Author's response to reviews

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

Authors:

Annatina Schmidheiny (a.schmidheiny@gmx.ch)
Jaap Swanenburg (jaap.swanenburg@usz.ch)
Dominik Straumann (dominik.straumann@usz.ch)
Eling D de Bruin (eling.debruin@hest.ethz.ch)
Ruud H Knols (ruud.knols@usz.ch)

Version: 3 Date: 24 July 2015

Author's response to reviews:

Dr Elaine Zhang
Journal Editorial Office
BioMed Central

Revised version: MS: 4850285901546823 Discriminant validity and test re-test reliability of a gait assessment in patients with Vestibular disorders.

24.07.2015
Dear Dr. Zhang,

We greatly appreciated the opportunity to edit and revise our manuscript 4850285901546823.

Please find enclosed a copy of the revised article for review.

With this letter we also would like to thank the reviewers for their comments and recommendations. Below we have listed the reviewers' comments and recommendations in a point-by-point reply.

We hope that this revision of the manuscript is now acceptable for publication in BMC Ear Nose and Throat Disorders.

Sincerely, also on behalf of the co-authors,

Eling D. de Bruin

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

1. 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

1a. 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

1b. 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.

Our answer 1a & 1b.

We thank the reviewer for this helpful comment. We have followed the advice and changed the terms reliability, reproducibility (test-retest), measurement error (SEM and SDC), relative & absolute reliability and hypothesis testing, according to the COSMIN taxonomy for measurement properties throughout the abstract and manuscript.

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

Our answer 2

This recommendation is well received and has been implemented in the text to read as: Several studies evaluated the psychometric properties of the GAITRite® system and demonstrated good reproducibility. Hollman et al.(24) reported excellent ICCs 2.1 (> 0.84) for velocity and cadence in older people under single and dual task walking conditions. (Page 3)

In stroke patients, test re-test reproducibility measures for the GAITRite® system were consistent with ICC (2.1) values varying from 0.72 to 0.98 (16,25).
Measurement error reflects the magnitude of the differences between two measures (26). Examples of these measures are the standard error of measurement [SEM], calculated as the square root of the error variance and the smallest detectable change [SDC] (27,28) (Page 3).

3. 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’.

Our answer 3.

We thank the reviewer for this feedback which we have adopted. It now reads in the text as: To be clinically useful, measurement error needs to be considered in relation to meaningful change or clinically important differences (29). (Page 3)

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

Our answer 4.

We adapted the comment of the reviewer. This is now to read in the Discussion (page 9): Secondly, the sample size was relatively small and may have affected the values of the reproducibility and measurement error. A sample size of at least 50 is generally seen as adequate for the assessment of the agreement parameter, based on a general guideline by Altman (1990) (55). The sample size we used of 35 patients with vestibular disorders and 27 controls is, however, a realistic group size to find first estimates for the assumed relation between vestibular disorders and gait and to identify differences between patients and healthy controls.

5. 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.

Our answer 5.

We have followed the advice and changed the background to read as:

Measurement error reflects the magnitude of the differences between two measures (26). Examples of these measures are the standard error of measurement [SEM], calculated as the square root of error variance and the smallest detectable change [SDC] (27,28). To be clinically useful, measurement error needs to be considered in relation to meaningful change or clinically important differences (29).

And in the last para to read as: In this study, the hypothesis was tested if the GAITRite® system could discriminate patients with vestibular disorders from healthy participants for the outcomes of self-selected walking speed, cadence
and step length. Based on the study of Menz, et al. (13), a magnitude of 10% or larger difference in outcomes was defined. Further, patients were evaluated twice to determine the test re-test reproducibility (ICC 2.1) and the measurement error (SEM, SDC) of walking behaviour as assessed with the GAITRite®. We conducted this study to (a) investigate the degree to which the scores of a gait analysis performed with the GAITRite® differ between patients with vestibular disorders and healthy participants, (b) identify the reproducibility of gait parameters measured with GAITRite® in patients with vestibular disorders walking under single and dual-task conditions, and (c) identify the measurement error (precision). (page 3).

6. 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.

Our answer 6.

The background section was edited to improve the clarity of the manuscript. (Page 2 and 3).

Methods

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

Our answer 7.

This paragraph was drastically shortened and revised to make it more consistent regarding the use of terminology in the remainder of the manuscript. It now reads: A cross-sectional design was chosen. (page 4)

8. 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.

