BMJ Open

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-032560
Article Type:	Research
Date Submitted by the Author:	24-Jun-2019
Complete List of Authors:	Granström, Fredrik; Uppsala University, Centre for Clinical Research Sörmland Hedlund, Mattias; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and rehabilitation, Physiotherapy Lindström, Britta; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and Rehabilitation, Physiotherapy Eriksson, Staffan; Uppsala University, Centre for Clinical Research Sörmland; Uppsala University, Department of Neuroscience, Physiotherapy
Keywords:	Stroke < NEUROLOGY, measurement, Reliability, Clinical assessment, hand function, twenty-five-hole peg test

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Fredrik Granström[1], Mattias Hedlund[2], Britta Lindström[2], Staffan Eriksson[1, 3]

[1]Centre for Clinical Research Sörmland, Uppsala University, Sweden

[2]Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Sweden

[3]Department of Neuroscience, Physiotherapy, Uppsala University, Sweden

CORRESPONDING AUTHOR: Staffan Eriksson, Centre for Clinical Research Sörmland, Uppsala University, Kungsgatan 41, 631 88 Eskilstuna, Sweden. Telephone number: +46 16 105251, +46 73 949 99 33. Email: staffan.eriksson@germed.umu.se

KEY WORDS: clinical assessment, measurement, reproducibility of results, reliability, hand function, stroke, twenty-five-hole peg test

WORD COUNT 3920 words, 4045 words including “strengths and limitations of this study”.

ABSTRACT

Objectives: The twenty-five-hole peg test (TFHPT) is similar to the nine-hole peg test (NHPT), but the larger number of available pegs makes it straightforward to count the number of pegs inserted, during a stipulated time frame (50 seconds), as the result. The objective was to assess the test-retest reliability of the TFHPT when testing persons with stroke, a special focus was placed on the absolute reliability as quantified by the smallest real difference (SRD). Complementary aims were to investigate possible implications for the use of the TFHPT and for how the SRD of the TFHPT performance should be expressed.

Design: This study employed a test-retest design including 3 trials; the pause between trials was approximately 10-120 seconds.

Participants, setting and outcome measure: Thirty-one participants who had suffered a stroke were recruited from a group designated for constraint-induced movement therapy (CIMT) at outpatient clinics. The result of the TFHPT was expressed as the number of pegs inserted.

Methods: Absolute reliability was quantified by the SRD, including random and systematic error for a single trial, $SRD_{2.1}$, and for an average of three trials, $SRD_{2.3}$. For the SRD measures, the corresponding smallest real difference percentage (SRD%) measures were also reported.

Results: The differences in the number of pegs necessary to detect a change in the TFHPT for $SRD_{2.1}$ and $SRD_{2.3}$ were 4.0 and 2.3, respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 36.5% and 21.3%, respectively.

Conclusions: The smallest change that can be detected in the TFHPT should be just above 2 pegs for a test procedure including an average of three trials when systematic error is also

considered ($SRD_{2.3}$). The use of an average of three trials compared to a single trial ($SRD_{2.1}$ vs $SRD_{2.3}$) reduces the measurement error substantially.

Trial registration: ISRCTN registry, reference number ISRCTN24868616.

ARTICLE SUMMARY

Strengths and limitations of this study

- There were some issues in this study regarding the generalizability of the results, as the participants were selected because they should benefit from CIMT, and few scored above 20 pegs during the 50-second trial duration.
- Among other measures of reliability, the SRD percentage was reported, which is a good measure for comparisons between different tests, scales and populations.
- The results were presented with several different reliability measures, which helps to gain some knowledge about the source of the measurement error.
- As the test-retest trials were performed within minutes, the possible day-to-day variation was not captured.
- The intended practice trial was included as one of three trials in the analyses, which appears to have contributed to the present learning effect.

INTRODUCTION

For a comprehensive upper limb assessment among persons with stroke it is important to combine a measure of proximal upper limb function with a measure of manual dexterity.[1] However, only approximately 25% of studies regarding upper limb interventions include a specific measure of manual dexterity.[1] The nine-hole peg test (NHPT) is a common test of fine manual dexterity and the most common in research.[1-3] The reliability of the NHPT has been investigated in two studies of stroke populations, revealing a large discrepancy in the reliability reported.[2, 3]

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]

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. 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;[8] however, this test has not been investigated for reliability.

In Sweden, a similar peg test, a twenty-five-hole peg test (TFHPT), has been used in clinical practice. The larger number of available pegs makes it straightforward to count the number of pegs inserted, during a stipulated time frame of 50 seconds, as the result. Thus, the TFHPT measures the motor function on a numerical scale, with low floor effects and reasonable ceiling effects. Concomitantly, in the two other studies in which the reliability of the NHPT has been investigated, individuals with worse motor impairment, compared to what is possible to test with the TFHPT, were excluded due to floor effects.[2, 3] The TFHPT is not previously described in the literature, and its reliability has not been investigated. Due to the similarity of the NHPT and the TFHPT, the underlying skill assessed with these tests is most likely the same. 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]

Measurements with the TFHPT are quantified 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.

The overall aim of this study was to assess the test-retest reliability of the TFHPT for persons suffering from stroke; a special focus was placed on the absolute reliability, measured as the SRD. Complementary aims were to investigate possible implications for the use of the TFHPT and for how the SRD of the TFHPT performance should be expressed.

METHOD

Participants

The participants in this study were consecutively recruited in the process of screening patients eligible for inclusion in a multicentre randomized controlled trial (RCT), reference number

ISRCTN24868616 at the ISRCTN registry. The patients were considered for inclusion because they were to undergo constraint-induced movement therapy (CIMT) at one of the clinics participating in the RCT. The clinics were outpatient rehabilitation clinics in the public health care system in Sweden. Data were collected at the clinics. The sample in this study consisted of included and excluded participants in the multicentre RCT. The participants were included if they had one stroke or more registered in the medical record and if TFHPT data were available from three trials before and three trials after the CIMT. Moreover, related to the outcome measure, a minimum of one peg and a maximum of 24 pegs inserted was necessary for inclusion. This was to avoid an untrue low measurement error from participants stable at 0 or 25 pegs inserted.

A minimum of 30 participants were included to obtain a sufficient number for a reliability study.[9]

Procedure and measurements

The TFHPT has twenty-five holes and pegs.[6] The test used in this study consisted of a rectangular 21 cm × 45 cm board with a box containing pegs on one side and an elevated 18 cm × 18 cm area with holes on the other side. The holes were 9 mm wide, 18 mm deep and spaced 20 mm apart. The box had a base of 13 cm × 18 cm and was 5 cm deep. The pegs were 40 mm long and 8 mm in diameter.

The TFHPT was administered as the second test in a battery of different tests. The preceding test required approximately 30-60 minutes to administer. The tests were administered in an examination room, in which only the participant and the physiotherapist were present. For the TFHPT, three trials were performed with each hand. The participants started with the less affected hand, followed by the more affected hand, i.e., the hand of investigation in this test-retest study. The pause between trials was approximately 10-120 seconds. The board was

placed at a distance favoured by the participant with the centre row of holes centred towards the navel and the box side oriented towards the tested hand. The starting position was with both hands on the board, and the time keeping was begun upon first hand contact with the pegs. We gave the following instructions to the participants:

1. I want you to pick up one peg at a time and insert them in the holes of the board.
2. Use only the right/left hand; you can only use the other hand to steady the board.
3. You can fill the holes in any order you desire.
4. We start with a practice trial.
5. You have got 50 seconds to insert as many pegs as you can. After 50 seconds, the trial is terminated.
6. Are you ready? Ready, set, go!
7. After the practice trial: This was practice, now come the two actual test trials where the results are noted down. Repeat step 6.

The test-retest reliability of the TFHPT was assessed on two separate occasions, i.e., before and after a two-week training period (the CIMT). The same procedure was used on both of these occasions, and for each participant, all tests were administered by the same physiotherapist. The assessment after the CIMT period was performed as an internal validation.

Two physiotherapists, SE and BL, administered the tests in this study. SE has general experience with persons suffering from stroke and experience administering the original NHPT. BL has extensive experience with persons suffering from stroke, including administering the original NHPT.

Background data were collected by the staff at the clinics, except for data on the dominant hand before the stroke and the Fugl-Meyer test.[10]

Statistics

All three trials were used in the analyses, although the first trial was introduced as a practice trial to the participants. The exception was the Bland-Altman plots, for which only trials two and three were used. Analyses of pre-intervention data and post-intervention data were performed separately.

Bland-Altman plots provided a graphic description of the variability of the data. The mean of trials two and three was plotted against the difference between trials three and two for each subject. 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.

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).

Relative reliability refers to the consistency of the positions of measurements relative to those of others within the tested group and was quantified by using several intra-class correlation coefficients (ICCs).[5 11] Concomitantly, in ICCs, the within-subject variability is compared to the between-subject variability.[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]

Three separate measures of relative reliability including 95% confidence intervasl (CIs), namely, $ICC_{2.1}$, $ICC_{2.3}$, and $ICC_{3.3}$, 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. 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 $ICC_{2.3}$ are calculated from a two-way random effect model and incorporate both systematic and random error, whereas $ICC_{3.3}$ is calculated from a two-way fixed effect model and incorporates only random error.[5, 13] Thus, the less the systematic error contributes to the total error, the closer $ICC_{2.3}$ is to $ICC_{3.3}$.[5] Furthermore, ICCs were calculated for single measures, i.e., 2.1, and average measures, i.e., 2.3 and 3.3. An ICC for single measures represents the reliability for a test procedure in which the subject is tested with a single trial on a test occasion.[5] An ICC for average measures represents the reliability for a test procedure in which the subject is tested with two or more trials on a test occasion and the score is expressed as the average of these trials.

To estimate the absolute reliability, the standard error of measurement (SEM), SRD and SRD percentage (SRD%) were calculated. These three measures of absolute reliability were calculated from each of the three different ICC-measures: $SEM_{2.1}$, $SRD_{2.1,}$ $SRD\%_{2.1}$, $SEM_{2.3}$, $SRD_{2.3}$, $SRD\%_{2.3,}$ $SEM_{3.3}$, $SRD_{3.3,}$ and $SRD\%_{3.3}$. 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] The SEM was calculated according to $SEM = \mathrm{SD}$

$\sqrt{(1-ICC)}$, where SD was calculated from the total sum of squares (SS_{TOTAL}) in the ANOVA table generated in the ICC analyses as $\sqrt{\text{SStotal}/(n-1)}$.[5] SRD can be interpreted as an extension of SEM, in which a 95% CI of the measurement error for both test occasions in a test-retest situation has been incorporated in the measure.[5] The SRD can be seen as the smallest difference that can be detected with 95% certainty in an individual using a test instrument.[5] The SRD was calculated using the formula 1.96 × SEM × $\sqrt{2}$, where 1.96 is related to the 95% CI and $\sqrt{2}$ refers to the error of two measurements.[5] The SRD% was calculated by dividing the SRD value by the grand mean multiplied by 100.[2, 9] This value is independent of measurement units and is indexed to the mean value of the observations from which it was derived and is therefore a good measure for comparisons between different tests, scales and populations.[9, 11, 12] An SRD% of 30% has been suggested as an acceptable level of reliability.[15]

Because estimates of absolute reliability vary with the type of ICC value, some caution is warranted when comparing them with measures from other studies.[5] Therefore, $SEM_{\text{mean square error term (MSE)}} = \sqrt{MSE}$ was also calculated, where MSE (this term is called residual error by Hopkins and mean square residual in the SPSS-output) was taken from the ANOVA table of the ICC calculation.[5, 11] This SEM measure represents the reliability of a test procedure in which the subject is tested with a single trial on a test occasion and is a pure measure of random error.[5] SRD_{MSE} and $SRD\%_{MSE}$ were also derived from SEM_{MSE}.

The analysis of test-retest reliability was pre-planned. SPSS version 21 was used to calculate ICC and ANOVA. The alpha level was set to 0.05.

Patient and public involvement

No patients or public were involved in the development or design of this study.

