PEER REVIEW HISTORY

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

ARTICLE DETAILS

TITLE (PROVISIONAL)	Test-retest reliability of the twenty-five-hole peg test in stroke patients
AUTHORS	Granström, Fredrik; Hedlund, Mattias; Lindström, Britta'; Eriksson, Staffan

VERSION 1 – REVIEW

REVIEWER	Benjamin Mentiplay La Trobe University, Australia
REVIEW RETURNED	02-Aug-2019

GENERAL COMMENTS	This study examines a new test of upper limb function and manual dexterity in people following stroke. The study extends the existing nine hole peg test to examine a twenty-five hole peg test. This study is well written. I have some concerns that I have described below to improve the manuscript. Abstract Line 6: It is not clear (in the abstract at least) why it is easier to count pegs in the 25 hole test compared to 9 hole test. You explain in the Introduction the issues with the 9 hole test in terms of floor and ceiling effects so I suggest including this information in the abstract. Introduction Line 20-21 (page 4): Can you provide a brief summary of the reliability statistics in these two previous studies? Has the nine hole test been shown to be an important measure post-stroke? As in, does it have a relation with other functional measures or is it predictive of recovery or anything else? I think this information would be important to include. Line 26 (page 4): I am not sure that this requires its own paragraph. Most readers would have an idea of what reliability is; however, they me be unfamiliar with the SRD, which I imagine could be defined in the Methods section. Line 45 (page 4): How long is the stipulated time? Line 47 (page 4): I suggest rewording the following part of the sentence: 'great majority useful to test'. Line 3-26 (page 5): Can you confirm that this test has not been used previously in the literature? Why is 50 seconds used as the time limit? If this is the first time this test has been used in research, I would suggest including a picture of the test to enhance clarity for the readers.

	Methods Line 22 (page 6): The removal of 0/1 or 25 scores is an interesting decision, and I think needs further consideration/justification. Line 35 (page 6): The reference number 6 you have used here – did this study examine the 25 peg test? You mentioned in the Introduction that no study had examined this before. Also, this paragraph would be nice to include a figure of the board, box and pegs. Line 50 (page 6): what was the other test? Line 56 (page 7): I suggest rewording this sentence – are you saying that you did not collect the dominant hand and the Fugl-Meyer test information? Line 3 (page 8): this is a very lengthy statistical analysis section – I suggest trying to be more concise with the information in this section to enhance readability. I am also concerned with the number of statistics generated for a relatively small sample. Would any of your statistics need to be adjusted for multiplicity? Results Table 1 footnote: suggest spelling out SD Table 1: provide a unit of measurement for the TFHPT. I also suggest replacing 'min-max' with 'range'. Line 3 (page 12): I think you need to quantify the 'slight association' that you mention, or provide further explanation. Discussion Line 30 (page 14): what does the reference at the end of this sentence indicate? It seems you are reporting results of your studies so it is unclear why this reference is needed. Line 20 (page 16): But you removed this 0 result from your analysis. Limitations: You have reported a new test – but only examined one component of its properties (i.e. reliability). You have not examined other psychometric properties of the test and I would suggest acknowledging this further in your limitations. Not only have you not examined day-to-day variation, but you have not examined other properties such as validity, sensitivity or specificity.

REVIEWER	I-Ping Hsueh National Taiwan University Taiwan, Republic of China
REVIEW RETURNED	24-Aug-2019

GENERAL COMMENTS	Since test-retest reliability is a prerequisite for clinicians and

researchers to appropriately interpret change scores in repeated assessments, this study shows good contribution for measurement of manual dexterity. In general, this study shows two advantages: (1) the test-retest reliabilities of a single score and averaged scores for 3 trials were examined simultaneously, which can be helpful to select a more reliable index of manual dexterity; and (2) the SRD is a practical index of random measurement error, which can be useful in determining whether examinees' changes are real or just the consequences of measurement errors. However, this study also has 3 weaknesses: (1) the relatively small sample size (*n* = 31); (2) the short intervals (only a few minutes) with test-retest assessments; and (3) the readability of this manuscript may not be satisfactory. Thus, some efforts would be needed before it can be acceptable and published.

For the whole manuscript.

1. The 'comma' appears to be used inappropriately in the whole manuscript. Specifically, lots of commas were lost, which may have hampered the readability.
2. Improvement of readability for most sections is needed, especially for the introduction, results, and discussions. The specific suggestions have been provided in the corresponding sections.

Introduction

3. In page 4, lines 14 to 16 ("*However, only approximately 25% of studies regarding upper limb interventions include a specific measure of manual dexterity.[1]…*"), the relationships between "low proportion of studies included measures of manual dexterity" and "the TFHPT lacks validations of its test-retest reliability" are unclear. In particular, the lacks of evidence for test-retest reliability appears to be the main rationale of this study.
4. In page 4, lines 26 to 33 ("*Reliability is a term that describes how the result of a measurement with an instrument is affected by measurement error.[4] The concept of absolute reliability refers to the consistency of measurements within individuals and can be quantified by, for example, the smallest real difference (SRD) and by SRD%.[5]*"), two issues may be concerned.
   i. The concept of reliability may be inappropriate to be addressed here. Specifically, it would disrupt the linking between the previous and the next paragraphs, hampering the readability.
   ii. The SRD and SRD% may not be the index of reliability. Since the concepts of SRD and SRD% are more close to the "amount of variations due to measurement error", the SRD and SRD% may be more appropriate to be indicators of measurement error but not reliability.
5. In page 4, lines 36 to 52 ("*A weakness with the NHPT is that many persons with stroke cannot reach the lower limit; i.e., a floor effect arises. Furthermore, because there are only nine pegs, measures must be taken to avoid ceiling effects. Therefore, in the original test, the result is expressed as the time to complete the test, including inserting and removing all of the pegs.[2, 3, 6, 7] This, however, aggravates the floor effects because tests that are not completed during the stipulated time are excluded.[2, 3] The*

maximum time could be prolonged to include the great majority useful to test. However, this would be time consuming and possibly unethical due to the possibility of a non-completed test after a lengthy attempt..."), two issues may be concerned.

i. The rationales of validating the TFHPT (but not the NHPT) are not clear. If the modified NHPT has successfully overcome the problems of floor and ceiling effects, the remaining problem is lacks of evidence for its reliability.
However, the straightforward solution for unknown reliability would be validating it, but not validating another measure (i.e., the TFHPT).

ii. Readability of this paragraph can be improved. For example, the purposes of using the time of completing the test as an index can be strengthened (e.g., "to minimize ceiling effect, the time of completing the whole test was suggested."). In addition, the advantages and disadvantages of every version of the modified NHPT can be described in a clearer way.

6. In page 4, lines 52 to 56 ("*A modified NHPT is used to mitigate the floor effect while avoiding the ceiling effect, in which the result is expressed as the number of inserted pegs (not removed) per unit of time, i.e., the frequency…*"), why using the "number of inserted pegs" as an index can mitigate the floor effect? Further clarification would be needed to ensure readers can fully understand it.
7. In page 5, lines 9 to 12 ("*Thus, the TFHPT measures the motor function on a numerical scale, with low floor effects and reasonable ceiling effects*"), three issues may be concerned.

i. The original NHPT, if the number of inserted pegs was adopted, can also provide the numerical score. Thus, this advantage may not be unique for the TFHPT.

ii. References for the "low floor effect and reasonable ceiling effect" could be added to support that the TFHPT is a promising measure to validate.

iii. Please be consistent for the terms used in the whole manuscript. For example, the "upper limb assessment", "fine manual dexterity", and "motor function" stated in the first 3 paragraphs in the Introduction section seem to be interchangeable. However, their scopes and meanings may be different and thus, may result in confusion.

