting’. Please emphasize this more in the background and objectives, than it is not necessary to explicitly state this in the discussion.

Paragraph 11:
You don’t need to state that “In addition, the inclusion of both unilateral and bilateral balance tests, and the large numbers of participants are strength of the current study.” Moreover, in view of the confidence intervals of the CCC values, and the intra-subject variability that obscure comparison between NWB and AMTI, the number of participants is not very large.

Conclusion:
As mentioned before, conclusions should be based on the right analyses and interpretation.

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

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests' below.