RESULTS

In this study, participants were recruited between January 2011 and September 2014. Of 60 eligible patients, 29 were excluded for any of the following reasons: not suffering a stroke, missing data, and yielding either below the minimum number of inserted pegs or above the maximum number of inserted pegs (Figure 1). This yielded 31 participants, 21 men and 10 women, for inclusion in the analysis, with a mean ± SD age of 66 ± 9 years (Table 1). The two eligible patients who were excluded because they exceeded the permitted maximum number of pegs inserted completed the 25 pegs in their best trial within 49.3 seconds and 39.4 seconds. Of the eight patients who were excluded because they fell below the minimum number of inserted pegs, five could insert at least one peg in one of the trials. Data were collected from 17 and 14 participants, respectively, by the two physiotherapists (first and last author) at seven clinics.

Table 1. Characteristics of participants at pre-intervention trials

Participants	**N=31**
Age (years), mean ± SD[a]	66 ± 9
Men/women, n[b]	21/10
Time since stroke (months), median (IQR[d]), (min–max)	17 (8–24), (2–70)
Previous dominant hand more affected by stroke, n	19
TFHPT[c], mean of three trials, mean ± SD, (min–max)	10.8 ± 6.8, (1–22.7)
Fugl-Meyer test (score), median (IQR[d]), (min–max)	46 (41–53), (29–62)
More than one stroke, n	3

[a]SD, [b]number of participants, [c]twenty-five-hole peg test, [d]interquartile range

A graphic description of the variability of the data can be seen in the Bland-Altman plots (Figure 2a and b). A slight association between the random error and the magnitude of the measurements can

be observed in the pre-intervention trials (Figure 2a).

For *pre-intervention* trials, the mean values ± SDs for trials 1, 2 and 3 were 10.0 ± 6.5, 11.0 ± 7.1, and 11.5 ± 6.9, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 10.9 and $p < 0.001$. Post hoc tests revealed differences between trials 2 and 1 and between trials 3 and 1, with mean differences (95% CIs) of 1.0 (0.3–1.6) and 1.5 (0.9–2.2), respectively.

For *post-intervention* trials, the mean values ± SD for trials 1, 2 and 3 were 11.8 ± 6.5, 12.4 ± 6.7, and 12.5 ± 6.8, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 4.1 and $p = 0.027$. Post hoc tests revealed a difference between trials 3 and 1, with a mean difference (95% CIs) of 0.6 (0.2–1.1).

For *pre-intervention* trials, the ICCs incorporating random and systematic error, the $ICC_{2.1}$ (95% CI) for single measures and the $ICC_{2.3}$ (95% CI) for average measures, were 0.96 (0.92–0.98) and 0.99 (0.97–0.99), respectively (Table 2). The SRDs incorporating random and systematic error, the $SRD_{2.1}$ for single measures and the $SRD_{2.3}$ for average measures, were 4.0 and 2.3 pegs, respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 36.5% and 21.3%, respectively. The SRD only incorporating random error, the $SRD_{3.3}$ for average measures, was 2.0 pegs.

Table 2. Results of reliability measures for pre-intervention trials

	ICC[a] (95% CI)	**SEM[b], n[c]**	**SRD[d], n[c]**	**SRD%[e]**
$ICC_{2.1}$	0.96 (0.90–0.98)	1.4	4.0	36.5
$ICC_{2.3}$	0.99 (0.97–0.99)	0.8	2.3	21.3
$ICC_{3.3}$	0.99 (0.98–0.99)	0.7	2.0	18.3

Derived from MSE[f]	1.3	3.5	32.1

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

For *post-intervention* trials, the ICCs incorporating random and systematic error, the $ICC_{2.1}$ (95% CI) for single measures and the $ICC_{2.3}$ (95% CI) for average measures, were 0.97 (0.95-–0.98) and 0.99 (0.98–1.0), respectively (Table 3). The SRDs incorporating random and systematic error, the $SRD_{2.1}$ for single measures and the $SRD_{2.3}$ for average measures, were 3.2 and 1.8 pegs, respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 25.9% and 15.0%, respectively. The SRD only incorporating random error, the $SRD_{3.3}$ for average measures, was 1.8 pegs.

Table 3. Results of reliability measures for post-intervention trials

	ICC[a] (95% CI)	**SEM[b], n[c]**	**SRD[d], n[c]**	**SRD%[e]**
$ICC_{2.1}$	0.97 (0.95–0.98)	1.1	3.2	25.9
$ICC_{2.3}$	0.99 (0.98–1.0)	0.7	1.8	15.0
$ICC_{3.3}$	0.99 (0.98-1.0)	0.7	1,8	15.0
Derived from MSE[f]		1.1	3.1	25.5

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

DISCUSSION

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 ($SRD_{2.3}$ vs $SRD_{2.1)}$. Moreover, the absolute test-retest reliability of the TFHPT was at a level that can be considered acceptable for measures representing an average of three trials and incorporating systematic error, i.e., $SRD_{2.3}$ and $SRD\%_{2.3}$.

Comparing $SRD_{2.1}$ to $SRD_{2.3}$, revealed that the use of an average of three trials reduced the measurement error by approximately 1.5 pegs compared to the use of a single trial.[5] The result of the ANOVA indicated the presence of systematic error. Comparing $SRD_{2.3}$ to $SRD_{3.3}$, where $SRD_{3.3}$ incorporates only random error, revealed that the contribution of the systematic error was approximately 0.3 pegs of the total 2.3 pegs when the average of three trials was used.[5] Although the systematic error was small compared to the random error it was not small enough to be overlooked in the assessment of reliability. A measure of absolute reliability, expressed as an absolute number of pegs for the mean of a study population, can over- or underestimate the number of pegs necessary to demonstrate an improvement for an individual.[11 12]. 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). To remedy this, the use of a relative measure of absolute reliability, such as SRD%, has been proposed.[11 12] However, because the heteroscedasticity was modest (Figure 2a), for the TFHP, the plain SRD appears to be a better choice for use in individuals [12]. 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 $SRD_{2.3}$.

The results for $SRD_{2.3}$ and $SRD\%_{2.3}$ were 2.3 pegs and 21.3%, respectively. The value of $SRD\%_{2.3}$ fell within the 30% level that has been suggested as acceptable.[15] The 30% level seems high in this context, with persons affected by stroke in a chronic stage; from a clinical viewpoint, our opinion is that the results of 21.3% and 2.3 pegs in this study indicate a barely acceptable level of absolute reliability. For a favourable level, we believe that a mean number consisting of approximately 1.5 pegs is desired. The relative test-retest reliability, as measured by $ICC_{2.3,}$ was 0.99, which seems excellent. The discrepancy between the level of the relative and the absolute reliability is most likely caused by the heterogeneity in this study population (Figure 2a, Table 1) which inflates the relative reliability.[5]

The level of the relative absolute test-retest reliability (SRD%), the most comparable measure, observed for the TFHPT in this study (21,3%) is better than what Chen et al.[2] reported (54%), and, is at approximately the same level as Ekstrand et al.[3] reported (24%) for the NHPT. 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] Several methodological differences between these 3 studies could have affected the results.[2, 3] *First*, the results of the TFHPT and NHPT were measured using different scales, where the use of time for completion of the test in the NHPT should accommodate more variability compared to the peg count in the TFHPT. However, the SRD% results should still be comparable between the TFHPT and the NHPT because this relative measure of absolute reliability adjusts for different scales and study populations.[11, 12] *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. A systematic error may have originated in possible recovery from stroke because the time since stroke was 3 months or less for a quarter of the study sample [17].

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.

There was a tendency towards improved reliability after the CIMT period which, was due to decreased systematic error and decreased random error. 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 and in the main effects of the trial in the ANOVA results.[5] The decreased systematic error is most likely due to a decreased learning effect, when the participants had previous experience in the test. This is indicated by the increases in the mean values over the trials, especially over trials 1-2, and by the less pronounced increase in the post intervention trials.[11, 12] The lower random error can be observed from the lower $SRD_{3.3}$ results in the post-intervention trials.[5] The cause of the decreased random error is less clear, but it could also be attributed to the decreased systematic error.[11] Furthermore, it is likely that the $SRD_{2.3}$ result of 2.3 pegs for TFHPT could, in reality, be adjusted downwards. A peg test is often used to evaluate a rehabilitation period; because the error is smaller in the post-intervention trials, the "true" SRD may be somewhere between those of the pre- and post-intervention trials (2.3 vs 1.8).

There were some weaknesses in this study, including the relatively low number of participants, few observations above 20 pegs and a study population in which the participants were selected

because they should benefit from CIMT.[9, 11] Both of these latter objections may, to some degree, hinder the generalization of the results to other groups of people suffering from stroke. The SEM and SRD are not as population-independent as the SRD% but are still considered rather robust.[5] In addition, the intended practice trial was included as one of three trials in the analyses which appears to have contributed to systematic error through an increased learning effect, indicated by a large increase in the mean values between trials 1 and 2.[5, 11, 12] Thus, to mitigate the learning effect, a practice trial preceding regular trials is recommended. Moreover, the possible day-to-day variation was not captured in the present study design. The advantage of this approach is that it yields a pure result for measurement error for the instrument in this population; the disadvantage is that the result is less clinically applicable.[11, 16]

In conclusion, our results suggest that the smallest difference that can be detected using a test procedure with an average of three trials ($SRD_{2.3}$) conducted by a single tester should be just above 2 pegs with the TFHPT. Furthermore, to reach an acceptable level of measurement error, the use of the average of multiple trials is crucial. Future research should focus on optimizing the number of trials.

ACKNOWLEDGEMENTS

The authors thank the staff at the clinics for the extra work associated with participating in the RCT. The authors thank Håkan Littbrand for input on the conception of the study and Monika Edström for administrative work with the study.

AUTHOR STATEMENT

Design of the study and data interpretation SE, FG, BL and MH. Acquisition of the data SE, BL and MH. Statistical analysis SE and FG. Drafting and finaliszation of the manuscript SE. Critical revision of the manuscript FG, BL and MH. Final approval of the submitted manuscript FG, BL and MH.

COMPETING INTERESTS

None declared.

FUNDING

The Norrbacka-Eugenia Foundation; The Swedish STROKE-Association; The Stroke Foundation of Northen Sweden; and the Centre for Clinical Research Sörmland, Uppsala University.

DATA SHARING

No additional data are available.

PATIENT CONSENT

Written informed consent was obtained from the participants.

ETHICS APPROVAL

The Regional Ethical Review Board in Umeå, reference 09-104M, with additional approval Dnr 2010/314-32M, Dnr 2011-244-32M, and 2012-235-32M.

REFERENCES

1 Santisteban L, Teremetz M, Bleton JP, et al. Upper limb outcome measures used in stroke rehabilitation studies: a systematic literature review. *PLoS One* 2016;11:e0154792.

2 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;23:435-40.

3 Ekstrand E, Lexell J, Brogardh C. Test-retest teliability and convergent validity of three manual dexterity measures in persons with chronic stroke. *PM R* 2016;8:935-43.

4 Carter RE, Lubinsky J, Domholdt E. Rehabilitation Research: Principles and Applications. St. Louis: Elsevier-Saunders 2011: 237-239.

5 Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231-40.

6 Mathiowetz V, Weber K, Kashman N, et al. Adult norms for the nine hole peg test of finger dexterity. *Occup Ther J Res* 1985;5:25-38.

7 Grice KO, Vogel KA, Le V, et al. Adult norms for a commercially available nine hole peg test for finger dexterity. *Am J Occup Ther* 2003;57:570-73.

8 Heller A, Wade DT, Wood VA, et al. Arm function after stroke: measurement and recovery over the first three months. *J Neurol Neurosurg Psychiatry* 1987;50:714-19.

9 Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil* 2005;84:719-23.

10 Fugl-Meyer AR, Jaasko L, Leyman I, et al. The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance. *Scand J Rehabil Med* 1975;7:13-31.

11 Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30:1-15.

12 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217-38

13 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.