8. In page 5, lines 24 to 26 ("*…However, since the tests have completely different stop criteria – a time limit for the TFHP vs. all pegs inserted for the NHPT – equal reliability cannot be taken for granted.[6]…*"), the "TFHP" appears to be a typo.
9. In page 5, lines 29 to 34 ("*Measurements with the TFPHT and quantification on a numerical scale and can be used on a large portion of persons suffering from stroke. Thus, depending on the magnitude of measurement error, this test may be useful, both in clinical practice and in research.*"), three issues may be concerned.

i. Please remove the duplicated information (e.g., the numerical scale) that had mentioned in the other paragraphs.

ii. Please provide references for accessibility of the TFPHT in patients with stroke.

iii. The meanings of "*depending on the magnitude of* *measurement error*” is unclear.

Methods

10. In page 8, lines 8 to 11 (“*…The exception was the Bland-Altman plots, for which only trials two and three were used…*”), why the Bland-Altman plots were provided only for comparisons between the trial 2 and trial 3?
11. In page 8, lines 16 to 25 (about the Bland-Altman plots and heteroscedasticity), since heteroscedasticity can also be examined by checking whether the Pearson correlation coefficient for the absolute value of differences between two assessments and the average score exceeded 0.3, such method may be adopted to provide objective judgment of the heteroscedasticity.
12. In page 8, lines 28 to 44 (*“Measurement error can be either random or systematic. In random error, there is no pattern to the variability, whereas in systematic error the measurement varies in a non-random way, i.e., the mean values between the trials differ.[5] To investigate whether there was a systematic error in test scores, one-way repeated measures ANOVA was used to detect potential between trial effects. Fisher’s LSD post hoc tests between trials were performed when the main effect for trials was significant. The assumption of sphericity was met in all trials according to Mauchly’s test, whereas the assumption of normal distribution was violated in trial two’s pre-intervention according to the Kolmogorov-Smirnov test (p=0.043)”*), the descriptions about systematic error seem to be skipped.
13. In pages 8 to 9, lines 54 to 5 (“*…This makes ICCs sensitive to the degree of between-subject variability, and with all other things being equal, a more heterogeneous sample will produce higher ICC values. [5, 11] In addition, it is difficult to draw statistical inference from one sample to another.[12]…*”), two issues may be concerned.

    i. Thought the aforementioned description is correct for ICC, however, since the “within-subject variability” is the main target for examining test-retest reliability, such description appears not straightforward.
    ii. The relationships between the between- and within-subjective variabilities in ICC models are not clear in the article. Thus, the impacts of heterogeneous sample on ICC value are not easy to understand. Please consider to provide some explanation.

14. In page 9, lines 3 to 5 (“…*In addition, it is difficult to draw statistical inference from one sample to another.[12]*”), the statement may not be appropriate because if the statement is true, ICC seems not a suitable method for test-retest reliability and thus, should not be used in this study.
15. In page 9, lines 8 to 11 (*“Three separate measures of relative reliability including 95% confidence intervals (CIs), namely, ICC2.1, ICC2.3, and ICC3.3, were calculated…”*), two issues may be concerned.

    i. This statement may misguide the readers because thought the 95% C.I.s were provided for the 3 ICC values, but these

C.I. values were not named ICC(2,1), ICC(2,3), and ICC(3,3).

ii. The rationales for using ICC(2,1) remain unclear. If the authors aimed to examine systematic error for the single score, the ICC(3,1) would also be included.

16. In page 9, lines 15 to 29 ("*…The first figure in the ICC designation represents the type of intraclass correlation coefficient (ICC model), while the second figure represents single or average measures, where "1" represents single measures and "2" or higher represents the number of trials from which the average is calculated.[5] ICC 2.1 and ICC2.3 are calculated from a two-way random effect model and incorporate both systematic and random error, whereas ICC3.3 is calculated from a two-way fixed effect model and incorporates only random error.[5, 13]…*"), two issues may be concerned.

i. Readability of this paragraph can be improved. Specifically, introductions of a specific topic can be provided just after the concept was mentioned. For example, since the authors addressed that "the first figure in the ICC designation represent the models of the ICC", the explanations of "the number 2 indicated the two-way random model, which incorporated both systematic and random error; whilst the number 3 represented the two-way fixed effect model, which considered only the random error." can be addressed.

ii. Abbreviations (e.g., ICC) should be used for the whole text since they have been defined for the first time.

17. In page 9, lines 46 to 48 ("*To estimate the absolute reliability, the standard error of measurement (SEM), SRD and SRD percentage (SRD%) were calculated…*"), concepts of the "relative reliability" and "absolute reliability" can be provided to improve the readability.

18. In page 9, lines 53 to 60 ("*…SEM is the within-subject standard deviation calculated from repeated tests.[11, 14] The variation of repeated tests can be thought of as the error around a true value; concomitantly, the within-subject standard deviation is used as a measure of measurement error.[14]…*"), these descriptions can be integrated with the introductions of ICC to achieve a top-down structure, which tends to easy to understand for most readers.

Results

19. In pages 12 to 13, lines 28 to 34 (for both paragraphs about results), two issues may be concerned.

i. Please summarize the main findings of this study because conceptually, descriptions in the results section should not be duplicated to those in the tables.

ii. In addition, please remove the redundant explanations (e.g., for the single measures) when reporting the main findings to remain sentences clear and concise.

Discussions

20. In pages 14 to 17, for all paragraphs in the discussions section, please put the similar findings in the same paragraph and targeting only one issue in a single paragraph. For example, if the reliability of a single score and the average scores was the main issue, then the findings of repeated ANOVA (systematic error) would be moved to other places. By doing so, the readability of the current discussions section can be substantial imrproved.
21. In page 14, lines 16 to 21 ("*This study indicated that in a selected group of persons suffering from stroke, the use of an average of three trials reduced the measurement error substantially compared to a single trial (SRD2.3 vs SRD2.1).*"), two issues may be concerned.

    i. Since the main purposes of this study were not reducing the measurement error, describing the main findings in this way appears to misguide readers.
    ii. The authors seem to suggest the users to adopt the average scores rather than a single score. However, such suggestions seem to be stated indirectly, which may not be clear for most readers.

22. In page 14, lines 46 to 51 ("*…This limitation arises because the random error of measurements often increases with the magnitude of the measurements (i.e. heteroscedasticity), [11 12] which also, to some extent, was evident in this study (Figure 2a)….*"), the meanings of this sentence are unclear.
23. In pages 14 to 15, lines 60 to 5 ("*Thus, to capture the systematic error and express the absolute reliability as an absolute number of pegs representing an average of three trials, the most accurate measure investigated in this study for assessing the absolute reliability of the TFHPT is SRD2.3*"), the conclusion cannot be drawn from the previous sentence. Specifically, since the previous sentence only indicated that heteroscedasticity (stability of random error for patients with different levels of manual dexterity) may not be a major concern for the TFHPT, concluding that the SRD(2,3) was suggested to capture the systematic error are confusing. Accordingly, the flow of this paragraph needs to be reorganized.
24. In page 15, lines 36 to 41 ("*…Even though the SRD% measures reported in the studies by Chen et al. and by Ekstrand et al. were calculated in different ways compared to the SRD%2.3 reported in this study, the measures used in these three studies are fairly equivalent.[5]…*"), it would be better to specify the differences between the two methods and the possible impacts on the findings.
25. In pages 15 to 16, lines 55 to 8 ("*…Second, in this study of the TFHPT, the test and retest trials were performed within minutes compared to within days in the studies of the NHPT, which may have resulted in seemingly worse reliability for the NHPT because of possible random error from day-to-day variation in performance.[11, 16] Third, the 3-5 days between test and retest trials in the study by Chen et al.[2] may also have resulted in seemingly worse reliability in that study because of systematic error…*"), the impact of differences in the findings are unclear. To clarify it, the authors may describe the impacts of the findings directly (e.g., the test-retest reliability in this study may have been overestimated); or comparing the results with

the previous studies (e.g., yielding higher test-retest reliability of the TFNPT than the NHPT).