Our answer 8. This is a relevant remark of the reviewer. We did not intend to include patients with benign paroxysmal positional vertigo without symptoms. Therefore, we edited the manuscript in the background to read as: The current management of vestibular disorders includes vestibular rehabilitation, pharmacological treatment, surgery, manual therapy and positioning manoeuvres for a specific diagnostic group of benign paroxysmal positional vertigo (8-10). (Page 2)

In the Methods – Patients and Participants section we wrote: “Participants were excluded when diagnosed with benign paroxysmal positional vertigo without vertigo symptoms after position manoeuvres” (Page 4)

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

Our answer 9.

We deleted this sentence from the text with the aim to improve clarity of the manuscript.
10. What do you mean by “parameter calculations were designed”?

Our answer 10.
We deleted this sentence from the text with the aim to improve clarity of the manuscript.

11a. “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.

Our answer 11a.
We acknowledge this remark from the reviewer. Because we already referred to the ‘Functional Gait Assessment’ in the paragraph ‘procedure’ we deleted the sentence: “A more detailed description of the protocol can be found elsewhere [29]” from the text.

11b. 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).

Our answer 11b.
We thank the reviewer for this important comment. The manuscript has been edited accordingly and now reads in the Background section: Gait abnormalities may be assessed with the Functional Gait Assessment (FGA), a 10-item assessment based on the Dynamic Gait Index (11). Although the FGA is a practical and functional assessment tool, it does not quantify temporal and spatial gait parameters beyond a sum score. Quantification of gait parameters while performing the FGA would, however, add more sophisticated information to a gait assessment. (Page 2)

and

Based on the FGA, a gait protocol was developed to evaluate 7 of 10 FGA tasks with the GAITRite® system. (Page 3)

and in the Methods section:

The gait protocol included the following FGA tasks: [1] self-selected walking speed, [2] gait with horizontal head turns, [3] gait with vertical head turns, [4] gait with narrow base of support (with tandem steps), [5] gait with closed eyes, [6] walking backwards. In addition we tested gait with a dual tasking paradigm (counting backwards in steps of 7 from 100). The latter task was added as we expected differences in temporal and spatial gait parameters between patients and healthy participants (39). The FGA tests: Change in gait speed, gait and
pivot turn, Step over obstacle and steps on stairs where not recorded as it was deemed not feasible or useful to be measured with the GAITRite® system. (Page 5)

12. 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?

Our answer 12.

The manuscript section has been edited and now reads: Fifthly, the short time break of 10 minutes between measurements could influence the reproducibility and measurement error data in this study. However, one study reported good to excellent ICC’s (0.87-.097) for self-selected walking speed, cadence and step length and SEM’s with a 15 minute break between measures (57). Furthermore, the internal consistency of the FGA as determined with Cronbach alpha is with 0.79 (11) rather good, which further indicates that no behavioural response in gait is to be expected when gait is measured. Currently, there is no standardization for an optimal time break between reproducibility measures. Thus, researchers and clinicians have to choose an optimal time frame when designing a study. (Page 10)

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

Our answer 13.

This was edited and now stated as: Data processing The recorded measurements were analysed immediately after each walking attempt on the GAITRite® system. (Page 5)

14. “…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.

Our answer 14.

This was edited and now stated as: Further, in order to minimize environmental variability, walking evaluations were conducted in the same hallway for each test. (Page 5)

15a. 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.

Our answer 15a.

This important comment was included in the background section of the manuscript, which was edited accordingly and is now to read: “When a novel instrument is introduced for clinical use in a patient population it is important to
evaluate the degree in which scores of different relevant groups deviate with a feasible measurement protocol (18). Thus; publication of study results will establish the robustness of an assessment. We hypothesised that gait assessed with the GAITRite® system would reveal differences for self-selected walking speed, cadence and step length between patients with vestibular disorders and healthy age-matched adults (13).”(Page 2)

15b. Then you can discuss whether your study findings support the hypothesis, and hence the validity of the outcome measures, or not.

Our answer 15b.
This reads in the discussion section as: The results of this study suggest that the walking assessment protocol performed on the GAITRite® system yields good discriminant validity between patients with vestibular disorders and healthy participants. (Page 7).

15c. 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?

Our answer 15c.
The information related to the subgroup analysis was deleted from the manuscript because of the small sample size.

Results
16. 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.