14 Bland JM, Altman DG. Measurement error. *BMJ* 1996;313:744.

15 Huang SL, Hsieh CL, Lin JH, et al. Optimal scoring methods of hand-strength tests in patients with stroke. *Int J Rehabil Res* 2011;34:178-80.

16 Sim J, Wright C. Research in Health Care: Concepts, Designs and Methods. Cheltenham: Nelson Thornes Ltd 2002: 133-134.

17 Kwakkel G, Kollen B, Twisk J. Impact of time on improvement of outcome after stroke. *Stroke* 2006;37:2348-53.

FIGURE LEGENDS

Figure 1. Flowchart of the recruitment process in the study. *Constraint-induced movement therapy.

Figure 2. Bland-Altman plots of numbers of pegs from pre-intervention (a) and post-intervention trials (b). The mean of trials 2 and 3 was plotted against the difference of trials 3 and 2 for each subject. The centre line displays the mean difference for the group between trials 3 and 2. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

Figure 1.

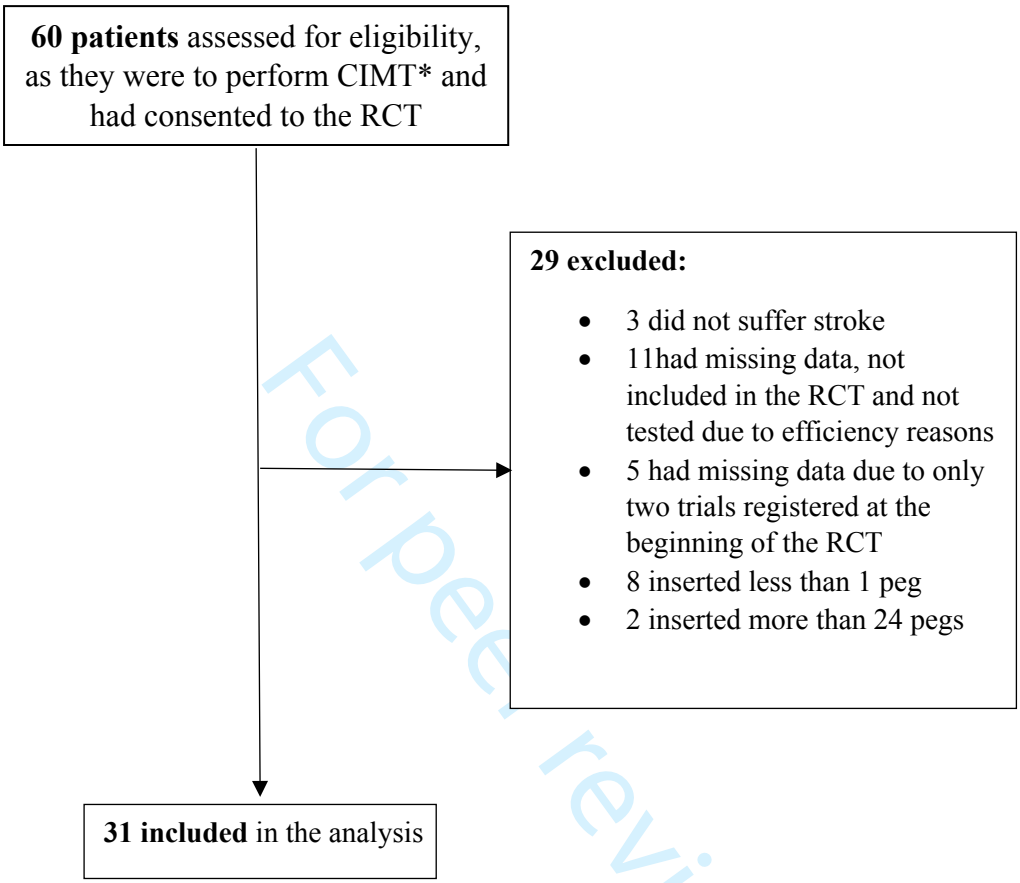


Figure 2a.

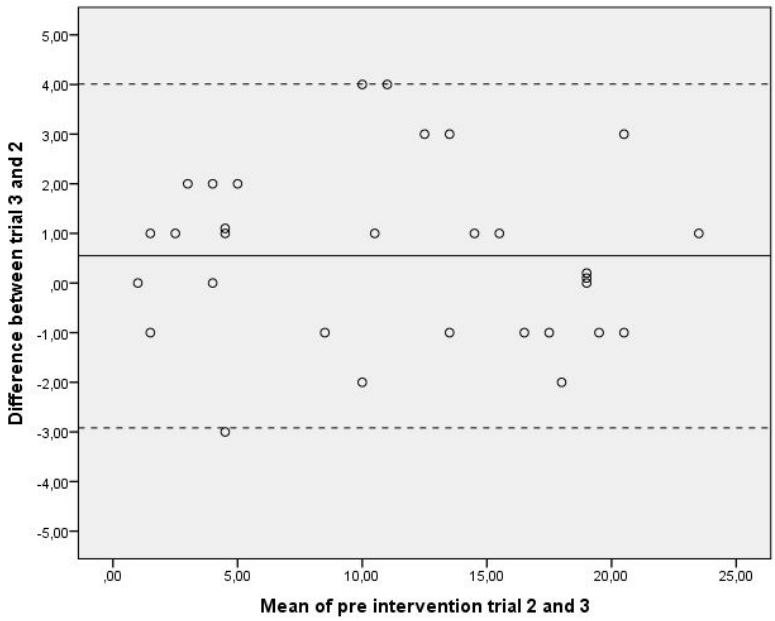
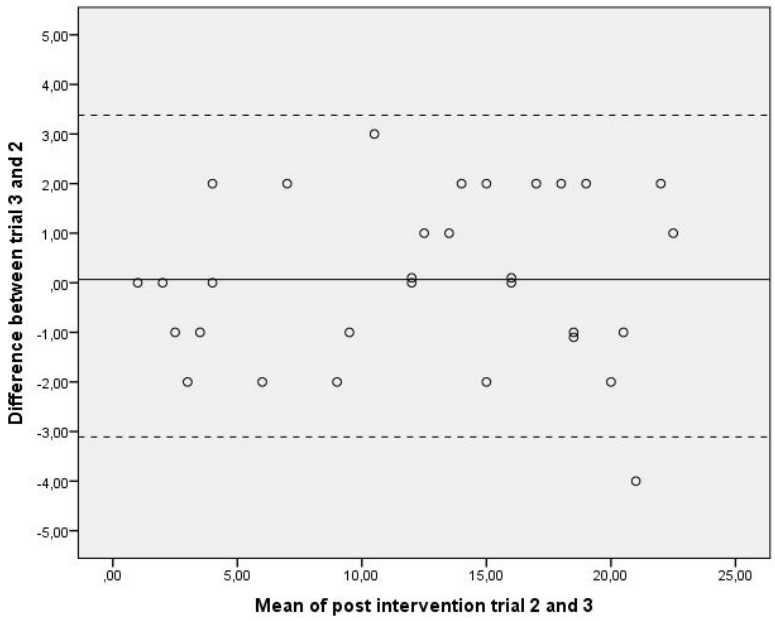


Figure 2b.



EDITORIAL CERTIFICATE

This document certifies that the manuscript listed below was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at American Journal Experts.

Manuscript title:

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Authors:

Staffan Eriksson, Fredrik Granström, Mattias Hedlund, Britta Lindström

Date Issued:

June 13, 2019

Certificate Verification Key:

3D8C-EEE0-E006-60A1-591E

This certificate may be verified at www.aje.com/certificate. This document certifies that the manuscript listed above was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at American Journal Experts. Neither the research content nor the authors' intentions were altered in any way during the editing process. Documents receiving this certification should be English-ready for publication; however, the author has the ability to accept or reject our suggestions and changes. To verify the final AJE edited version, please visit our verification page. If you have any questions or concerns about this edited document, please contact American Journal Experts at support@aje.com.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	(Reliability), 1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4-5
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	10, registered, refnr: ISRCTN24868616
Participants	**6**	Eligibility criteria	5-6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	5-6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	Setting: 6 Dates: 11 Exact locations of the clinics not included.
	9	Whether participants formed a consecutive, random or convenience series	5
Test methods	**10a**	Index test, in sufficient detail to allow replication	Not applicable
	10b	Reference standard, in sufficient detail to allow replication	n.a.
	11	Rationale for choosing the reference standard (if alternatives exist)	n.a.
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	n.a.
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	n.a.
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	n.a.
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	Reliability measures 8-10
	15	How indeterminate index test or reference standard results were handled	n.a.
	16	How missing data on the index test and reference standard were handled	n.a.
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	n.a.
	18	Intended sample size and how it was determined	6
RESULTS			
Participants	**19**	Flow of participants, using a diagram	11
	20	Baseline demographic and clinical characteristics of participants	10
	21a	Distribution of severity of disease in those with the target condition	Table 1, figure 2.
	21b	Distribution of alternative diagnoses in those without the target condition	n.a.
	22	Time interval and any clinical interventions between index test and reference standard	6
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Plots instead, figure 2.
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Page 12, table 2 and 3
	25	Any adverse events from performing the index test or the reference standard	n.a.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	16-17

	27	Implications for practice, including the intended use and clinical role of the index test	16
OTHER INFORMATION			
	28	Registration number and name of registry	ISRCTN registery; referene number: ISRCTN24868616
	29	Where the full study protocol can be accessed	At the registery (above), but not detailed.
	30	Sources of funding and other support; role of funders	18

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-032560.R1
Article Type:	Original research
Date Submitted by the Author:	24-Oct-2019
Complete List of Authors:	Granström, Fredrik; Uppsala University, Centre for Clinical Research Sörmland Hedlund, Mattias; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and rehabilitation, Physiotherapy Lindström, Britta; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and Rehabilitation, Physiotherapy Eriksson, Staffan; Uppsala University, Centre for Clinical Research Sörmland; Uppsala University, Department of Neuroscience, Physiotherapy
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Neurology
Keywords:	Stroke < NEUROLOGY, measurement, Reliability, Clinical assessment, hand function, twenty-five-hole peg test

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Fredrik Granström[1], Mattias Hedlund[2], Britta Lindström[2], Staffan Eriksson[1,3]

[1]Centre for Clinical Research Sörmland, Uppsala University, Sweden

[2]Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Sweden

[3]Department of Neuroscience, Physiotherapy, Uppsala University, Sweden

CORRESPONDING AUTHOR

Staffan Eriksson, Centre for Clinical Research Sörmland, Uppsala University, Kungsgatan 41, 631 88 Eskilstuna, Sweden. Telephone number: +46 16 105251, +46 73 949 99 33. Email: staffan.eriksson@germed.umu.se

KEY WORDS: clinical assessment, measurement, reproducibility of results, reliability, hand function, stroke, twenty-five-hole peg test

WORD COUNT 3991 words, 4105 words including “strengths and limitations of this study”.

ABSTRACT

Objectives: In the nine-hole peg test (NHPT), tests that are not completed within the stipulated time are excluded, resulting in floor effects. A modified NHPT exists in which the result is expressed as the number of inserted pegs per unit of time, which might be difficult to comprehend. In the twenty-five-hole peg test (TFHPT), the larger number of available pegs makes it straightforward to count the number of pegs inserted as the result. It thus provides a comprehendible result and low floor effects, as zero pegs is a valid result. The objective was to assess the test-retest reliability of the TFHPT when testing persons with stroke. A particular focus was placed on the absolute reliability, as quantified by the smallest real difference (SRD). Complementary aims were to investigate possible implications for how the TFHPT should be used and for how the SRD of the TFHPT performance should be expressed.

Design: This study employed a test-retest design including three trials.

Participants, setting and outcome measure: Thirty-one participants who had suffered a stroke were recruited from a group designated for constraint-induced movement therapy at outpatient clinics. The TFHPT result was expressed as the number of pegs inserted.

Methods: Absolute reliability was quantified by the SRD, including random and systematic error for a single trial, $SRD_{2.1}$, and for an average of three trials, $SRD_{2.3}$.