26. In page 16, lines 13 to 22 ("*…One advantage with the TFHPT, compared to the NHPT, is that persons with worse motor function can be tested.[2, 3] In the study by Ekstrand et al.[3], those who did not complete the NHPT in 180 seconds were excluded. This would correspond to inserting and removing a minimum of 2.5 pegs in 50 seconds, whereas 0 pegs inserted in 50 seconds is a valid result with the TFHPT…*"), since the NHPT can have the similar utility to the TFHPT if the index of number of inserted pegs is used, this may not be the unique advantage for all versions of the NHPT. If the authors aimed to strengthen this advantage, the specific versions of the NHPT should be mentioned.
27. In page 16, lines 46 to 48 ("*…The cause of the decreased random error is less clear, but it could also be attributed to the decreased systematic error.[11]…*"), why the reduction of random error is correlated to that of the systematic error? More information would be needed to address it.
28. In pages 16 to 17, for the paragraph about limitations, please specify the number of limitations that would be addressed at the beginning. Such information can be helpful for readers to capture the whole picture of limitations.

Conclusion

29. In page 17, lines 27 to 31 ("*…In conclusion, our results suggest that the smallest difference that can be detected using a test procedure with an average of three trials (SRD2.3) conducted by a single tester should be just above 2 pegs with the TFHPT…*"), the actual value of SRD appears to be more important here rather than the code of the ICC model adopted. Thus, why not directly point out the changes of 3 pegs between two assessments may indicate a real change.

Others

30. Ceiling and floor effects appear to be the main reasons for using the TFNPT. However, they were not examined in this study, remains whether this measure can solve the original problem unknown. The rationales of this study need to be revised.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Benjamin Mentiplay
Institution and Country: La Trobe University, Australia
Please state any competing interests or state 'None declared': None declared

This study examines a new test of upper limb function and manual dexterity in people following stroke. The study extends the existing nine hole peg test to examine a twenty-five hole peg test. This study is well written. I have some concerns that I have described below to improve the manuscript.

Abstract
1. Line 6: It is not clear (in the abstract at least) why it is easier to count pegs in the 25 hole test compared to 9 hole test. You explain in the Introduction the issues with the 9 hole test in terms of floor and ceiling effects so I suggest including this information in the abstract.

Answer: Now we have tried to explain this in the abstract. However, in the very limited space, in the abstract, it is not easy to explain the advantages with the TFHPT. We had to delete some of the less important information under the design, method, result and conclusion sections, and still the advantages with 25-holes, as we see it, is not fully explained.

Introduction
2. Line 20-21 (page 4): Can you provide a brief summary of the reliability statistics in these two previous studies?

Answer: Yes, this is now incorporated in the introduction, *page 4, line 73*.

3. Has the nine hole test been shown to be an important measure post-stroke? As in, does it have a relation with other functional measures or is it predictive of recovery or anything else? I think this information would be important to include.

Answer: We have now incorporated information on correlation to other measures of hand function (i.e. criterion validity), *page 4, line 73-77*.

4. Line 26 (page 4): I am not sure that this requires its own paragraph. Most readers would have an idea of what reliability is; however, they me be unfamiliar with the SRD, which I imagine could be defined in the Methods section.

Answer: We think you make a very good point. All information on reliability is now removed from the introduction. However, some of the information on reliability is not removed entirely, but moved to the method section, to increase the readability of that section. You bring up an important question when you mention the SRD. However, one consequence of removing the section about reliability, is that we now mention SRD in our aim, without having explained it in the introduction.

5. Line 45 (page 4): How long is the stipulated time?

Answer: Sixty seconds and 180 seconds have been used, this is now incorporated in the introduction, *page 4, lines 84-85*.

6. Line 47 (page 4): I suggest rewording the following part of the sentence: 'great majority useful to test'.

Answer: Yes, this is now changed, *page 4, line 85-86*.

7. Line 3-26 (page 5): Can you confirm that this test has not been used previously in the literature?

Answer: To our knowledge it has not been used in the litterature. We have performed a search in Pubmed and a librarian performed a separate search without any trace of it (Pubmed, Pedro and Cinahl).

8. Why is 50 seconds used as the time limit?

Answer: We made a reasonable estimate of a time limit that would give the largest variability between subjects, i.e. allow as many as possible to insert at least one peg, and, at the same time, prevent as many as possible to reach 25 inserted pegs. For this assessment we had access to data on NHPT from patients performing Constraint induced movement therapy. We had data on more than a few patients, but I don´t remember the exact number. We also tested a subject with an intact nervous system in this process.

9. If this is the first time this test has been used in research, I would suggest including a picture of the test to enhance clarity for the readers.

Answer: Yes, that is a very good suggestion. It is now included on *page 6, line 136*.

Methods

10. Line 22 (page 6): The removal of 0/1 or 25 scores is an interesting decision, and I think needs further consideration/justification.

Answer: Because these two are wider intervals, measurements at them should be more stable compared to other intervals. A person can be far under the floor or high over the ceiling and measurements of such persons would be very stable, i.e. rendering a very small measurement error, or, no measurement error at all. Hence, including tests, at those two intervals, in the study/analysis would most likely yield a better reliability for TFHPT than the "true" reliability, i.e. the reliability would be overestimated. We have now explained this better in the manuscript with an additional sentence, *page 6, lines 131-132*.

11. Line 35 (page 6): The reference number 6 you have used here – did this study examine the 25 peg test? You mentioned in the Introduction that no study had examined this before.

Answer: Thanks, you are very observant, there should be no reference there.

12. Also, this paragraph would be nice to include a figure of the board, box and pegs.

Answer: Yes, we provide a picture (photo), *page 6, line 136*.

13. Line 50 (page 6): what was the other test?

Answer: BL-motor assessment. We have now included this in the manuscript, *page 7, line 142.*

14. Line 56 (page 7): I suggest rewording this sentence – are you saying that you did not collect the dominant hand and the Fugl-Meyer test information?

Answer: No, we collected data on the dominant hand and the Fugl-Meyer test, but the staff at the clinics did not. We have now rewritten this sentence in an easier way, however, with less information, *page 8, line168*.

15. Line 3 (page 8): this is a very lengthy statistical analysis section – I suggest trying to be more concise with the information in this section to enhance readability. I am also concerned with the number of statistics generated for a relatively small sample. Would any of your statistics need to be adjusted for multiplicity?

Answer: This section is now substantially rewritten in a more efficient way and to give a top-down structure. However, as we moved some information from the introduction (i.e. reliability) to this section and also incorporated a test on heteroscedasticity we also had to remove some less important information to make the section shorter. The section is now shortened by approximately 60 words*. Starts at page 8, line 170*.

We removed information on the construction on the Bland-Altman plots and the ANOVA-statistics. The Bland-Altman plots are now less important (the tests of heteroscedasticity gives the most of the information the plots were intended to give) and the information can still be found in the figure legends. The following two sections has been removed: 1) The centre line displayed the mean difference for the group between trials three and two. The upper and lower confidence limits were calculated as the mean difference ± standard deviation (SD) of the mean difference × 1.96. 2). 2)

Fisher’s LSD post hoc tests between trials were performed when the main effect for trials was significant. The assumption of sphericity was met in all trials according to Mauchly’s test, whereas the assumption of normal distribution was violated in trial two’s pre-intervention according to the Kolmogorov-Smirnov test (p=0.043).

Adjustment for multiplicity is of interest in relation to hypothesis testing (significance testing), and then to avoid a false positive finding (i.e. Type 1 error, for instance, indicating a false positive effect of a treatment due to a significant result by chance). In this study we only performed significance testing when we investigated possible presence of systematic error, with the use of ANOVA, and heteroscedasticity. However, in these cases the finding of a significant difference is a negative result, for the features of TFHPT (i.e. systematic error and/or heteroscedasticity is detected). Hence, in these cases, not adjusting for multiplicity is the conservative way, as the chance of detecting a negative finding increases.[1] Conversely, adjusting for multiplicity would decrease the chance of finding a negative result, i.e. systematic error or heteroscedasticity. This is because, when adjusting for multiplicity you basically lower the standard 5% significance limit.

