Reviewer’s report

Title: Discriminant validity and test re-test reliability of a gait assessment in patients with vestibular dysfunction

Version: 2 Date: 9 April 2015

Reviewer: Thorlene Egerton

Reviewer’s report:

General comments

The study appears to be well conducted and the methods for determining the quality of the measurement properties are mostly robust. Overall, the paper is a little difficult to follow and could be improved by some restructuring and re-writing of sections and some changes to what is reported. For example, it would be helpful if it was clear from the start that the walking conditions being tested were taking in part from a previously developed performance test. In addition, the property of validity requires a specific hypothesis to be stated and then tested.

Major Compulsory Revisions

Background

Terminology is somewhat inconsistent and clarity could be improved in the introduction and throughout the manuscript. I recommend using the taxonomy proposed by the COSMIN group whereby the terms reliability, or even better, reproducibility (test-retest), measurement error (SEM and SDC), and hypothesis testing (known groups or discriminative), are used in relation to your study. Finding differences between groups does not prove validity, it contributes to evidence for or against validity which is an ongoing process.

Information on the COSMIN taxonomy for measurement properties can be found at http://www.cosmin.nl/cosmin-taxonomy.html

In addition, the terms relative and absolute reliability are often used in the literature differently to how you have used them. Using the terms reproducibility and measurement error should improve clarity.

Please specify which ICC and which SEM calculations you used and why.

Measurement error needs to be considered in relation to meaningful change or clinically important differences. It is not sufficient to say the error was ‘small’.

Sample size is smaller than the recommended 50 for measurement error and reproducibility tests.

The final paragraph of the Background section is not clear. Validity appears in aim A along with non-specific ‘reliability’, then measurement error is given separately but without mention of the conditions this test is carried out for. Finally
SDC is given as a separate aim where previously it was suggested to be part of “absolute reliability”. SEM is not mentioned separately but for the first time called measurement error. The statement of study aims needs to be much clearer and directly related to the presentation of methods and results.

While the Background is generally easy to read and makes sense. It would benefit from some re-organisation to make it clearer and more concise.

Methods
In the first paragraph, the terminology changes again (“discriminative validity” and “test-retest reliability”).

Why were patients “diagnosed with benign paroxysmal positional vertigo” excluded? I found this surprising since in the Background, “positioning manoeuvres” was one of the treatments mentioned for vestibular problems which implied to me that BPPV could be part of the study group.

What do you mean by “and placed 1.72 cm on center”?

What do you mean by “parameter calculations were designed”?

“A more detailed description of the protocol can be found elsewhere [29]” The reference cited here describes the Functional Gait Assessment but does not describe procedures for testing gait with a gaitrite walking system and therefore does not appear to provide a more detailed description of the protocol used in this study, as suggested. It is not clear at this point in the paper that you have used the FGA items as conditions for the gaitrite walks. The FGA needs to be introduced in the Background and a rationale for using FGA items as the walking conditions for the gaitrite measurements needs to be provided. Also there are 10 items in the FGA and only 7 conditions tested (including one which is not part of the FGA).

The 10 minutes pause between tests seems rather short. Is there any data (pilot or otherwise) to show that a 10 minutes pause equivalent to a longer gap that is more likely to be used in practice?

What do you mean by: “Data analysis of recorded attempts on the GAITRite® system was conducted consecutively.”

“…evaluations were conducted at the same place and time for each test”. This statement seems to be confusing given it has already been stated that the tests were carried out 10 minutes apart.

Establishing validity in this way requires an hypothesis to be stated upfront including why you think the two groups will give different results on all the outcome measures, the direction of the difference and ideally an indication of the expected magnitude of the difference. Then you can discuss whether your study findings support the hypothesis, and hence the validity of the outcome measures, or not. You should also try and state an hypothesis for the subgroup analyses. Do you expect to see significant differences in the gait scores between people
with different diagnoses? How are the findings going to support or otherwise the validity of the measures?

Results

I’m not sure that the missing values (negative step lengths and refusal to perform dual task condition) should have been replaced by group average values for this study. There were unequal numbers in the groups with more in the vestibular group anyway, and the imputed values would likely have negatively distorted the reliability results.

“There were no significant differences between the five different subgroups (according to diagnosis); gait speed (p=0.194), cadence step (p=0.277), and step length (p=0.383) assessed with the Kruskal Wallis test.” Presumably this refers to just the preferred walking speed condition. Why is this information not given with the results of the other discriminative validity results?

Before concluding whether discriminative validity is supported, it would be helpful to confirm that the direction (and magnitude) of the difference were as expected (hypothesised).

The limits of agreement from the Bland Altman plots are determined from the data, therefore you cannot turn it around and say that 95% of data fell within the LOA, because by definition, 5% will be outside.

Discussion

How do you know the discriminative validity was “good” if you didn’t (a priori) set criteria for what constitutes good or otherwise?

It would be good if you could comment a bit more on what the results mean for group-level comparisons versus individual-level change.

Paragraph on Discriminative Validity: The whole paragraph is difficult to follow.

The recommendation to not use tandem walking condition for gaitrite analysis seems reasonable. However, it seems a shame that during the processing of the gaitrite trials, data from trials where two footfalls were combined as one were accepted as valid trials. I would suggest that during processing, such trials should either be manually processed so that the separate footfalls are identified as separate footfalls, or the trial is excluded as invalid.

In this study the heterogeneity of the patient population is not necessarily a problem because your reliability results will therefore be more generalizable to the broader population. However, the sample is not ideal for characterising the gait of vestibular patients in general for the reasons you stated, therefore I would not emphasise the group results but focus on the measurement properties of the outcome variables, which was your study aim.

“However, when using pressure walkways stop and go movements introduce transients in the stride trajectories that have the potential to bias variability estimates as well [43].” This sentence is difficult to follow. Do you mean variability
of gait (which is not used as an outcome variable in your study) or variability in terms of reliability/measurement error? I think you can justify using one walking trial in your study because it has practical advantages in your population of interest and your study actually aims to establish reliability/measurement error estimates for YOUR gait testing protocol. However, I think the 10-minute gap between the tests is a major limitation that should be mentioned.

Minor Essential Revisions

There are several typing errors that need to be corrected (eg. ‘tests’ instead of ‘test’, and ‘Vestibular’ instead of ‘vestibular’) so the manuscript would benefit from further proof reading.

Tables - Suggest presenting mean difference and 95% CI for differences between groups in Table 2.

Table 2 – consider using just one decimal place unless you believe your measurement system can accurately detect to 0.01 of a centimetre.

In the results section, two decimal places are probably not necessary and overestimate the precision of the gaitrite outcomes.

Terminology: self-selected or self-defined?

Minor issues not for publication

Some further editing to remove some repetition and improve clarity and conciseness would be worthwhile.

Tables - Tables are not particularly easy to read. Suggest using spacing and cell alignment to improve readability.

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

Declaration of competing interests:

I declare that I have no competing interests