Results: The differences in the number of pegs necessary to detect a change in the TFHPT for $SRD_{2.1}$ and $SRD_{2.3}$ were 4.0 and 2.3, respectively.

Conclusions: The smallest change that can be detected in the TFHPT should be just above two pegs for a test procedure including an average of three trials. The use of an average of three trials compared to a single trial substantially reduces the measurement error.

Trial registration: ISRCTN registry, reference number ISRCTN24868616.

ARTICLE SUMMARY

Strengths and limitations of this study

- The generalizability of the results may be limited: the participants were selected because they should benefit from CIMT, and few scored above 20 pegs during the 50-second trial duration.
- Among other measures of reliability, the SRD percentage was reported; this is a good measure for comparisons between different tests, scales and populations.
- The results were presented with several different reliability measures to offer knowledge about the source of the measurement error.
- As the test-retest trials were performed within minutes, possible day-to-day variation was not captured.
- The intended practice trial was included as one of three trials in the analyses, which appears to have contributed to the learning effect.

INTRODUCTION

For a comprehensive upper limb assessment among persons with stroke, it is important to combine a measure of proximal upper limb function with a measure of hand function.[1] However, only approximately 25% of studies regarding upper limb interventions include a specific measure of hand function.[1] The nine-hole peg test (NHPT) is a common test of hand function focusing on fine manual dexterity, and it is the most common such test used in research.[1-3] Two studies of stroke populations investigated the reliability of the NHPT, and there was a large discrepancy in the reliability reported: SRD% 24% vs. 52%.[2, 3] The NHPT has mostly shown moderate to excellent correlations (0.55-0.97) with other tests and self-reports focusing on hand function, including the Action Research Arm Test, the Jebsen-Taylor Hand Function Test, and the Stroke Impact Scale (hand function domain).[4, 5] The exception is the Motor Activity Log, for which a low correlation has been reported.[5]

A weakness of the NHPT is that many persons with stroke cannot reach the lower limit; i.e., a floor effect arises. Furthermore, if the number of completed pegs is used as an outcome measure, a test with only nine pegs can measure only a narrow range of hand function, resulting in profound ceiling effects.[6] Therefore, to widen the scale and avoid ceiling effects, the original NHPT expresses the result as the time needed to complete the test (including inserting and removing all the pegs).[2, 3, 7, 8] However, this approach aggravates the floor effects because tests that are not completed during the stipulated time (limits of 60 and 180 seconds has been used) are excluded.[2, 3] The maximum time could be prolonged; however, this would be time consuming, mentally strenuous and therefore possibly unethical due to the possibility of a non-completed test after a lengthy attempt. A modified NHPT is used to mitigate the floor effect while avoiding the ceiling effect; in this modified version, the result is expressed as the number of inserted pegs per unit of time, i.e., the frequency.[9] This modified test includes only peg insertion and not peg removal. It is thus possible also to

include tests that were not completed within the stipulated time limit and still measure performance on the same task across the entire range of hand function. However, it may be difficult both to interpret the frequency and to communicate it to other staff members and patients, especially to those suffering from a brain injury. The reliability of this modified test has not been investigated.

In Sweden, a similar peg test, a twenty-five-hole peg test (TFHPT), has been used in clinical practice. The larger number of available pegs makes it straightforward to count the number of pegs inserted during a stipulated time frame of 50 seconds, as the test result. Thus, the TFHPT measures fine manual dexterity on a numerical scale that is easy to comprehend, with low floor effects and presumably reasonable ceiling effects (based on pre-study data). Moreover, compared to the individuals whom the original NHPT can test, individuals with worse hand function can be tested with the TFHPT.[2, 3] Of the two studies investigating the reliability of the NHPT, the one with the most generous time limit excluded all tests that were not completed in 180 seconds.[3] This limit corresponds 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. The TFHPT has not been previously described in the literature, and its reliability has not been investigated. Due to the similarity of the NHPT and the TFHPT, the underlying skill assessed with these tests is most likely the same. However, since the tests have completely different stop criteria – a time limit for the TFHPT vs. the insertion of all pegs for the NHPT – equal reliability cannot be taken for granted.[7] Thus, if the size of the measurement error related to the TFHPT is shown to be acceptable, this test may be useful in both clinical practice and research.

The overall aim of this study was to assess the test-retest reliability of the TFHPT for persons suffering from stroke. A particular focus was placed on the absolute reliability, as quantified by the smallest real difference (SRD). Complementary aims were to investigate possible

implications for how the TFHPT should be used and for how the SRD of the TFHPT performance should be expressed.

METHOD

Participants

The participants in this study were consecutively recruited in the process of screening patients eligible for inclusion in a multicentre randomized controlled trial (RCT), reference number ISRCTN24868616 at the ISRCTN registry. The patients were considered for inclusion because they were to undergo constraint-induced movement therapy (CIMT) at one of the clinics participating in the RCT. The clinics were outpatient rehabilitation clinics in the public health care system in Sweden. Data were collected at the clinics. The sample in this study consisted of included and excluded participants in the multicentre RCT. The participants were included if they had one stroke or more registered in the medical record and if TFHPT data were available from three trials before and three trials after the CIMT. Moreover, with regard to the outcome measure, a minimum of one peg and a maximum of 24 pegs inserted was necessary for inclusion. This was to avoid an untrue low measurement error from participants stable at 0 or 25 pegs inserted. Because these two intervals are wider, measurements at these intervals should be more stable.

A minimum of 30 participants were included to obtain a sufficient number for a reliability study.[10]

Procedure and measurements

The TFHPT has twenty-five holes and pegs (Figure 1). The test used in this study consisted of a rectangular 21 cm × 45 cm board with a box containing pegs on one side and an elevated 18 cm × 18 cm area with holes on the other side. The holes were 9 mm wide and18 mm deep,

and they were spaced 20 mm apart. The box had a base of 13 cm × 18 cm and was 5 cm deep. The pegs were 40 mm long and 8 mm in diameter.

The TFHPT was administered as the second test in a battery of different tests. The preceding test, BL motor assessment, required approximately 30-60 minutes to administer.[11, 12] The tests were administered in an examination room in which only the participant and the physiotherapist were present. For the TFHPT, three trials were performed with each hand. The participants started with the less affected hand, followed by the more affected hand, i.e., the hand of investigation in this test-retest study. The pause between trials was approximately 10-120 seconds. The board was placed at a distance favoured by the participant with the centre row of holes centred towards the navel and the box side oriented towards the tested hand. The starting position was with both hands on the board, and time keeping began upon first hand contact with the pegs. We gave participants the following instructions:

1. I want you to pick up one peg at a time and insert them in the holes of the board.
2. Use only the right/left hand; you can only use the other hand to steady the board.
3. You can fill the holes in any order you desire.
4. We start with a practice trial.
5. You have 50 seconds to insert as many pegs as you can. After 50 seconds, the trial is terminated.
6. Are you ready? Ready, set, go!
7. After the practice trial: This was practice; now come the two actual test trials, where the results are recorded. Repeat step 6.

The test-retest reliability of the TFHPT was assessed on two separate occasions, i.e., before and after a two-week training period (the CIMT). The same procedure was used on both of these occasions, and for each participant, all tests were administered by the same

physiotherapist. The assessment after the CIMT period was performed as an internal validation.

Two physiotherapists, SE and BL, administered the tests in this study, including the Fugl-Meyer test.[13] SE has general experience with persons suffering from stroke and experience administering the original NHPT. BL has extensive experience with persons suffering from stroke, including administering the original NHPT. Background data from medical records were collected by staff at the clinics.

Statistics

All three trials were used in the analyses, although the first trial was introduced to the participants as a practice trial. Analyses of pre-intervention data and post-intervention data were performed separately.

Bland-Altman plots of trials one and two provided a graphic description of the data variability. The mean of trials one and two was plotted against the difference between trials two and one for each subject. Heteroscedasticity – i.e., an association between the random error and the magnitude of measurements [14] – was investigated with pairwise comparisons of trials using Koenker's [15] studentized test, which is useful for small samples and skewed data. Heteroscedasticity is indicated by a significant result.

Measurement error can be either random or systematic. In random error, there is no pattern of variability between trials, whereas in systematic error, the measurements varies in a non-random way; i.e., the mean values between the trials differ.[16] To investigate whether there was a systematic error in test scores, one-way repeated measures ANOVA was used to detect potential between trial effects.

Reliability is a term that describes how the measurement result of an instrument is affected by measurement error.[6, 14] Reliability can be quantified as either relative or absolute.[6]

Relative reliability refers to the consistency of the positions of measurements relative to those of others within the tested group, and it is quantified using several intra-class correlation coefficients (ICCs).[16, 17] In ICCs, between-subject variability is related to the within-subject variability by a ratio.[16] Thus, ICCs are sensitive to the degree of between-subject variability, and with all other things being equal, a more heterogeneous sample (i.e., a larger between-subject variability) produces higher ICC values.[16, 17] The concept of *absolute reliability* refers to the consistency of measurements within individuals.[6, 16] Measurement error, quantified as within-subject standard deviations in repeated tests, is a common measure of absolute reliability [6, 14, 16-18] and is called the standard error of measurement (SEM). SRD is an extension of the SEM, and it can be seen as the smallest detectable difference, with 95% certainty, using a test instrument on an individual.[16]

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. 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. *The first* figure in the ICC designation represents the type of ICC model.[16] $ICC_{2.1}$ and $ICC_{2.3}$ are calculated from a two-way random effect model and incorporate both systematic and random error, whereas $ICC_{3.3}$ is calculated from a two-way fixed effect model and incorporates only random error.[16, 19] Thus, the less systematic error contributes to the total error, the closer $ICC_{2.3}$ is to $ICC_{3.3}$.[16] *The second* figure in the ICC designation 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.[16] $ICC_{2.1}$ represents the reliability of a test procedure in which the subject is tested with a single trial on a test occasion.[16] $ICC_{2.3}$ and $ICC_{3.3}$ represents the reliability of a test procedure in which the subject is tested with three trials on a test occasion and the score is expressed as the average of these trials.

To estimate *absolute reliability*, the SEM, SRD and SRD percentage (SRD%) were calculated for each of the three different ICC measures ($ICC_{2.1}$, $ICC_{2.3}$, and $ICC_{3.3}$), resulting in the corresponding properties $SEM_{2.1}$, $SRD_{2.1,}$ $SRD\%_{2.1}$, $SEM_{2.3}$, $SRD_{2.3}$, $SRD\%_{2.3,}$ $SEM_{3.3}$, $SRD_{3.3,}$ and $SRD\%_{3.3}$. The SEM was calculated according to $SEM = SD\sqrt{(1 - ICC)}$, where SD was calculated from the total sum of squares (SS_{TOTAL}) in the ANOVA table generated in the ICC analyses as $\sqrt{SStotal/(n-1)}$.[16] The SRD was calculated using the formula $1.96 \times SEM \times \sqrt{2}$, where 1.96 is related to the 95% CI and $\sqrt{2}$ refers to the error of two measurements.[16] The SRD% was calculated by dividing the SRD value by the grand mean multiplied by 100.[2, 10] This value is independent of measurement units and is indexed to the mean value of the observations from which it was derived. It is therefore a good measure for comparisons between different tests, scales and populations.[10, 14, 17] An SRD% of 30% has been suggested as an acceptable level of reliability.[20]

Because estimates of absolute reliability vary with the type of ICC value, some caution is warranted when comparing them with measures from other studies.[16] Therefore, $SEM_{mean\ square\ error\ term\ (MSE)} = \sqrt{MSE}$ was also calculated, where MSE (this term is called residual error by Hopkins and the mean square residual in the SPSS output) was taken from the ANOVA table of the ICC calculation.[16, 17] This SEM measure represents the reliability of a test procedure in which the subject is tested with a single trial on a test occasion, and it is a pure measure of random error.[16] SRD_{MSE} and $SRD\%_{MSE}$ were also derived from SEM_{MSE}.