Results

16. Table 1 footnote: suggest spelling out SD.

Answer: Yes, this is now corrected.

17. Table 1: provide a unit of measurement for the TFHPT. I also suggest replacing ‘min-max’ with ‘range’.

Answer: We have now provided a unit, number of pegs. However, we prefer to use “min-max” because it gives a clearer picture of the low end and high end, which is interesting in relation to floor and ceiling effects.

18. Line 3 (page 12): I think you need to quantify the ‘slight association’ that you mention, or provide further explanation.

Answer: Yes, this is a very good suggestion. We have now quantified this by the use of the Koenker’s studentized test, *page 8, line 176-179, page 12, line 250 and table 2*. No heteroscedasticity was detected. Please, see also our response to comment 11 by reviewer 2.

Discussion

19. Line 30 (page 14): what does the reference at the end of this sentence indicate? It seems you are reporting results of your studies so it is unclear why this reference is needed.

Answer: it was a reference to what the two SRD-measures includes and, hence, to what the difference between them mean. However, it is now removed.

20. Line 20 (page 16): But you removed this 0 result from your analysis.

Answer: Yes, we removed it because it improves the study design and should give a more “true” estimate of the reliability, but in clinical practice 0 pegs is a valid result. We have discussed the rationale for this above, in connection to your comment no 10. This section in the discussion has been moved to the introduction to make the advantages with the TFHPT clearer, *page 5, lines 104-105*.

Limitations:

21. You have reported a new test – but only examined one component of its properties (i.e. reliability). You have not examined other psychometric properties of the test and I would

suggest acknowledging this further in your limitations. Not only have you not examined day-to-day variation, but you have not examined other properties such as validity, sensitivity or specificity.

Answer: Yes, one clear limitation is that we did not investigate sensitivity to change. Neither did we investigate the criterion validity, which would have been good. However, as we have explained in the manuscript under limitations, the criterion validity is most likely at a similar level as for the NHPT, *page 17, lines 356-361*. Regarding sensitivity and specificity, those measures are mostly of interest for test of diagnostic tests,[2] or, if used in tests with a larger number of intervals, there should at least be a clear cut-off value to use for the tests of sensitivity and specificity,

Reviewer 2

Since test-retest reliability is a prerequisite for clinicians and researchers to appropriately interpret change scores in repeated assessments, this study shows good contribution for measurement of manual dexterity. In general, this study shows two advantages: (1) the test-retest reliabilities of a single score and averaged scores for 3 trials were examined simultaneously, which can be helpful to select a more reliable index of manual dexterity; and (2) the SRD is a practical index of random measurement error, which can be useful in determining whether examinees' changes are real or just the consequences of measurement errors. However, this study also has 3 weaknesses: (1) the relatively small sample size (*n* = 31); (2) the short intervals (only a few minutes) with test-retest assessments; and (3) the readability of this manuscript may not be satisfactory. Thus, some efforts would be needed before it can be acceptable and published.

For the whole manuscript.

1. The 'comma' appears to be used inappropriately in the whole manuscript. Specifically, lots of commas were lost, which may have hampered the readability.

Answer: Yes, it seems that commas has been lost. We have been going over them again.

2. Improvement of readability for most sections is needed, especially for the introduction, results, and discussions. The specific suggestions have been provided in the corresponding sections.

Answer: Your suggestions have been very useful to improve the readability.

Introduction

3. In page 4, lines 14 to 16 ("*However, only approximately 25% of studies regarding upper limb interventions include a specific measure of manual dexterity.[1]…*"), the relationships between "low proportion of studies included measures of manual dexterity" and "the TFHPT lacks validations of its test-retest reliability" are unclear. In particular, the lacks of evidence for test-retest reliability appears to be the main rationale of this study.

Answer: there are some links in this chain, the chain between a low proportion of studies including a test of manual dexterity and that the TFHPT is not tested for reliability. We simply start out wide and describe the current situation regarding measurements of manual dexterity in research of stroke rehabilitation, and point out that this is generally a neglected area. Manual dexterity should be tested more often, and the TFHPT is a test of manual dexterity.

4. In page 4, lines 26 to 33 ("*Reliability is a term that describes how the result of a measurement with an instrument is affected by measurement error.[4] The concept of absolute reliability refers to the consistency of measurements within individuals and can be quantified by, for example, the smallest real difference (SRD) and by SRD%.[5]*"), two issues may be concerned.

i. The concept of reliability may be inappropriate to be addressed here. Specifically, it would disrupt the linking between the previous and the next paragraphs, hampering the readability.

Answer: Yes, we agree that the placing of this section hampers the flow. This section is now removed from the introduction, some remnants of it has been moved to the “statistics” section under methods, *page 9, line 192*.

ii. The SRD and SRD% may not be the index of reliability. Since the concepts of SRD and SRD% are more close to the “amount of variations due to measurement error”, the SRD and SRD% may be more appropriate to be indicators of measurement error but not reliability.

Answer: We are not sure whether you only object to the use of SRD as a measure of absolute reliability, or, to all measures related to within-subject variation as measures of absolute reliability. We suppose the latter, as we consider SRD as an extension of the SEM.[1] This is a confusing area as the terminology in reliability research has been used differently between researchers and research areas.[1] However, we decided on this terminology because it is used in several of the central articles that we refer to.[1 3 4] The definition of reliability used by Atkinson: Reliability can be defined as the consistency of measurements, or of an individual’s performance, on a test; or ‘the absence of measurement error‘.[3] Hence, we consider reliability to be the label of the overall concept, which can be subdivided in relative reliability and absolute reliability.[5] The amount of measurement error, quantified as the variability in repeated measurements, is both an explanation for the concept of absolute reliability and the foundation for quantifying it. However, the terms measurement error and absolute reliability are often used interchangeably. We have used the term measurement error when we have discussed the meaning of the results. Here are some examples from our central references that corroborate the terminology of SRD as an index of reliability, however, indirectly through other measures related to within-subject variation: **1)** “The most common methods of analyzing absolute reliability are the SEM and the CV” ([3], page 229). **2)** “Within-subject variation (equal to SEM in Hopkins`s vocabulary) is the most important type of reliability measure for researchers, …..” ([4], page 2). **3)** “…,the SEM provides an absolute index of reliability.” ([1], page 237).

5. In page 4, lines 36 to 52 (“*A weakness with the NHPT is that many persons with stroke cannot reach the lower limit; i.e., a floor effect arises. Furthermore, because there are only nine pegs, measures must be taken to avoid ceiling effects. Therefore, in the original test, the result is expressed as the time to complete the test, including inserting and removing all of the pegs.[2, 3, 26, 7] This, however, aggravates the floor effects because tests that are not completed during the stipulated time are excluded.[2, 3] The maximum time could be prolonged to include the great majority useful to test. However, this would be time consuming and possibly unethical due to the possibility of a non-completed test after a lengthy attempt...*”), two issues may be concerned.

i.The rationales of validating the TFHPT (but not the NHPT) are not clear. If the modified NHPT has successfully overcome the problems of floor and ceiling effects, the remaining problem is lacks of evidence for its reliability. However, the straightforward solution for unknown reliability would be validating it, but not validating another measure (i.e., the TFHPT).

Answer: The result of the modified NHPT can be considered difficult to comprehend, especially for persons suffering from a brain injury. This has now been incorporated in the introduction, *page 5, lines 92-94.*

ii. Readability of this paragraph can be improved. For example, the purposes of using the time of completing the test as an index can be strengthened (e.g., “to minimize ceiling effect, the time of completing the whole test was suggested.”). In addition, the advantages and disadvantages of every version of the modified NHPT can be described in a clearer way.

Answer: the introduction has been substantially rewritten and we hope the readability is improved. However, in this particular sentence we have only used a part of your suggestion, the information that insertion and removal of pegs is included in the test is still there, but within brackets. It is considered important information as it is an important difference compared to both the TFHPT and the modified NHPT*, page 4, lines 81-83*.