Our answer 16.
Thank you for this comment. We recalculated the discriminant validity and test-retest reproducibility results using only the complete datasets. (see Table 2 and 3). Additionally, the following was added to the text in the results section;
The measurements of three patients yielded invalid values for negative step length and could not be used for the reproducibility analysis. Therefore, the analyses of 32 patients were performed for self-selected walking speed (see table 3). Furthermore, for discriminant validity and reproducibility, one patient could not perform the task ‘walking under dual-task conditions, as he was afraid to lose balance or fall. The analyses were, therefore, performed with 34 patients (Tables 2 and 3). (Page 6).

17. “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?

Our answer 17.

As stated earlier, the subgroup analysis was deleted from the manuscript. The discussion based on a sub-group analysis would be too hypothetically.

18. 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).

Our answer 18.

We thank the reviewer for this important comment. The manuscript was edited and now reads:

Data for temporal and spatial gait parameters (gait speed, cadence and step length) and the specific walking conditions are presented in Table 2. For patients with vestibular disorders, the values in the 6 different walking tasks for gait speed varied from 0.7 to 1.2 m/s, for cadence from 88 to 108 steps/min, and for step length from 40 to 64 cm. See table 2 for the mean values across all walking tasks. For healthy control subjects, data for gait speed varied from 1.0 to 1.4 m/s, for cadence from 101 to 115 steps/min, and for step length from 54 to 73 cm. All tests showed significant differences between the two groups (p#0.01) with differences generally above 10% between patients and healthy participants for self-selected gait speed and step length. (Page 7).

There are basically two ways to achieve the goal of contrasting different groups without inflating the Type I error rate: [A] to break down the variance accounted for by the model in component parts, and [B] to compare every group (as if conducting several t-tests) but to use a stricter acceptance criterion such that the family-wise error rate does not rise above .05. The first option (A) can be done using planned comparisons (also known as planned contrasts) whereas the latter option (B) can be done using post hoc comparisons. The difference between planned comparisons and post hoc tests can be likened to the difference between one- and two-tailed tests in that planned comparisons are done when you have specific hypotheses that you want to test, whereas post hoc tests are done when you have no specific hypotheses.

As can be seen from our submission we hypothesized about differences (non-directional) between groups we expected. This, furthermore, explains why we tested with a two-sided focus. For clarity we have slightly modified the introduction where we give our hypotheses. We now write our hypotheses in a more explicit form and in more detail.

19. 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.

Our answer 19.

The manuscript is edited accordingly and is now to read as: Most of the data
were between 2 standard deviations in the Bland-Altman plots, with the exception of a few outliers (1-2) in gait speed and cadence for self-selected walking speed. The Bland and Altman plots for step length yielded 4 data points outside the 2 standard deviations. The Bland-Altman plots for gait speed, cadence, and step length of the task self-selected walking speed for patients are illustrated in Figure 1. (Page 7)

Discussion

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

Our answer 20.

Schniepp et al. (17) determined the variability of gait parameters using the GAITRite® system in patients with cerebellar ataxia, patients with vestibular disorders and healthy participants. Self-selected walking speed for healthy participants was 1.11 ± 0.19 m/s, for cerebral ataxia 1.0 ± 0.2 m/s and for patients with bilateral vestibular disorders 1.0 ± 0.2 m/s, indicating a difference of approximately 10% between healthy participants and patients with vestibular disorders. (Page 2).

And in the Discussion

In this study, an approximate 10% difference between patients and healthy controls in favour of healthy controls was observed in normal walking for the mean values of the parameters gait speed, cadence and step length. Differences for GAITRite® parameters between patients and healthy participants with a comparable magnitude were also measured in a report on individuals with cerebellar ataxia (49), patients with unilateral peripheral vestibular loss [UVL] and patients with bilateral peripheral vestibular loss [BVL] (40). (Page 7-8)

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

Our answer 21.

The manuscript was edited in the discussion and now reads:

Parameters of measurement error will be more stable over different population samples than reproducibility parameters. Reproducibility parameters are highly dependent on the variation in the population sample and are only generalizable with samples of a similar variation. It is clearly a characteristic of the performance of an instrument in a certain group sample. Measurement error is more a characteristic of the measurement instrument itself. Measurement error parameters are preferable in all situations in which the instrument will be used for evaluation purposes, which is often the case in medical research in a clinical setting. Researchers and clinicians should be eager to apply and interpret the parameters of measurement error (on an individual level) and reproducibility (on a group level) correctly (44). (Page 8)

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

Our answer 22.