The analysis of test-retest reliability was pre-planned. SPSS version 21 was used to calculate ICC and ANOVA. The alpha level was set to 0.05.

Patient and public involvement

No patients or members of the public were involved in the development or design of this study.

RESULTS

In this study, participants were recruited between January 2011 and September 2014. Of 60 eligible patients, 29 were excluded for any of the following reasons: not suffering a stroke, missing data, and yielding either below the minimum or above the maximum number of inserted pegs (Figure 2). This yielded 31 participants (21 men and 10 women) for inclusion in the analysis, with a mean ± SD age of 66 ± 9 years (Table 1). The two eligible patients who were excluded because they exceeded the permitted maximum number of pegs inserted completed the 25 pegs in their best trial within 49.3 seconds and 39.4 seconds. Of the ten patients who were excluded because they fell below the minimum number of inserted pegs, five inserted at least one peg in one of the trials. Data were collected from 17 and 14 participants by the two physiotherapists (third and last author, respectively) at seven clinics.

Table 1. Characteristics of participants at pre-intervention trials

Participants	**N=31**
Age (years), mean ± SD[a]	66 ± 9
Men/women, n[b]	21/10
Time since stroke (months), median (IQR[d]), (min-max)	17 (8–24), (2–70)
Previous dominant hand more affected by stroke, n	19
TFHPT[c], mean of three trials (number of pegs), mean ± SD, (min-max)	10.8 ± 6.8, (1–22.7)
Fugl-Meyer test (score), median (IQR[d]), (min-max)	46 (41–53), (29–62)
More than one stroke, n	3

[a]Standard deviation, [b]number of participants, [c]twenty-five-hole

peg test, [d]interquartile range

A graphic description of the data variability can be seen in the Bland-Altman plots (Figures 3 and 4). According to Koenker's studentized test, the measurement error was not affected by heteroscedasticity (Table 2).

Table 2. Results of the Koenker's studentized test, n=31

Pairwise test of trials	**Pre-intervention trials**		**Post-intervention trials**	
	Chi-square	P-value	Chi-square	P-value
1-2	1.33	0.25	0.41	0.52
2-3	0.05	0.83	1.38	0.24
1-3	0.28	0.60	0.20	0.66

For *pre-intervention* trials, the mean values ± SDs for trials 1, 2 and 3 were 10.0 ± 6.5, 11.0 ± 7.1, and 11.5 ± 6.9, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 10.9 and $p < 0.001$. Post hoc tests revealed differences between trials 2 and 1 and between trials 3 and 1, with mean differences (95% CIs) of 1.0 (0.3–1.6) and 1.5 (0.9–2.2), respectively.

For *post-intervention* trials, the mean values ± SD for trials 1, 2 and 3 were 11.8 ± 6.5, 12.4 ± 6.7, and 12.5 ± 6.8, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 4.1 and $p = 0.027$. Post hoc tests revealed a difference between trials 3 and 1, with a mean difference (95% CIs) of 0.6 (0.2–1.1).

For *pre-intervention* trials, $ICC_{2.3}$ (95% CI) was 0.99 (0.97–0.99) (Table 3). The SRDs incorporating random and systematic error, $SRD_{2.1}$ and $SRD_{2.3}$, were 4.0 and 2.3 pegs,

respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 36.5% and 21.3%, respectively. The SRD incorporating only random error, $SRD_{3.3}$, was 2.0 pegs.

For *post-intervention* trials, $SRD_{2.1}$ and $SRD_{2.3}$ were 3.2 and 1.8 pegs, respectively (Table 4). $SRD_{3.3}$, was 1.8 pegs.

Table 3. Results of reliability measures for pre-intervention trials

	ICC[a] (95% CI)	SEM[b], n[c]	SRD[d], n[c]	SRD%[e]
$ICC_{2.1}$	0.96 (0.90–0.98)	1.4	4.0	36.5
$ICC_{2.3}$	0.99 (0.97–0.99)	0.8	2.3	21.3
$ICC_{3.3}$	0.99 (0.98–0.99)	0.7	2.0	18.3
Derived from MSE[f]		1.3	3.5	32.1

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

Table 4. Results of reliability measures for post-intervention trials

	ICC[a] (95% CI)	SEM[b], n[c]	SRD[d], n[c]	SRD%[e]
$ICC_{2.1}$	0.97 (0.95–0.98)	1.1	3.2	25.9
$ICC_{2.3}$	0.99 (0.98–1.0)	0.7	1.8	15.0
$ICC_{3.3}$	0.99 (0.98-1.0)	0.7	1,8	15.0
Derived from MSE[f]		1.1	3.1	25.5

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

DISCUSSION

This study indicated that in a selected group of persons suffering from stroke, the absolute test-retest reliability of the TFHPT was at a level that can be considered acceptable for measures representing an average of three trials and incorporating systematic error.

To assess implications for the use of the TFHPT and to determine which SRD measure best captures the absolute reliability, three issues were considered: 1) whether to use single or average measures, 2) whether to include systematic error in the assessments, and 3) whether to take heteroscedasticity into account.

Comparing $SRD_{2.1}$ to $SRD_{2.3}$ revealed that the use of an average of three trials reduced the measurement error by approximately 1.5 pegs compared to the use of a single trial. This finding suggests that the reliability of the TFHPT is substantially improved when an average of three trials is used.

Comparing $SRD_{2.3}$ to $SRD_{3.3}$, where $SRD_{3.3}$ incorporates only random error, revealed that the contribution of the systematic error was approximately 0.3 pegs of the total 2.3 pegs when the average of three trials was used.[16] Although the systematic error was small compared to the random error it was not small enough to be overlooked in the assessment of reliability. Therefore, $SRD_{2.3}$ is preferable to $SRD_{3.3}$ for measuring the reliability of the TFHPT.

The choice of SRD% instead of SRD is dependent on whether the measurement error is affected by heteroscedasticity. A measure of absolute reliability, expressed as an absolute number of pegs, can over- or underestimate the number of pegs necessary to demonstrate an improvement for an individual.[14, 17] The reason is that the random error of measurements often increases with the magnitude of the measurements (i.e., heteroscedasticity).[14, 17] As a remedy, the use of a relative measure of absolute reliability, such as SRD%, has been proposed.[14, 17] However, the lack of heteroscedasticity detection suggests that both $SRD_{2.3}$ and $SRD\%_{2.3}$ are appropriate measures of reliability for the TFHPT.[14]

The results for $SRD_{2.3}$ and $SRD\%_{2.3}$ were 2.3 pegs and 21.3%, respectively. The value of $SRD\%_{2.3}$ fell within the 30% level, which has been suggested as acceptable.[20] The 30% level seems high in this context, with persons affected by stroke in a chronic stage; from a clinical viewpoint, our opinion is that the results of 21.3% and 2.3 pegs in this study indicate a barely acceptable level of absolute reliability. For a favourable level, we believe that a mean number consisting of approximately 1.5 pegs is desirable.

The relative test-retest reliability, as measured by $ICC_{2.3,}$ was 0.99, which seems excellent. The discrepancy between the level of the relative and the absolute reliability is most likely caused by the heterogeneity in this study population (Figure 2a, Table 1) which inflates the relative reliability.[16]

The level of the relative absolute test-retest reliability (SRD%), the most comparable measure, observed for the TFHPT in this study (21.3%) is better than what Chen et al.[2] reported (54%), and, is at approximately the same level as Ekstrand et al.[3] reported (24%) for the NHPT. Although the SRD% measures reported in the studies by Chen et al. and by Ekstrand et al. were calculated in different ways than the $SRD\%_{2.3}$ reported in this study, the measures used in these three studies are fairly equivalent.[16] Several methodological differences between these three studies could have affected the results.[2, 3] *First*, the results of the TFHPT and NHPT were

measured using different scales, where the use of time for completion of the test in the NHPT should accommodate more variability than the peg count used in the TFHPT. However, the SRD% results should still be comparable between the TFHPT and the NHPT because this relative measure of absolute reliability adjusts for different scales and study populations.[14, 17] *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. Thus, the TFHPT may seem more reliable because of possible random error from day-to-day variation in performance which was not captured in this study.[17, 21] *Third*, the longer time since stroke in this study of the TFHPT compared to the study of NHPT by Chen et al.[2] may have resulted in seemingly better reliability for the TFHPT because of a more stable level of hand function. In the study by Chen et al., a systematic error may have originated in recovery from stroke in the 3-5 days between the test and retest trials because the time since stroke was 3 months or less for a quarter of the study sample.[22]

It seems that the ceiling effects in the TFHPT can be considered acceptable. Only two of the persons assessed for eligibility inserted 25 pegs, and only one of them actually did hit the ceiling because according to this individual's best times for completion of the 25 pegs, he/she would have been able to insert more pegs if available. This occured in a sample where approximately a quarter of the included participants suffered from a mild impairment of arm and hand function, as judged by the Fugl-Meyer test.[23, 24]

There was a tendency towards improved reliability after the CIMT period, which was due to decreased systematic error and decreased random error. The decreased systematic error can be observed in the main effects of trial in the ANOVA results.[16] The decreased systematic error is most likely due to a decreased learning effect when the participants had previous experience in the test. The learning effect is indicated by the increases in the mean values over the trials, especially over trials 1-2, and the decreased learning effect is indicated by the less pronounced

increase in the post-intervention trials.[14, 17] The lower random error can be observed from the lower $SRD_{3.3}$ results in the post-intervention trials.[16] The cause of the decreased random error is less clear, but it could also be attributed to the decreased systematic error.[17] This is because, the magnitude of the learning effect probably differs between individuals, which will show as random error. Furthermore, it is likely that the $SRD_{2.3}$ result of 2.3 pegs for TFHPT could, in reality, be adjusted downwards. A peg test is often used to evaluate a rehabilitation period; because the error is smaller in the post-intervention trials, the "true" SRD may be somewhere between the SRDs of the pre- and post-intervention trials (2.3 vs 1.8).

Four weaknesses of this study should be considered. The sample included a relatively low number of participants with few observations above 20 pegs, and participants that were selected because they should benefit from CIMT.[10, 17] These sample qualities may thus, to some degree, hinder the generalization of the results to other groups of people suffering from stroke. In addition, the intended practice trial was included as one of three trials in the analyses which appears to have contributed to systematic error through an increased learning effect, indicated by a large increase in the mean values between trials 1 and 2.[14, 16, 17] Thus, to mitigate the learning effect, a practice trial preceding regular trials is recommended. Moreover, the possible day-to-day variation was not captured in the present study design. The advantage of this approach is that it yields a pure result for measurement error for the instrument in this population; the disadvantage is that the result is less clinically applicable.[17, 21] Finally, in this study, sensitivity to change and validity were not examined. However, the criterion validity for NHPT has mostly shown a moderate to excellent level [4, 5] and the underlying skill assessed with the TFHPT is most likely the same. A high reliability level is a prerequisite for a high validity, and because the reliability of the TFHPT was at the same level as that of the NHPT, the criterion validity should also be similar.[21]

In conclusion, our results suggest that the smallest detectable difference between two assessments using a test procedure with an average of three trials conducted by a single tester should be just above two pegs with the TFHPT. Furthermore, to reach an acceptable level of measurement error, the use of the average of multiple trials is crucial. Future research should focus on optimizing the number of trials.

ACKNOWLEDGEMENTS

The authors thank the staff at the clinics for the extra work associated with participating in the RCT. The authors thank Håkan Littbrand for input on the conception of the study and Monika Edström for administrative work with the study.

AUTHOR CONTRIBUTIONS

Study design and data interpretation: SE, FG, BL and MH. Data acquisition: SE, BL and MH. Statistical analysis: SE and FG. Drafting and finaliszation of the manuscript: SE. Critical revision of the manuscript: FG, BL and MH. Final approval of the submitted manuscript: FG, BL and MH.

COMPETING INTERESTS

None declared.

FUNDING

The Norrbacka-Eugenia Foundation; The Swedish STROKE-Association; The Stroke Foundation of Northen Sweden; and the Centre for Clinical Research Sörmland, Uppsala University.