Describing the advantages and disadvantages of every version of the modified NHPT would be and overwhelming task, as in clinical practice there is a range of versions. However, the main features, including advantages and disadvantages, of the modified NHPT are outlined in the introduction*, pages 4-5, lines 89-95.* The advantages: only insertion – not removal of pegs. This makes it possible to terminate the test after a stipulated time frame. This makes it possible to measure the performance of the same task across the entire range of hand function. This is enabled since only <u>one task</u> is tested. Conversely, if the task would be to insert as well as remove pegs (as in the original NHPT), a stipulated time frame cause a bias problem. In this case, people at the higher end of hand function would perform both insertion and removal of pegs while people at the lower end of hand function would only perform insertion of pegs – which is more difficult. The bias is the reason why trials not fully completed (both insertion – and removal) have to be excluded in the original NHPT, which elevates the floor effect.

6. In page 4, lines 52 to 56 ("*A modified NHPT is used to mitigate the floor effect while avoiding the ceiling effect, in which the result is expressed as the number of inserted pegs (not removed) per unit of time, i.e., the frequency…*"), why using the "number of inserted pegs" as an index can mitigate the floor effect? Further clarification would be needed to ensure readers can fully understand it.

Answer: it is <u>inserted</u> pegs, as opposed to, <u>inserted and removed</u> pegs, per unit of time that mitigates the floor effects. This is now better explained in the introduction, *page 4-5, lines 89-92.* Please, see also our response to your comment 5ii, it is also relevant here.

7. In page 5, lines 9 to 12 ("*Thus, the TFHPT measures the motor function on a numerical scale, with low floor effects and reasonable ceiling effects*"), three issues may be concerned.

i. The original NHPT, if the number of inserted pegs was adopted, can also provide the numerical score. Thus, this advantage may not be unique for the TFHPT.

Answer: We have now explained the shortcomings with such a result (number of inserted pegs), in a test with only nine pegs, introduction, *page 4, lines 79-81*.

ii. References for the "low floor effect and reasonable ceiling effect" could be added to support that the TFHPT is a promising measure to validate.

Answer: Because no data on the TFHPT has been published before, we cannot provide such references. We can only reason about expected advantages regarding floor and ceiling effects. Please, see also our response to your comment 30.

iii. Please be consistent for the terms used in the whole manuscript. For example, the "upper limb assessment", "fine manual dexterity", and "motor function" stated in the first 3 paragraphs in the Introduction section seem to be interchangeable. However, their scopes and meanings may be different and thus, may result in confusion.

Answer: The intention was not to use these terms interchangeably, however, "motor function" and "motor impairment" was used in that way.

An "upper limb assessment" focuses as much on the arm as on the hand, so, it has a broader meaning than the terms "hand function", "manual dexterity", and "fine manual dexterity", which are focused on the hand. Regarding "hand function" and "manual dexterity" they are broader terms focusing on hand function and they should be possible to use rather interchangeable. Regarding "fine manual dexterity", it is used to describe a finer skill/grasp of the hand (for example tested in the NHPT and the TFHPT), including more finger coordination, as compared to "gross manual dexterity" which is used to describe a rougher grasp of the hand (for example tested in the Box and block test).[6 7]

To be consistent, we have now replaced “manual dexterity” and “motor function” etc with “hand function”. However, we have kept “upper limb function” and “upper limb assessment” to distinguish, arm and hand function/assessment, from, mainly hand function/assessment. We have also kept the term “fine manual dexterity”, specifically, for describing what NHPT and TFHPT measures.

8. In page 5, lines 24 to 26 (“…*However, since the tests have completely different stop criteria – a time limit for the TFHP vs. all pegs inserted for the NHPT – equal reliability cannot be taken for granted.[6]…*”), the “TFHP” appears to be a typo.

Answer: Yes, thanks, this is now corrected.

9. In page 5, lines 29 to 34 (“*Measurements with the TFPHT and quantification on a numerical scale and can be used on a large portion of persons suffering from stroke. Thus, depending on the magnitude of measurement error, this test may be useful, both in clinical practice and in research.*”), three issues may be concerned.

i. Please remove the duplicated information (e.g., the numerical scale) that had mentioned in the other paragraphs.

Answer: Yes, this is now removed.

ii. Please provide references for accessibility of the TFPHT in patients with stroke.

Answer: Please, see our response to your comments 7ii and 30.

iii. The meanings of “*depending on the magnitude of measurement error*” is unclear.

Answer: To make it clearer this sentence has now been rewritten: “Thus, if the size of the measurement error related to the TFHPT is shown to be acceptable, this test may be useful, both in clinical practice and in research.”, *page 5, lines 110-112*.

Methods

10. In page 8, lines 8 to 11 (“*…The exception was the Bland-Altman plots, for which only trials two and three were used…*”), why the Bland-Altman plots were provided only for comparisons between the trial 2 and trial 3?

Answer: You make a very good point. We have now provided the Bland-Altman plots for trial 1 and 2, instead of trial 2 and 3. Because we have now provided measures of heteroscedasticity we believe that these plots are less important. Otherwise, we should have provided plots for the other trial pairs too, *page 8, lines 174-176*.

11. In page 8, lines 16 to 25 (about the Bland-Altman plots and heteroscedasticity), since heteroscedasticity can also be examined by checking whether the Pearson correlation coefficient for the absolute value of differences between two assessments and the average score exceeded 0.3, such method may be adopted to provide objective judgment of the heteroscedasticity.

Answer: This is a very good point. We have now incorporated an assessment of heteroscedasticity by use of the Koenker’s studentized test, page 8, line 176-179, page 12, line 250 and table 2.. No heteroscedasticity was detected. We also performed the analysis you suggest and got similar results, only one of six pairs of trials showed an r >0.3. For the pre-intervention trials, the Pearson correlation coefficient for the pairs, 1-2, 2-3, and 1-3, was, 0.31, -0.06, and 0.23, respectively. For the post-intervention trials, the Pearson correlation coefficient for the pairs, 1-2, 2-3, and 1-3, was, 0.16, 0.27, and 0.22, respectively.

12. In page 8, lines 28 to 44 (*“Measurement error can be either random or systematic. In random error, there is no pattern to the variability, whereas in systematic error the measurement varies in a non-random way, i.e., the mean values between the trials differ.[5] To*

***investigate whether there was a systematic error in test scores, one-way repeated measures ANOVA was used to detect potential between trial effects. Fisher's LSD post hoc tests between trials were performed when the main effect for trials was significant. The assumption of sphericity was met in all trials according to Mauchly's test, whereas the assumption of normal distribution was violated in trial two's pre-intervention according to the Kolmogorov-Smirnov test (p=0.043)"*), the descriptions about systematic error seem to be skipped.**

Answer: The sentence may be somewhat unclear, but the information is there (above in red). To make the sentence clearer we have incorporated at little more information, the rewritten sentence is as follows: In random error, there is no pattern of the variability **between trials**, whereas in systematic error the measurements varies in a non-random way, i.e., the mean values between the trials differ. *Page 8, line 180-182*.

13. In pages 8 to 9, lines 54 to 5 ("*…This makes ICCs sensitive to the degree of between-subject variability, and with all other things being equal, a more heterogeneous sample will produce higher ICC values. [5, 11] In addition, it is difficult to draw statistical inference from one sample to another.[12]…*"), two issues may be concerned.

i. Thought the aforementioned description is correct for ICC, however, since the "within-subject variability" is the main target for examining test-retest reliability, such description appears not straightforward.

Answer: That the "within-subject variability" is the main measure in test-retest reliability may not be evident to all readers. We included this information to explain why there is little focus on the ICCs in the manuscript.

ii. The relationships between the between- and within-subjective variabilities in ICC models are not clear in the article. Thus, the impacts of heterogeneous sample on ICC value are not easy to understand. Please consider to provide some explanation.