The paragraph was edited and is now to read as:

**Discriminant validity**

In this study, an approximate 10% difference between patients and healthy controls in favour of healthy controls was observed in normal walking for the mean values of the parameters gait speed, cadence and step length. Differences for GAITRite® parameters between patients and healthy participants with a comparable magnitude were also measured in a report on individuals with cerebellar ataxia (49), patients with unilateral peripheral vestibular loss [UVL] and patients with bilateral peripheral vestibular loss [BVL] (40). The small sample size and the heterogeneity of our sample (central vestibular dysfunction, M. Menière, and vestibular migraine) may have led to these differences. Conversely, other studies did not show significant difference for the parameters gait speed and cadence of self-selected walking between BVL patients and healthy controls (14,17,50). However, the differences between groups found in this study are supported by small SDC values that were smaller than the differences found between patients and healthy adults (the exception being tandem walking), thus, indicating good discriminant validity of the protocol. (Pages 7-8)

23. 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.

Our answer 23.

The manuscript was edited in the results and now reads:

For the parameter tandem walking, the GAITRite® software had considerable difficulty automatically detecting footfalls. Human intervention was required to process the data from two footfalls to one. As this is clinically not feasible, the measurements were declared invalid and not presented in the manuscript and tables. (Page 6)

24. 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.

Our answer 24.

We deleted the sentence: “and of subgroups that included several Vestibular diagnoses) was not homogenous and too small for subgroup analysis.” from the manuscript and discussed the measurement properties of the Gaitrite to read in
the discussion as: Firstly, the lack of a standardized measurement protocol for the GAITRite® system limits the interpretation of gait variability from evaluative and prognostic studies. The differences found in our study between patients with vestibular disorders and healthy subjects does not in itself prove validity for the GAITRite® system, however rather contributes to the evidence for or against validity which is an ongoing process. Further research is needed to standardize testing procedures and establish validity, reproducibility and measurement error for confident use of GAITRite® walking system in patients with vestibular disorders. (Page 9).

25. “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.

Our answer 25.

The manuscript was edited and now reads: However when using pressure walkways, stop and go movements introduce transients in the stride trajectories that have the potential to bias variability in terms of reproducibility and measurement error (52). (Pages 9-10)

Furthermore, a fifth paragraph was added concerning the 10-minute interval time. This is to read as: Fifthly, the short time break of 10 minutes between measurements could influence the reproducibility and measurement error data in this study. However, one study reported good to excellent ICC’s (0.87-.097) for self-selected walking speed, cadence and step length and SEM’s with a 15 minute break between measures (57). Furthermore, the internal consistency of the FGA as determined with Cronbach alpha is with 0.79 (11) rather good, which further indicates that no behavioural response in gait is to be expected when gait is measured. Currently, there is no standardization for an optimal time break between reproducibility measures. Thus, researchers and clinicians have to choose an optimal time frame when designing a study. (Page 10)

Minor Essential Revisions

26. 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.

Our answer 26.

This was corrected throughout the manuscript.

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

Our answer 27.
Mean difference and 95% CI are added to Tables.

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

Our answer 28.
The outcomes in table 2 are presented with 1 decimal.

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

Our answer 29.
The outcomes in the results are now presented with 1 decimal.

30. Terminology: self-selected or self-defined?

Our answer 30.
Self-defined was changed in self-selected throughout the manuscript.

Minor issues not for publication
31. Some further editing to remove some repetition and improve clarity and conciseness would be worthwhile.

Our answer 31.
The manuscript was edited accordingly in all sections.

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

Our answer 32:
The tables were improved

Reviewer’s report
Title: Discriminant validity and test re-test reliability of a gait assessment in patients with Vestibular dysfunction
Version: 2 Date: 13 June 2015
Reviewer: Robin Criter

Reviewer’s report:
Minor Essential Revisions:
1. Page 2, Line 77: Please do not use abbreviations such as “e.g.” in manuscript text unless in parentheses. (Also Page 4, Line 148.)
Our answer 1.
The abbreviations have been deleted from the text as suggested.

2. Page 3, Methods – Patients and participants: Age is discussed later on as a possible confounding factor when comparing the patient and healthy control groups. I would like to see age range reported for both groups.

Our answer 2.
The age range is added in Table 1

3. Page 3, Line 117: Please change “Data was uploaded…” to “Data were uploaded…”.

Our answer 3.
The manuscript was edited accordingly.