DATA SHARING

No additional data are available.

PATIENT CONSENT

Written informed consent was obtained from the participants.

ETHICS APPROVAL

The Regional Ethical Review Board in Umeå, reference 09-104M, with additional approval Dnr 2010/314-32M, Dnr 2011-244-32M, and 2012-235-32M.

REFERENCES

1. Santisteban L, Teremetz M, Bleton JP, et al. Upper limb outcome measures used in stroke rehabilitation studies: a systematic literature review. *PLoS One* 2016;11:e0154792.
2. 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;23:435–40.
3. Ekstrand E, Lexell J, Brogardh C. Test-retest reliability and convergent validity of three manual dexterity measures in persons with chronic stroke. *PM R* 2016;8:935–43.
4. 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;33:96–103.
5. Lin KC, Chuang LL, Wu CY, et al. Responsiveness and validity of three dexterous function measures in stroke rehabilitation. *J Rehabil Res Dev* 2010;47:563–71.
6. Carter RE, Lubinsky J, Domholdt E. *Rehabilitation Research: Principles and Applications*. St. Louis, MO: Elsevier-Saunders, 2016:239–244.
7. Mathiowetz V, Weber K, Kashman N, et al. Adult norms for the Nine-Hole Peg Test of Finger Dexterity. *Occup Ther J Res* 1985;5:25–38.
8. Grice KO, Vogel KA, Le V, et al. Adult norms for a commercially available Nine Hole Peg Test for finger dexterity. *Am J Occup Ther* 2003;57:570–3.
9. Heller A, Wade DT, Wood VA, et al. Arm function after stroke: measurement and recovery over the first three months. *J Neurol Neurosurg Psychiatry* 1987;50:714–9.
10. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil* 2005;84:719–23.

11. Lindmark B, Hamrin E. Evaluation of functional capacity after stroke as a basis for active intervention. Validation of a modified chart for motor capacity assessment. *Scand J Rehabil Med* 1988;20:111–5.

12. Lindmark B, Hamrin E. Evaluation of functional capacity after stroke as a basis for active intervention. Presentation of a modified chart for motor capacity assessment and its reliability. *Scand J Rehabil Med* 1988;20:103–9.

13. Fugl-Meyer AR, Jaasko L, Leyman I, et al. The post-stroke hemiplegic patient. 1. A method for evaluating of physical performance. *Scand J Rehabil Med* 1975;7:13-31.

14. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217–38.

15. Koenker R. A note on studentizing a test for heteroscedasticity. *J Econom* 1981;17:107–12.

16. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231–40.

17. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30:1–15.

18. Bland JM, Altman DG. Measurement error. *BMJ* 1996;313:744.

19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.

20. Huang SL, Hsieh CL, Lin JH, et al. Optimal scoring methods of hand-strength tests in patients with stroke. *Int J Rehabil Res* 2011;34:178–80.

21. Sim J, Wright C. *Research in Health Care: Concepts, Designs and Methods*. Cheltenham, England: Nelson Thornes Ltd 2002: 123–33.

22. Kwakkel G, Kollen B, Twisk J. Impact of time on improvement of outcome after stroke. *Stroke* 2006;37:2348–53.

23. Duncan PW, Goldstein LB, Horner RD, et al. Similar motor recovery of upper and lower extremities after stroke. *Stroke* 1994;25:1181–8.

24. 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;98:456–62.

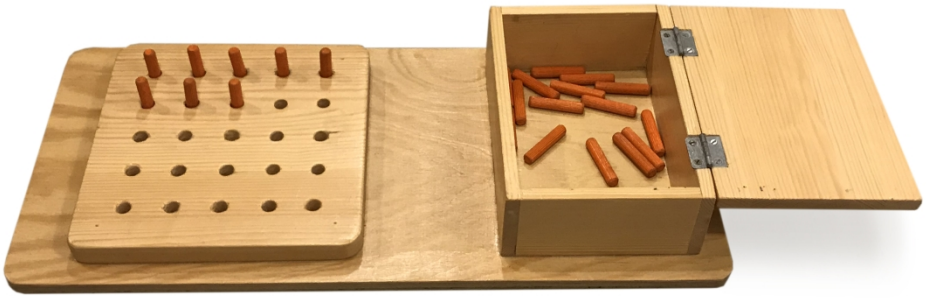
FIGURE LEGENDS

Figure 1. The twenty-five-hole peg test.

Figure 2. Flowchart of the recruitment process in the study. *Constraint-induced movement therapy.

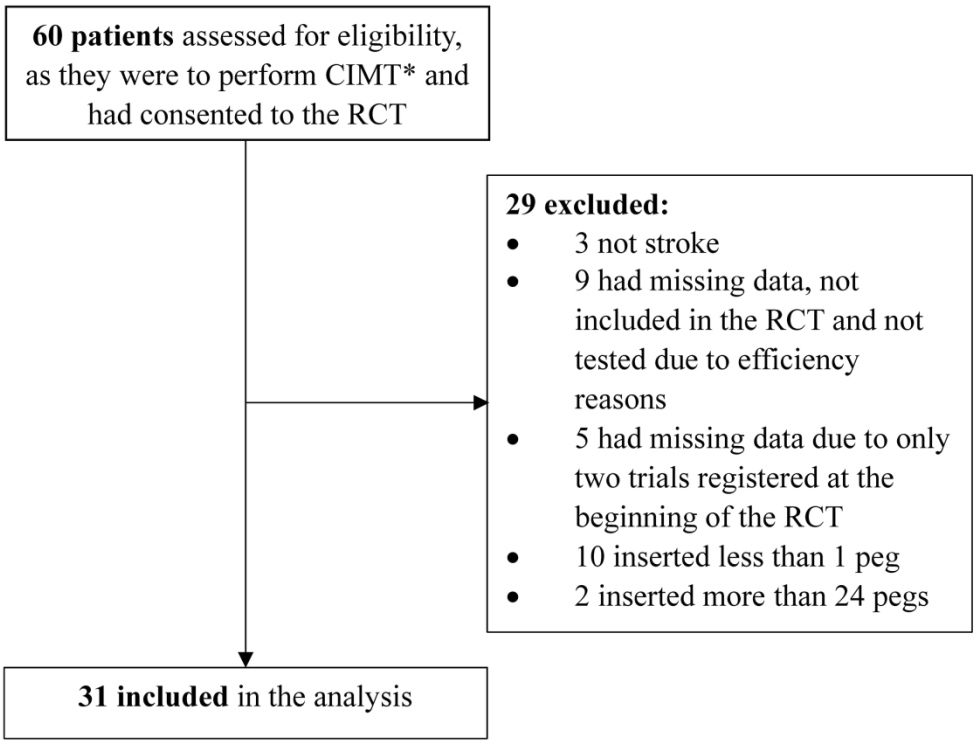
Figure 3. Bland-Altman plots of numbers of pegs from pre-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

Figure 4. Bland-Altman plots of numbers of pegs from post-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.



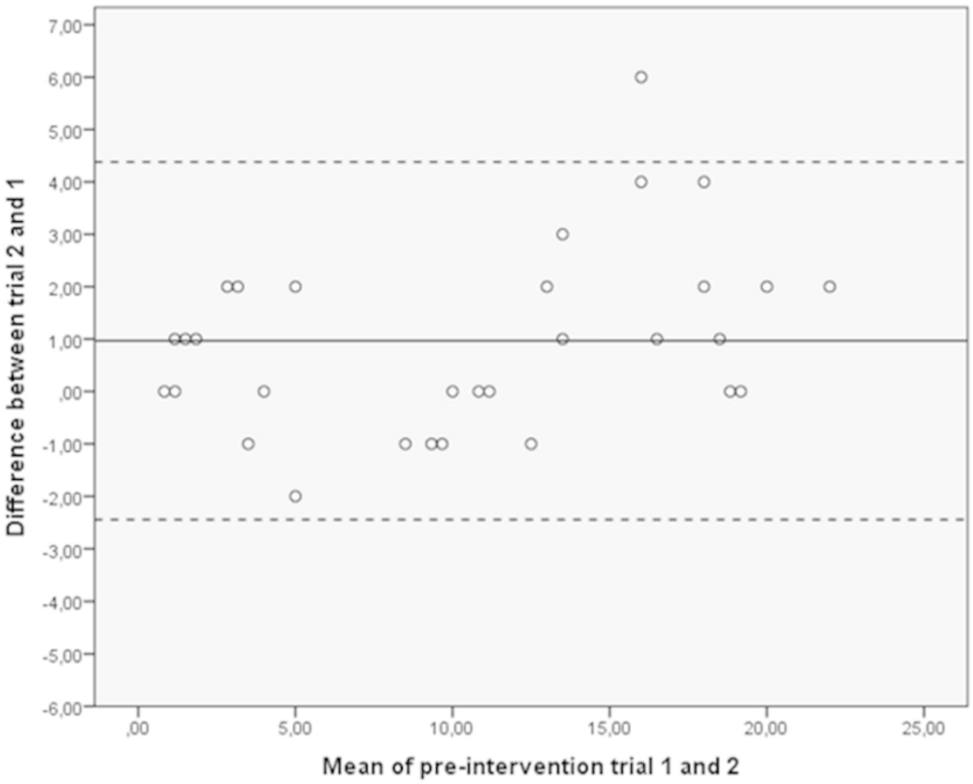
The twenty-five-hole peg test.

208x140mm (300 x 300 DPI)



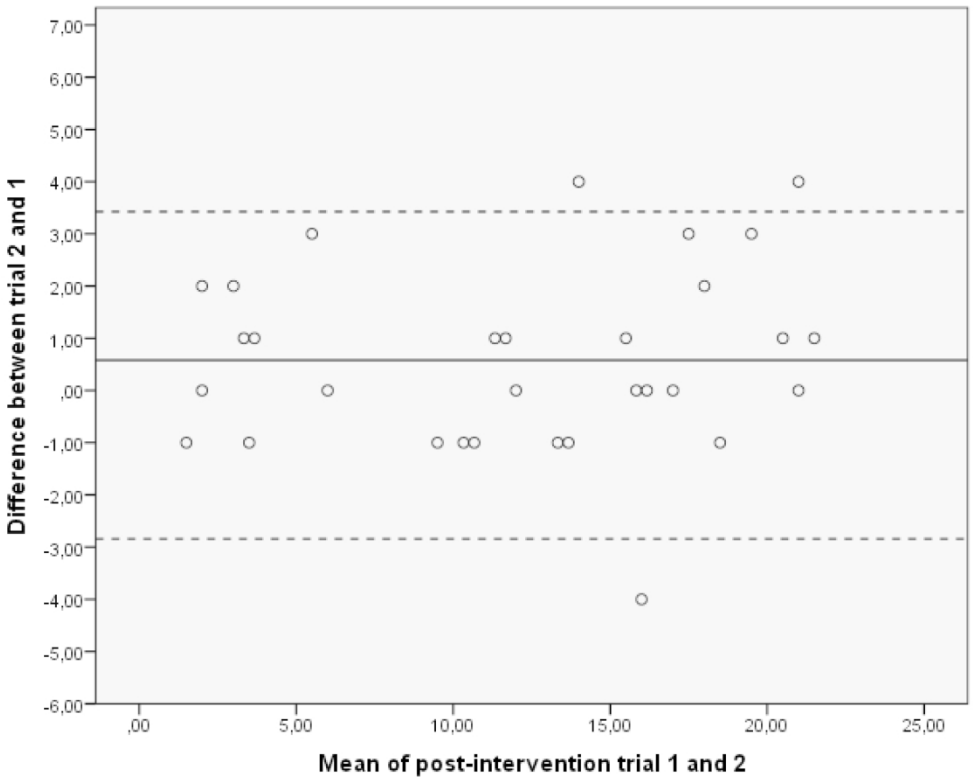
Flowchart of the recruitment process in the study. *Constraint-induced movement therapy.