Answer: Yes, the reasoning regarding this subject was not clear. We have now incorporated more information in two sentences to make this clearer, *page 9, line 189 and lines 191-192*.

14. In page 9, lines 3 to 5 ("…*In addition, it is difficult to draw statistical inference from one sample to another.[12]*"), the statement may not be appropriate because if the statement is true, ICC seems not a suitable method for test-retest reliability and thus, should not be used in this study.

Answer: Also this information was included to explain why we don´t focus on the ICCs in the manuscript; however, the sentence is now removed.

15. In page 9, lines 8 to 11 (*"Three separate measures of relative reliability including 95% confidence intervals (CIs), namely, ICC2.1, ICC2.3, and ICC3.3, were calculated…"*), two issues may be concerned.

i. This statement may misguide the readers because thought the 95% C.I.s were provided for the 3 ICC values, but these C.I. values were not named ICC(2,1), ICC(2,3), and ICC(3,3).

Answer: Yes, we have rewritten this sentence to make it clearer, but we do not think it is necessary to name the CIs after the different ICCs, the rewritten sentence: Three separate measures of *relative reliability*, i.e., $ICC_{2.1}$, $ICC_{2.3}$, and $ICC_{3.3}$, including 95% confidence intervals (CIs), were calculated. *Page 9, lines 198-199*.

ii. The rationales for using ICC(2,1) remain unclear. If the authors aimed to examine systematic error for the single score, the ICC(3,1) would also be included.

Answer: Because the ANOVA-results indicated that the measurements were affected by systematic error the $ICC_{2.X}$-measures and, above all, the $SRD_{2.X}$-measures were our starting point. By comparing the $SRD_{2.1}$ and $SRD_{2.3}$, the difference between measures representative of a single trial and measures representative of an average of three trials, including systematic error, could be examined. By comparing the $SRD_{2.3}$ and $SRD_{3.3}$, the influence of systematic error could be estimated, in measures representativof an average of three trials.

In less detail, this is explained in the following section, *page 9, line 199-201*: "Three separate measures of *relative reliability*, namely, $ICC_{2.1}$, $ICC_{2.3}$, and $ICC_{3.3}$, including 95% confidence intervals (CIs) were calculated. This panel of measures was used to compare the results representative of single and average measures and to obtain an estimate of the influence of systematic error."

16. In page 9, lines 15 to 29 ("*…The first figure in the ICC designation represents the type of intraclass correlation coefficient (ICC model), while the second figure represents single or average measures, where "1" represents single measures and "2" or higher represents the number of trials from which the average is calculated.[5] ICC 2.1 and ICC2.3 are calculated from a two-way random effect model and incorporate both systematic and random error, whereas ICC3.3 is calculated from a two-way fixed effect model and incorporates only random error.[5, 13]…*"), two issues may be concerned.

i. Readability of this paragraph can be improved. Specifically, introductions of a specific topic can be provided just after the concept was mentioned. For example, since the authors addressed that "the first figure in the ICC designation represent the models of the ICC", the explanations of "the number 2 indicated the two-way random model, which incorporated both systematic and random error; whilst the number 3 represented the two-way fixed effect model, which considered only the random error." can be addressed.

Answer: Yes, this has been addressed and this section has also been moved to accomplish the top-down structure you suggested, *page 9, line 201-211*.

ii. Abbreviations (e.g., ICC) should be used for the whole text since they have been defined for the first time.

Answer: Yes, this is now corrected.

17. In page 9, lines 46 to 48 ("*To estimate the absolute reliability, the standard error of measurement (SEM), SRD and SRD percentage (SRD%) were calculated…*"), concepts of the "relative reliability" and "absolute reliability" can be provided to improve the readability.

Answer: Yes, we have now incorporated information on the concepts of relative and absolute reliability, partly by moving information from the introduction to this section, *page 8, lines 185-186, and, page 9, lines 187 and 192*.

18. In page 9, lines 53 to 60 ("*…SEM is the within-subject standard deviation calculated from repeated tests.[11, 14] The variation of repeated tests can be thought of as the error around a true value; concomitantly, the within-subject standard deviation is used as a measure of measurement error.[14]…*"), these descriptions can be integrated with the introductions of ICC to achieve a top-down structure, which tends to easy to understand for most readers.

Answer: Yes, this is a very good point. We have rewritten parts of the statistics section, and moved the section that you mention, to achieve a top-down structure. One sentence mentioned above has been removed from the manuscript to make the statistics section shorter and more concise, removed: "*The variation of repeated tests can be thought of as the error around a true value; concomitantly,". Page 9, lines 192-197.*

Results
19. In pages 12 to 13, lines 28 to 34 (for both paragraphs about results), two issues may be concerned.

i. Please summarize the main findings of this study because conceptually, descriptions in the results section should not be duplicated to those in the tables.

Answer: Yes, we understand this, we have now removed less important results from the text.

ii. In addition, please remove the redundant explanations (e.g., for the single measures) when reporting the main findings to remain sentences clear and concise.

Answer: We have adhered partly to this suggestion, we have removed information about average and singles measures. However, we have kept information on random and systematic error; because, we believe this information makes it easier to read and understand the results.

Discussions
20. In pages 14 to 17, for all paragraphs in the discussions section, please put the similar findings in the same paragraph and targeting only one issue in a single paragraph. For example, if the reliability of a single score and the average scores was the main issue, then the findings of repeated ANOVA (systematic error) would be moved to other places. By doing so, the readability of the current discussions section can be substantial improved.

Answer: The ANOVA results and the SRD results are not isolated from each other, they describe the same phenomenon from different perspectives. However, you may be right, the mixing of ANOVA-results and SRD-results may be difficult to follow. We have removed such information where it can be considered redundant. In the 4th paragraph, a conclusion drawn from the ANOVA-results has been removed from *page 14, line 283*: “The result of the ANOVA indicated the presence of systematic error”. From the 10th paragraph, a discussion about systematic error based on SRD-results has been removed, *page 16, line 333*: “The decreased systematic error can be observed in the elimination of the difference between the $SRD_{2.3}$ that incorporates systematic error and the $SRD_{3.3}$ that does not in the post-intervention trials”.

Furthermore, the previous 2nd paragraph has been split into paragraphs 3-5, with one clear subject in each paragraph. The current 2nd paragraph is new and incorporated to enhance the readability of the paragraphs 3-5.

21. In page 14, lines 16 to 21 (“*This study indicated that in a selected group of persons suffering from stroke, the use of an average of three trials reduced the measurement error substantially compared to a single trial (SRD2.3 vs SRD2.1).*”), two issues may be concerned.

i. Since the main purposes of this study were not reducing the measurement error, describing the main findings in this way appears to misguide readers.

Answer: This sentence has been removed from this paragraph, we save this topic for later in the discussion.

ii. The authors seem to suggest the users to adopt the average scores rather than a single score. However, such suggestions seem to be stated indirectly, which may not be clear for mostreaders.

Answer: This sentence has now been removed, we save this topic for later in the discussion.

22. In page 14, lines 46 to 51 (“*…This limitation arises because the random error of measurements often increases with the magnitude of the measurements (i.e. heteroscedasticity), [11 12] which also, to some extent, was evident in this study (Figure 2a)….*”), the meanings of this sentence are unclear.

Answer: We are not sure which part of this sentence that is unclear? However, to make the sentence clearer, **and correct**, we have rewritten the first part and removed the last part, *page 15, lines 291-292.*

Regarding the first part: “This limitation” refers to “the possible over- or underestimation, of the number of pegs necessary to demonstrate an improvement”, mentioned in the preceding sentence. However, to make the sentence clearer we have replaced “This limitation arises because” with “The reason is that the”.

Regarding the last part: This part of the sentence has now been removed as it is not correct, the tests performed did not reveal heteroscedasticity, removed: ".which also, to some extent, was evident in this study (Figure 2a)…".