4. Page 4, Line 143: Please change “…test re-tests…” to “…test re-test…”

Our answer 4.
The manuscript was edited accordingly.


Our answer 5.
The manuscript was edited accordingly.

6. Page 8, Line 291: Please change “trail” to “trial”.

Our answer 6.
The manuscript was edited accordingly.

7. Tables: I found the tables difficult to read (extra spaces, uneven rows and columns); however, this could have been due to the conversion to pdf

Our answer 7.
The tables were improved.

Major Compulsory Revisions:

8. Page 5, Results: I would like to know more about how the patients were classified as having a Vestibular disorder (i.e., what tests or criteria were used, who made the diagnosis).

Our answer 8
The manuscript was edited and now reads:

Eligible patients were consecutively selected from July 2013 to February 2014 by the leading physician of the Department of Neurology. All patients reported chronic deficits > 6 month duration. Tests used for diagnosis were performed by
the horizontal head impulse test to both sides (search coil technique or a video based system) (34-36) and/or by video-oculography (caloric response absent or bilaterally diminished <5 deg/sec or absent (37).

Reviewer's report
Title: Discriminant validity and test re-test reliability of a gait assessment in patients with Vestibular dysfunction
Version: 2 Date: 16 June 2015
Reviewer: Avril Mansfield
Reviewer's report:
Overview
This is a generally well-written paper that aims to determine the test-retest reliability of instrumented gait assessment with the Gaitrite walking among individuals with Vestibular dysfunction. This is important foundational work required prior to using this test protocol in a clinical setting or in future research (e.g., to evaluate the effect of an intervention on characteristics of walking among individuals with Vestibular dysfunction).

Major compulsory revisions
1. The authors must specify in the methods section that only one pass was completed per condition. This is not stated until much later in the paper (limitations, line 321). Additionally, the authors must provide further justification for only including one pass per condition. It is likely that reliability could have been enhanced by averaging data from several passes – the authors had the opportunity to determine if one pass was sufficient or if an average of multiple passes should be taken and I believe failure to do so is a major limitation of the current work.

Our answer 1
We fully agree with the reviewer on this point. The manuscript was therefore edited and now reads in the method section: The patients and healthy subjects performed one walking trial for 7 protocol walking conditions. (Pages 4-5)

and in the discussion:
Contrarily, in a clinical setting time and resource constraints often prevent performance of extensive measurement protocols. Furthermore, our patients performed their measures during clinical visits to the University Hospital and we did not want to oblige them with a stressful program. (Page 10)

2. Further explanation is required regarding how the Gaitrite data were processed. It appears that the authors accepted the automatic footfall identification performed by the Gaitrite system (lines 117-118) with the exception of manually removing some footfalls (line 153). However, for two conditions (tandem and backwards walking) I suspect the Gaitrite software would have had
considerable difficulty detecting footfalls automatically (e.g., line 299-300) and human intervention would have been required to process the data. When such intervention is required, rater reliability cannot be assumed and, therefore, rater error or judgement may have influenced the results of the current study (e.g., see Wong et al., Gait & Posture, 2014).

Our answer 2.
We agree with the comment of the reviewer. The manuscript is edited in the results and now reads.
For the parameter tandem walking, the GAITRite® software had considerable difficulty automatically detecting footfalls. Human intervention was required to process the data from two footfalls to one. As this is clinically not feasible the measurements were declared invalid and not presented in the manuscript and tables.(Page 6)

3. What was the rationale for replacing missing data with the mean (line 198-199)? Consider just removing these data points from the analysis.

Our answer 3.
The data points were deleted from the analyses. Reproducibility and measurement error outcomes including Bland & Altman plots for selected walking speed (Gait speed, Cadence, step length) were recalculated and presented throughout the manuscript.

Minor essential revisions
3. The authors must specify in the introduction that they are interested in test-retest, rather than rater, reliability.

We thank the reviewer for her comment. We have followed the advice and changed the terms reliability, reproducibility (test-retest), measurement error (SEM and SDC), relative & absolute reliability and hypothesis testing, according to the COSMIN taxonomy for measurement properties throughout the abstract and manuscript.

4. m/s is the preferred unit for walking speed.

Our answer 4.
The manuscript and tables were changed accordingly.

5. Lines 208-210: clarify that these are the ranges in mean values across all conditions.

Our answer 5.
This was stated in text as:
See table 2 for the mean values across other walking conditions.