121x92mm (600 x 600 DPI)



Bland-Altman plots of numbers of pegs from pre-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

140x112mm (600 x 600 DPI)



Bland-Altman plots of numbers of pegs from post-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

140x112mm (600 x 600 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	(Reliability), 1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4-5
	4	Study objectives and hypotheses	5-6
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	10, registered, refnr: ISRCTN24868616
Participants	**6**	Eligibility criteria	6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	Setting: 6 Dates: 11 Exact locations of the clinics not included.
	9	Whether participants formed a consecutive, random or convenience series	6
Test methods	**10a**	Index test, in sufficient detail to allow replication	Not applicable
	10b	Reference standard, in sufficient detail to allow replication	n.a.
	11	Rationale for choosing the reference standard (if alternatives exist)	n.a.
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	n.a.
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	n.a.
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	n.a.
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	Reliability measures 8-10
	15	How indeterminate index test or reference standard results were handled	n.a.
	16	How missing data on the index test and reference standard were handled	n.a.
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	n.a.
	18	Intended sample size and how it was determined	6
RESULTS			
Participants	**19**	Flow of participants, using a diagram	11
	20	Baseline demographic and clinical characteristics of participants	11
	21a	Distribution of severity of disease in those with the target condition	Table 1, figure 3 and 4.
	21b	Distribution of alternative diagnoses in those without the target condition	n.a.
	22	Time interval and any clinical interventions between index test and reference standard	7
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Plots instead, figure 3 and 4.
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Page 12, table 3 and 4
	25	Any adverse events from performing the index test or the reference standard	n.a.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	17

	27	Implications for practice, including the intended use and clinical role of the index test	14-17
OTHER INFORMATION			
	28	Registration number and name of registry	ISRCTN registery; referene number: ISRCTN24868616
	29	Where the full study protocol can be accessed	At the registery (above), but not detailed.
	30	Sources of funding and other support; role of funders	18

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

BMJ Open

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-032560.R2
Article Type:	Original research
Date Submitted by the Author:	13-Nov-2019
Complete List of Authors:	Granström, Fredrik; Uppsala University, Centre for Clinical Research Sörmland Hedlund, Mattias; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and rehabilitation, Physiotherapy Lindström, Britta; Umeå Universitet Medicinska fakulteten, Department of Community Medicine and Rehabilitation, Physiotherapy Eriksson, Staffan; Uppsala University, Centre for Clinical Research Sörmland; Uppsala University, Department of Neuroscience, Physiotherapy
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Neurology
Keywords:	Stroke < NEUROLOGY, measurement, Reliability, Clinical assessment, hand function, twenty-five-hole peg test

Test-retest reliability of the twenty-five-hole peg test in stroke patients

Fredrik Granström[1], Mattias Hedlund[2], Britta Lindström[2], Staffan Eriksson[1,3]

[1]Centre for Clinical Research Sörmland, Uppsala University, Sweden

[2]Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Sweden

[3]Department of Neuroscience, Physiotherapy, Uppsala University, Sweden

CORRESPONDING AUTHOR

Staffan Eriksson, Centre for Clinical Research Sörmland, Uppsala University, Kungsgatan 41, 631 88 Eskilstuna, Sweden. Telephone number: +46 16 105251, +46 73 949 99 33. Email: staffan.eriksson@germed.umu.se

KEY WORDS: clinical assessment, measurement, reproducibility of results, reliability, hand function, stroke, twenty-five-hole peg test

WORD COUNT 4060 words, 4176 words including “strengths and limitations of this study”.

ABSTRACT

Objectives: Weaknesses of the nine-hole peg test include high floor effects and a result that might be difficult to interpret. In the twenty-five-hole peg test (TFHPT), the larger number of available pegs allows for the straightforward counting of the number of pegs inserted as the result. The TFHPT provides a comprehensible result and low floor effects. The objective was to assess the test-retest reliability of the TFHPT when testing persons with stroke. A particular focus was placed on the absolute reliability, as quantified by the smallest real difference (SRD). Complementary aims were to investigate possible implications for how the TFHPT should be used and for how the SRD of the TFHPT performance should be expressed.

Design: This study employed a test-retest design including three trials. The pause between trials was approximately 10-120 seconds.

Participants, setting and outcome measure: Thirty-one participants who had suffered a stroke were recruited from a group designated for constraint-induced movement therapy at outpatient clinics. The TFHPT result was expressed as the number of pegs inserted.

Methods: Absolute reliability was quantified by the SRD, including random and systematic error for a single trial, $SRD_{2.1}$, and for an average of three trials, $SRD_{2.3}$. For the SRD measures, the corresponding smallest real difference percentage (SRD%) measure was also reported.

Results: The differences in the number of pegs necessary to detect a change in the TFHPT for $SRD_{2.1}$ and $SRD_{2.3}$ were 4.0 and 2.3, respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 36.5% and 21.3%, respectively.

Conclusions: The smallest change that can be detected in the TFHPT should be just above two pegs for a test procedure including an average of three trials. The use of an average of three trials compared to a single trial substantially reduces the measurement error.

Trial registration: ISRCTN registry, reference number ISRCTN24868616.

ARTICLE SUMMARY

Strengths and limitations of this study

- The generalizability of the results may be limited: the participants were selected because they should benefit from CIMT, and few scored above 20 pegs during the 50-second trial duration.
- Among other measures of reliability, the SRD percentage was reported; this is a good measure for comparisons between different tests, scales and populations.
- The results are presented with several different reliability measures to offer knowledge about the source of the measurement error.
- As the test-retest trials were performed within minutes, possible day-to-day variation was not captured.
- The intended practice trial was included as one of three trials in the analyses, which appears to have contributed to the learning effect.

INTRODUCTION

For a comprehensive upper limb assessment among persons with stroke, it is important to combine a measure of proximal upper limb function with a measure of hand function.[1] However, only approximately 25% of studies regarding upper limb interventions include a specific measure of hand function.[1] The nine-hole peg test (NHPT) is a common test of hand function focusing on fine manual dexterity, and it is the most common such test used in research.[1-3] Two studies of stroke populations investigated the reliability of the NHPT, and there was a large discrepancy in the reliability reported: the smallest real difference percentage (SRD%) 24% vs. 52%.[2, 3] The NHPT has mostly shown moderate to excellent correlations (0.55-0.97) with other tests and self-reports focusing on hand function, including the Action Research Arm Test, the Jebsen-Taylor Hand Function Test, and the Stroke Impact Scale (hand function domain).[4, 5] The exception is the Motor Activity Log, for which low correlations have been reported (0.23-0.33).[5]

A weakness of the NHPT is that many persons with stroke cannot reach the lower limit; i.e., a floor effect arises. Furthermore, if the number of completed pegs is used as an outcome measure, a test with only nine pegs can measure only a narrow range of hand function, resulting in profound ceiling effects.[6] Therefore, to widen the scale and avoid ceiling effects, the original NHPT expresses the result as the time needed to complete the test (including inserting and removing all the pegs).[2, 3, 7, 8] However, this approach aggravates the floor effects because tests that are not completed during the stipulated time (limits of 60 and 180 seconds have been used) are excluded.[2, 3] The maximum time could be prolonged; however, this would be time consuming, mentally strenuous and therefore possibly unethical due to the possibility of a non-completed test after a lengthy attempt. A modified NHPT is used to mitigate the floor effect while avoiding the ceiling effect; in this modified version, the result is expressed as the number of inserted pegs per unit of time, i.e., the frequency.[9] This

modified test includes only peg insertion and not peg removal. It is thus possible also to include tests that were not completed within the stipulated time limit and still measure performance on the same task across the entire range of hand function. However, it may be difficult both to interpret the frequency and to communicate it to other staff members and patients, especially to those suffering from a brain injury. The reliability of this modified test has not been investigated.

In Sweden, a similar peg test, a twenty-five-hole peg test (TFHPT), has been used in clinical practice. The larger number of available pegs makes it straightforward to count the number of pegs inserted during a stipulated time frame of 50 seconds as the test result. Thus, the TFHPT measures fine manual dexterity on a numerical scale that is easy to comprehend, with low floor effects and presumably reasonable ceiling effects (based on pre-study data). Moreover, compared to the individuals whom the original NHPT can test, individuals with worse hand function can be tested with the TFHPT.[2, 3] Of the two studies investigating the reliability of the NHPT, the one with the most generous time limit excluded all tests that were not completed in 180 seconds.[3] This limit corresponds 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. The TFHPT has not been previously described in the literature, and its reliability has not been investigated. Due to the similarity of the NHPT and the TFHPT, the underlying skill assessed with these tests is most likely the same. However, since the tests have completely different stop criteria – a time limit for the TFHPT vs. the insertion of all pegs for the NHPT – equal reliability cannot be taken for granted.[7] Thus, if the size of the measurement error related to the TFHPT is shown to be acceptable, this test may be useful in both clinical practice and research.

The overall aim of this study was to assess the test-retest reliability of the TFHPT for persons suffering from stroke. A particular focus was placed on the absolute reliability, as quantified

by the smallest real difference (SRD). Complementary aims were to investigate possible implications for how the TFHPT should be used and for how the SRD of the TFHPT performance should be expressed.

METHOD

Participants

The participants in this study were consecutively recruited in the process of screening patients eligible for inclusion in a multicentre randomized controlled trial (RCT), reference number ISRCTN24868616 at the ISRCTN registry. The patients were considered for inclusion because they were to undergo constraint-induced movement therapy (CIMT) at one of the clinics participating in the RCT. The clinics were outpatient rehabilitation clinics in the public health care system in Sweden. Data were collected at the clinics. The sample in this study consisted of included and excluded participants in the multicentre RCT. The participants were included if they had one stroke or more registered in the medical record and if TFHPT data were available from three trials before and three trials after the CIMT. Moreover, with regard to the outcome measure, a minimum of one peg and a maximum of 24 pegs inserted was necessary for inclusion. This was to avoid an untrue low measurement error from participants stable at 0 or 25 pegs inserted. These two intervals are wider, a person can be far below the floor or high over the ceiling, so measurements at these intervals should be more stable.

A minimum of 30 participants were included to obtain a sufficient number for a reliability study.[10]

Procedure and measurements

The TFHPT has twenty-five holes and pegs (Figure 1). The test used in this study consisted of a rectangular 21 cm × 45 cm board with a box containing pegs on one side and an elevated 18 cm × 18 cm area with holes on the other side. The holes were 9 mm wide and18 mm deep,

and they were spaced 20 mm apart. The box had a base of 13 cm × 18 cm and was 5 cm deep. The pegs were 40 mm long and 8 mm in diameter.

A battery of different tests was administered in this study, including the Fugl-Meyer test [11] and the Birgitta Lindmark motor assessment (BL motor assessment).[12, 13] The TFHPT was administered as the second test. The preceding test, the BL motor assessment, required approximately 30-60 minutes to administer. The tests were administered in an examination room in which only the participant and the physiotherapist were present. For the TFHPT, three trials were performed with each hand. The participants started with the less affected hand, followed by the more affected hand, i.e., the hand of investigation in this test-retest study. The pause between trials was approximately 10-120 seconds. The board was placed at a distance favoured by the participant with the centre row of holes centred towards the navel and the box side oriented towards the tested hand. The starting position was with both hands on the board, and time keeping began upon first hand contact with the pegs. We gave participants the following instructions:

1. I want you to pick up one peg at a time and insert them in the holes of the board.
2. Use only the right/left hand; you can only use the other hand to steady the board.
3. You can fill the holes in any order you desire.
4. We start with a practice trial.
5. You have 50 seconds to insert as many pegs as you can. After 50 seconds, the trial is terminated.
6. Are you ready? Ready, set, go!
7. After the practice trial: This was practice; now come the two actual test trials, where the results are recorded. Repeat step 6.

The test-retest reliability of the TFHPT was assessed on two separate occasions, i.e., before and after a two-week training period (the CIMT). The same procedure was used on both of

these occasions, and for each participant, all tests were administered by the same physiotherapist. The assessment after the CIMT period was performed as an internal validation.

Two physiotherapists, SE and BL, administered the tests in this study. SE has general experience with persons suffering from stroke and experience administering the original NHPT. BL has extensive experience with persons suffering from stroke, including administering the original NHPT. Background data from medical records were collected by staff at the clinics.