23. In pages 14 to 15, lines 60 to 5 ("*Thus, to capture the systematic error and express the absolute reliability as an absolute number of pegs representing an average of three trials, the most accurate measure investigated in this study for assessing the absolute reliability of the TFHPT is SRD2.3*"), the conclusion cannot be drawn from the previous sentence. Specifically, since the previous sentence only indicated that heteroscedasticity (stability of random error for patients with different levels of manual dexterity) may not be a major concern for the TFHPT, concluding
that the SRD(2,3) was suggested to capture the systematic error are confusing. Accordingly, the flow of this paragraph needs to be reorganized.

Answer: You make a very good point. The conclusion was drawn from the whole paragraph, however, it was not clear. We have now restructured the paragraph, by splitting it in separate paragraphs (now paragraph 2-5), and the conclusion has been drawn step by step, or, paragraph by paragraph, *page 14-15, lines 275-295*.

24. In page 15, lines 36 to 41 ("*…Even though the SRD% measures reported in the studies by Chen et al. and by Ekstrand et al. were calculated in different ways compared to the SRD%2.3 reported in this study, the measures used in these three studies are fairly equivalent.[5]…*"), it would be better to specify the differences between the two methods and the possible impacts on the findings.

Answer: Our point is that the SRD%$_{2.3}$-measure presented in our study is fairly equivalent to the SRD%-measures presented in the studies by Ekstrand et al[8] and by Chen et al[9]. However, at first sight it may appear as if they are not, and to explain this requires a lot of space and hampers the flow of the discussion, therefore, we prefer to explain it below in this answer to you instead.

At first sight it can appear as if the SRD%-measures in the studies by Chen et al and by Ekstrand et al are representative of single measurements because in the study by Chen et al the measure was derived from ICC2.1 and in the study by Ekstrand et al it was directly derived from an ANOVA-table based on single measures[1]. However, in both those studies of the NHPT, the single measurements were actually pre-calculated averages of three trials. Hence, these results are representative of an average of 3 trials (as our SRD%$_{2.3}$). In addition, the SRD%-measure reported by Ekstrand et al was directly derived from a term in the ANOVA-table, and not, as in our study and in the study by Chen et al., from the ICC values. Usually when the SEM (and indirectly the SRD%) is calculated from an ANOVA-table, the mean square error term (residual) is used, which is a pure measure of random error. However, in this case they used "the square root of the total within subject variance" term from the ANOVA-table in the calculation, which renders a measure which incorporates both random and systematic error. As does our SRD%-measure derived from ICC2.3 and Chen's SRD%-measure derived from ICC2.1.

25. In pages 15 to 16, lines 55 to 8 ("*…Second, in this study of the TFHPT, the test and retest trials were performed within minutes compared to within days in the studies of the NHPT, which may have resulted in seemingly worse reliability for the NHPT because of possible random error from day-to-day variation in performance.[11, 16] Third, the 3-5 days between test and retest trials in the study by Chen et al.[2] may also have resulted in seemingly worse reliability in that study because of systematic error…*"), the impact of differences in the findings are unclear. To clarify it, the authors may describe the impacts of the findings directly (e.g., the test-retest reliability in this study may have been overestimated); or comparing the results with the previous studies (e.g., yielding higher test-retest reliability of the TFNPT than the NHPT).

Answer: Yes, we agree. We have tried to write this clearer by taking the perspective of the TFHPT, *page 16, lines 317-325.*

26. In page 16, lines 13 to 22 ("*…One advantage with the TFHPT, compared to the NHPT, is that persons with worse motor function can be tested.[2, 3] In the study by Ekstrand et al.[3], those who did not complete the NHPT in 180 seconds were excluded. This would correspond to inserting and removing a minimum of 2.5 pegs in 50 seconds, whereas 0 pegs inserted in 50 seconds is a valid result with the TFHPT…*"), since the NHPT can have the similar utility to the TFHPT if the index of number of inserted pegs is used, this may not be the unique advantage for all versions of the NHPT. If the authors aimed to strengthen this advantage, the specific versions of the NHPT should be mentioned.

Answer: Yes, you are correct. This is in comparison to the original NHPT which is now made clear. This section has been moved to the introduction to underline the lower floor effect of the TFHPT compared to the original NHPT, *page 5, line 101*.

27. In page 16, lines 46 to 48 ("*…The cause of the decreased random error is less clear, but it could also be attributed to the decreased systematic error.[11]…*"), why the reduction of random error is correlated to that of the systematic error? More information would be needed to address it.

Answer: Most likely, the magnitude of the systematic change differ between individuals, such differences between individuals will increase the SEM (and SRD) and, therefore, it appears as if the random error increases.[4] We have incorporated this information in the discussion, *page 17, line 340-342*.

28. In pages 16 to 17, for the paragraph about limitations, please specify the number of limitations that would be addressed at the beginning. Such information can be helpful for readers to capture the whole picture of limitations.

Answer: Yes, this information is now incorporated in the limitations section, *page 17, line 346*. In addition, to save space, we removed this limitation which we consider less important: The SEM and SRD are not as population-independent as the SRD% but are still considered rather robust.

Conclusion

29. In page 17, lines 27 to 31 ("*…In conclusion, our results suggest that the smallest difference that can be detected using a test procedure with an average of three trials (SRD2.3) conducted by a single tester should be just above 2 pegs with the TFHPT…*"), the actual value of SRD appears to be more important here rather than the code of the ICC model adopted. Thus, why not directly point out the changes of 3 pegs between two assessments may indicate a real change.

Answer: We have now removed the specific SRD-measure. However, we have kept the phrase "just above 2 pegs", not 3 pegs as suggested. Because we have come to the conclusion that an average of measurements is to prefer in front of a single measurement, we can express the result as parts of pegs.

Others

30. Ceiling and floor effects appear to be the main reasons for using the TFNPT. However, they were not examined in this study, remains whether this measure can solve the original problem unknown. The rationales of this study need to be revised.

Answer: Regarding low floor effects; we believe, based on logical reasoning, that the TFHPT has low floor effects compared to the original NHPT. To make this clearer in the introduction, we have moved a section regarding this subject from the discussion to the introduction, page 5, lines 102-106.

Regarding floor effects and ceiling effects: We made a reasonable estimate of a time limit (50 seconds) that would give the largest variability between subjects, i.e. allow as many as possible to insert at least one peg, and, at the same time, prevent as many as possible to reach 25 inserted pegs. For this assessment, we had access to data on the NHPT, from patients performing Constraint induced movement therapy. We had data from more than a few patients, but I don´t remember the exact number. We also tested a subject with an intact nervous system in this process.

Regarding ceiling effects: It was judged that 50 seconds (with 25 available pegs) would result in a low ceiling effect in this population. That is, people suffering from stroke that should benefit from an intensive period of hand/arm rehabilitation and where approximately a quarter of the sample had quite good hand function.[10 11] Indeed, our data did finally support that a time limit of 50 seconds give a very low ceiling effect in this population. Only one of the 60 persons that were screened for inclusion actually did hit the ceiling, *results, page 11, line 241-243, and, discussion, page 16, line 326-331.*

However, there is a potential ceiling effect with the test, e.g. if testing includes subjects with almost normal hand function. To completely rule out the possibility for a ceiling effect (and maintain a low floor effect), the test would have to be modified, e.g. with more available pegs and holes.

References

1. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19(1):231-40.
2. Sim J, Wright C. *Research in health care: concepts, designs and methods*. Cheltenham: Nelson Thornes Ltd, 2002: 184-85.
3. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26(4):217-38
4. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30(1):1-15
5. Carter RE, Lubinsky J, Domholdt E. *Rehabilitation research: principles and applications*. St. Louis, MO: Elsevier-Saunders, 2016: 243-44.
6. Beebe JA, Lang CE. Relationships and responsiveness of six upper extremity function tests during the first six months of recovery after stroke. *J Neurol Phys Ther* 2009;33(2):96-103.
7. Lin KC, Chuang LL, Wu CY, et al. Responsiveness and validity of three dexterous function measures in stroke rehabilitation. *J Rehabil Res Dev* 2010;47(6):563-71.
8. Ekstrand E, Lexell J, Brogardh C. Test-Retest Reliability and Convergent Validity of Three Manual Dexterity Measures in Persons With Chronic Stroke. *PM & R : the journal of injury, function, and rehabilitation* 2016;8(10):935-43.
9. Chen HM, Chen CC, Hsueh IP, et al. Test-retest reproducibility and smallest real difference of 5 hand function tests in patients with stroke. *Neurorehabil Neural Repair* 2009;23(5):435-40.
10. Duncan PW, Goldstein LB, Horner RD, et al. Similar motor recovery of upper and lower extremities after stroke. *Stroke* 1994;25(6):1181-8.
11. Woytowicz EJ, Rietschel JC, Goodman RN, et al. Determining Levels of Upper Extremity Movement Impairment by Applying a Cluster Analysis to the Fugl-Meyer Assessment of the Upper Extremity in Chronic Stroke. *Arch Phys Med Rehabil* 2017;98(3):456-62.

VERSION 2 – REVIEW

REVIEWER	Benjamin Mentiplay La Trobe University, Australia
REVIEW RETURNED	31-Oct-2019

GENERAL COMMENTS	I must thank the authors for their comprehensive revision of their manuscript. I believe the manuscript has been improved since the original submission. Only a few concerns remain from my perspective, which are detailed below. Abstract Thank you for your modifications to include the limitations of the nine-hole peg test, however I suggest shortening the first few

	sentences to simply state that the nine-hole peg test has known limitations such as floor and ceiling effects or potential difficulties in interpretation of results. This would allow more information to provided in the Results and Conclusions sections. Introduction Thank you for including details of the reliability reported in these previous studies. However, you have not yet defined the SRD% abbreviation in the manuscript yet, so perhaps you could either spell this abbreviation out here, or provide an easily interpretable explanation of the differences between studies. Can you provide the correlation value for the Motor Activity Log and the nine-hole peg test? Methods I still think you need to provide more justification in the manuscript as to why 0 and 25 scores were removed. Have you performed your analysis by leaving those scores in? Doing this would give you a good idea of if these scores had any impact on your results. I would imagine clinicians that want to use this test would still count scores of 0 or 25. Can you define what the BL in BL motor assessment is? Page 8, Line 165: This is the first time you have mentioned the Fugl-Meyer test. You should report earlier in the methods about this test if you used it in your results section – perhaps in your first paragraph of the methods to say that you collected participant characteristics such as age, time since stroke, Fugl-Meyer test etc. Discussion The Discussion is quite detailed in its arguments about the reliability and error associated with the TFHPT. However, I feel that the clinical implications from a practitioner's perspective are not very well articulated. How does a clinician/practitioner use the results of your study to inform their clinical practice?

REVIEWER	I-Ping Hsueh School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei.
REVIEW RETURNED	30-Oct-2019

GENERAL COMMENTS	This manuscript has been carefully revised based on the reviewers' suggests. No further common is available. Thank you very much!

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: I-Ping Hsueh
Institution and Country: School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei.
Please state any competing interests or state 'None declared': None declared.

Please leave your comments for the authors below
This manuscript has been carefully revised based on the reviewers' suggests. No further common is available. Thank you very much!

Reviewer: 1
Reviewer Name: Benjamin Mentiplay
Institution and Country: La Trobe University, Australia
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
I must thank the authors for their comprehensive revision of their manuscript. I believe the manuscript has been improved since the original submission. Only a few concerns remain from my perspective, which are detailed below.

Abstract
1. Thank you for your modifications to include the limitations of the nine-hole peg test, however I suggest shortening the first few sentences to simply state that the nine-hole peg test has known limitations such as floor and ceiling effects or potential difficulties in interpretation of results. This would allow more information to provided in the Results and Conclusions sections.

Answer: This is a very good point. This has been changed. Most of the information that was previously removed has been reintroduced and is presented in red text.

Introduction
2. Thank you for including details of the reliability reported in these previous studies. However, you have not yet defined the SRD% abbreviation in the manuscript yet, so perhaps you could either spell this abbreviation out here, or provide an easily interpretable explanation of the differences between studies.

Answer: Thank you for being so observant. The abbreviation has been spelled out on page 4, lines 74-75.

3. Can you provide the correlation value for the Motor Activity Log and the nine-hole peg test?

Answer: This is now included on page 4, line 79.

Methods
4. I still think you need to provide more justification in the manuscript as to why 0 and 25 scores were removed. Have you performed your analysis by leaving those scores in? Doing this would give you a good idea of if these scores had any impact on your results. I would imagine clinicians that want to use this test would still count scores of 0 or 25.

Answer: We have attempted to make this section clearer with a small change in the manuscript on page 6, lines 133-134. For other considerations, see our explanations below.

We did not perform our analyses with those scores included. It was predetermined to exclude persons with measurements at these intervals because this should be the most conservative way to estimate the reliability. Zero and 25 inserted pegs represent wider intervals than the rest of the scale, and a person may be far below the floor or high above the ceiling. By logical reasoning, the results of 0 and 25 pegs are therefore likely to be more stable. Hence, including them in the calculations of the reliability measures is likely to overestimate the reliability for a large part of the measurements with the TFHPT, i.e., measurements at 1-24 inserted pegs.

Our predetermined standpoint, that inclusion of these measurements would have an impact on the reliability, is confirmed by the raw data of this study. Of the 10 persons who were excluded because of

0 inserted pegs, 8 were stable at 0 inserted pegs over 3 trials pre-intervention, and 4 were stable at 0 inserted pegs over 3 trials post-intervention. Four were stable at 0 inserted pegs over 6 trials, both pre-intervention trials and post-intervention trials. These measurements are very stable compared to the measurements of the included participants. Two of the 31 included participants had stable measurements over 3 trials in the pre-intervention trials (at 1 and 19 inserted pegs), and 2 of 31 had stable measurements over 3 trials in the post-intervention trials (at 2 and 16 inserted pegs). These 10 excluded persons would constitute a large part of the study sample if they were included.

Yes, clinicians count scores of 0 inserted pegs, which we support. Our $SRD_{2.3}$ value of 2.3 pegs indicates a real change for participants at this level. However, for participants at this level, a lower SRD value should certainly also indicate a real change, but at how low of a value? A specific assessment of the reliability of this level would be very difficult to perform, and the result would depend heavily on the sample, specifically whether the participants are far below the floor (SRD will decrease towards an infinite low value), just below the floor, or at the floor (SRD will be closer to our reported value). For participants who insert 25 pegs pre-intervention, this is not the correct test; at least, improvements would not be possible to detect.

4. Can you define what the BL in BL motor assessment is?

Answer: BL stands for Birgitta Lindmark. It should not be confused with Britta Lindström who participated in this study. This has been clarified in the manuscript on page 7, line 144.

5. Page 8, Line 165: This is the first time you have mentioned the Fugl-Meyer test. You should report earlier in the methods about this test if you used it in your results section – perhaps in your first paragraph of the methods to say that you collected participant characteristics such as age, time since stroke, Fugl-Meyer test etc.

Answer: We do not believe that this information fits in the first paragraph of the methods section as that paragraph is about the setting and inclusion/exclusion criteria. We have moved it to the second paragraph under Procedure and measurements on page 7, line 143.

Discussion

6. The Discussion is quite detailed in its arguments about the reliability and error associated with the TFHPT. However, I feel that the clinical implications from a practitioner's perspective are not very well articulated. How does a clinician/practitioner use the results of your study to inform their clinical practice?

Answer: This has now been incorporated into the discussion on page 16, line 332-335.

VERSION 3 – REVIEW

REVIEWER	Benjamin Mentiplay La Trobe University, Australia
REVIEW RETURNED	14-Nov-2019

GENERAL COMMENTS	Thank you again for your careful edits to the manuscript. I believe this manuscript is ready for publication. My last point would be to mention in the first sentence of the abstract what these types of tests assess (i.e. hand function).