Statistics

All three trials were used in the analyses, although the first trial was introduced to the participants as a practice trial. Analyses of pre-intervention data and post-intervention data were performed separately.

Bland-Altman plots of trials one and two provided a graphic description of the data variability. The mean of trials one and two was plotted against the difference between trials two and one for each subject. Heteroscedasticity – i.e., an association between the random error and the magnitude of measurements [14] – was investigated with pairwise comparisons of trials using Koenker's [15] studentized test, which is useful for small samples and skewed data. Heteroscedasticity is indicated by a significant result.

Measurement error can be either random or systematic. In random error, there is no pattern of variability between trials, whereas in systematic error, the measurements vary in a non-random way; i.e., the mean values between the trials differ.[16] To investigate whether there was a systematic error in test scores, one-way repeated measures ANOVA was used to detect potential between trial effects.

Reliability is a term that describes how the measurement result of an instrument is affected by measurement error.[6, 14] Reliability can be quantified as either relative or absolute.[6] *Relative reliability* refers to the consistency of the positions of measurements relative to those of others within the tested group, and it is quantified using several intra-class correlation coefficients (ICCs).[16, 17] In ICCs, between-subject variability is related to the within-subject variability by a ratio.[16] Thus, ICCs are sensitive to the degree of between-subject variability, and with all other things being equal, a more heterogeneous sample (i.e., a larger between-subject variability) produces higher ICC values.[16, 17] The concept of *absolute reliability* refers to the consistency of measurements within individuals.[6, 16] Measurement error, quantified as within-subject standard deviations in repeated tests, is a common measure of absolute reliability [6, 14, 16-18] and is called the standard error of measurement (SEM). SRD is an extension of the SEM, and it can be seen as the smallest detectable difference, with 95% certainty, using a test instrument on an individual.[16]

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. 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. *The first* figure in the ICC designation represents the type of ICC model.[16] $ICC_{2.1}$ and $ICC_{2.3}$ are calculated from a two-way random effect model and incorporate both systematic and random error, whereas $ICC_{3.3}$ is calculated from a two-way fixed effect model and incorporates only random error.[16, 19] Thus, the less systematic error contributes to the total error, the closer $ICC_{2.3}$ is to $ICC_{3.3}$.[16] *The second* figure in the ICC designation 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.[16] $ICC_{2.1}$ represents the reliability of a test procedure in which the subject is tested with a single trial on a test occasion.[16] $ICC_{2.3}$ and $ICC_{3.3}$ represent the reliability of a test procedure in which the

subject is tested with three trials on a test occasion and the score is expressed as the average of these trials.

To estimate *absolute reliability*, the SEM, SRD and SRD percentage (SRD%) were calculated for each of the three different ICC measures ($ICC_{2.1}$, $ICC_{2.3}$, and $ICC_{3.3}$), resulting in the corresponding properties $SEM_{2.1}$, $SRD_{2.1,}$ $SRD\%_{2.1}$, $SEM_{2.3}$, $SRD_{2.3}$, $SRD\%_{2.3,}$ $SEM_{3.3}$, $SRD_{3.3,}$ and $SRD\%_{3.3}$. The SEM was calculated according to $SEM = \mathrm{SD}\sqrt{(1 - ICC)}$, where SD was calculated from the total sum of squares (SS_{TOTAL}) in the ANOVA table generated in the ICC analyses as $\sqrt{\mathrm{SStotal}/(n-1)}$.[16] The SRD was calculated using the formula 1.96 × SEM × $\sqrt{2}$, where 1.96 is related to the 95% CI and $\sqrt{2}$ refers to the error of two measurements.[16] The SRD% was calculated by dividing the SRD value by the grand mean multiplied by 100.[2, 10] This value is independent of measurement units and is indexed to the mean value of the observations from which it was derived. It is therefore a good measure for comparisons between different tests, scales and populations.[10, 14, 17] An SRD% of 30% has been suggested as an acceptable level of reliability.[20]

Because estimates of absolute reliability vary with the type of ICC value, some caution is warranted when comparing them with measures from other studies.[16] Therefore, $SEM_{\text{mean square error term (MSE)}} = \sqrt{MSE}$ was also calculated, where MSE (this term is called residual error by Hopkins and the mean square residual in the SPSS output) was taken from the ANOVA table of the ICC calculation.[16, 17] This SEM measure represents the reliability of a test procedure in which the subject is tested with a single trial on a test occasion, and it is a pure measure of random error.[16] SRD_{MSE} and $SRD\%_{MSE}$ were also derived from SEM_{MSE}.

The analysis of test-retest reliability was pre-planned. SPSS version 21 was used to calculate ICC and ANOVA. The alpha level was set to 0.05.

Patient and public involvement

No patients or members of the public were involved in the development or design of this study.

RESULTS

In this study, participants were recruited between January 2011 and September 2014. Of 60 eligible patients, 29 were excluded for any of the following reasons: not suffering a stroke, missing data, and yielding either below the minimum or above the maximum number of inserted pegs (Figure 2). This yielded 31 participants (21 men and 10 women) for inclusion in the analysis, with a mean ± SD age of 66 ± 9 years (Table 1). The two eligible patients who were excluded because they exceeded the permitted maximum number of pegs inserted completed the 25 pegs in their best trial within 49.3 seconds and 39.4 seconds. Of the ten patients who were excluded because they fell below the minimum number of inserted pegs, six inserted at least one peg in one of the trials. Data were collected from 17 and 14 participants by the two physiotherapists (third and last author, respectively) at seven clinics.

Table 1. Characteristics of participants at pre-intervention trials

Participants	**N=31**
Age (years), mean ± SD[a]	66 ± 9
Men/women, n[b]	21/10
Time since stroke (months), median (IQR[d]), (min-max)	17 (8–24), (2–70)
Previous dominant hand more affected by stroke, n	19
TFHPT[c], mean of three trials (number of pegs), mean ± SD, (min-max)	10.8 ± 6.8, (1–22.7)
Fugl-Meyer test (score), median (IQR[d]), (min-max)	46 (41–53), (29–62)

More than one stroke, n	3

[a]Standard deviation, [b]number of participants, [c]twenty-five-hole peg test, [d]interquartile range

A graphic description of the data variability can be seen in the Bland-Altman plots (Figures 3 and 4). According to Koenker's studentized test, the measurement error was not affected by heteroscedasticity (Table 2).

Table 2. Results of the Koenker's studentized test, n=31

Pairwise test of trials	**Pre-intervention trials**		**Post-intervention trials**	
	Chi-square	P-value	Chi-square	P-value
1-2	1.33	0.25	0.41	0.52
2-3	0.05	0.83	1.38	0.24
1-3	0.28	0.60	0.20	0.66

For *pre-intervention* trials, the mean values ± SDs for trials 1, 2 and 3 were 10.0 ± 6.5, 11.0 ± 7.1, and 11.5 ± 6.9, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 10.9 and $p < 0.001$. Post hoc tests revealed differences between trials 2 and 1 and between trials 3 and 1, with mean differences (95% CIs) of 1.0 (0.3–1.6) and 1.5 (0.9–2.2), respectively.

For *post-intervention* trials, the mean values ± SD for trials 1, 2 and 3 were 11.8 ± 6.5, 12.4 ± 6.7, and 12.5 ± 6.8, respectively. The one-way repeated measures ANOVA revealed a main effect between trials with F (2, 29) = 4.1 and $p = 0.027$. Post hoc tests revealed a difference between trials 3 and 1, with a mean difference (95% CIs) of 0.6 (0.2–1.1).

For *pre-intervention* trials, $ICC_{2.3}$ (95% CI) was 0.99 (0.97–0.99) (Table 3). The SRDs incorporating random and systematic error, $SRD_{2.1}$ and $SRD_{2.3}$, were 4.0 and 2.3 pegs, respectively. The corresponding SRD% values for $SRD_{2.1}$ and $SRD_{2.3}$ were 36.5% and 21.3%, respectively. The SRD incorporating only random error, $SRD_{3.3}$, was 2.0 pegs.

For *post-intervention* trials, $SRD_{2.1}$ and $SRD_{2.3}$ were 3.2 and 1.8 pegs, respectively (Table 4). $SRD_{3.3}$, was 1.8 pegs.

Table 3. Results of reliability measures for pre-intervention trials

	ICC[a] (95% CI)	SEM[b], n[c]	SRD[d], n[c]	SRD%[e]
$ICC_{2.1}$	0.96 (0.90–0.98)	1.4	4.0	36.5
$ICC_{2.3}$	0.99 (0.97–0.99)	0.8	2.3	21.3
$ICC_{3.3}$	0.99 (0.98–0.99)	0.7	2.0	18.3
Derived from MSE[f]		1.3	3.5	32.1

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

Table 4. Results of reliability measures for post-intervention trials

	ICC[a] (95% CI)	SEM[b], n[c]	SRD[d], n[c]	SRD%[e]
$ICC_{2.1}$	0.97 (0.95–0.98)	1.1	3.2	25.9
$ICC_{2.3}$	0.99 (0.98–1.0)	0.7	1.8	15.0
$ICC_{3.3}$	0.99 (0.98-1.0)	0.7	1,8	15.0
Derived from MSE[f]		1.1	3.1	25.5

[a]Intra-class correlation coefficient.

[b]Standard error of measurement derived from ICC2.1, 2.3, 3.3 and MSE.

[c]Number of pegs.

[d]Smallest real difference derived from ICC2.1, 2.3, 3.3 and MSE.

[e]SRD percentage derived from ICC2.1, 2.3, 3.3 and MSE.

[f]Mean square error term.

DISCUSSION

This study indicated that in a selected group of persons suffering from stroke, the absolute test-retest reliability of the TFHPT was at a level that can be considered acceptable for measures representing an average of three trials and incorporating systematic error.

To assess implications for the use of the TFHPT and to determine which SRD measure best captures the absolute reliability, three issues were considered: 1) whether to use single or average measures, 2) whether to include systematic error in the assessments, and 3) whether to take heteroscedasticity into account.

Comparing $SRD_{2.1}$ to $SRD_{2.3}$ revealed that the use of an average of three trials reduced the measurement error by approximately 1.5 pegs compared to the use of a single trial. This finding suggests that the reliability of the TFHPT is substantially improved when an average of three trials is used.

Comparing $SRD_{2.3}$ to $SRD_{3.3}$, where $SRD_{3.3}$ incorporates only random error, revealed that the contribution of the systematic error was approximately 0.3 pegs of the total 2.3 pegs when the average of three trials was used.[16] Although the systematic error was small compared to the random error it was not small enough to be overlooked in the assessment of reliability. Therefore, $SRD_{2.3}$ is preferable to $SRD_{3.3}$ for measuring the reliability of the TFHPT.

The choice of SRD% instead of SRD is dependent on whether the measurement error is affected by heteroscedasticity. A measure of absolute reliability, expressed as an absolute number of pegs, can over- or underestimate the number of pegs necessary to demonstrate an improvement for an individual.[14, 17] The reason is that the random error of measurements often increases with the magnitude of the measurements (i.e., heteroscedasticity).[14, 17] As a remedy, the use of a relative measure of absolute reliability, such as SRD%, has been proposed.[14, 17] However, the lack of heteroscedasticity detection suggests that both $SRD_{2.3}$ and $SRD\%_{2.3}$ are appropriate measures of reliability for the TFHPT.[14]

The results for $SRD_{2.3}$ and $SRD\%_{2.3}$ were 2.3 pegs and 21.3%, respectively. The value of $SRD\%_{2.3}$ fell within the 30% level, which has been suggested as acceptable.[20] The 30% level seems high in this context, with persons affected by stroke in a chronic stage; from a clinical viewpoint, our opinion is that the results of 21.3% and 2.3 pegs in this study indicate a barely acceptable level of absolute reliability. For a favourable level, we believe that a mean number consisting of approximately 1.5 pegs is desirable.

The relative test-retest reliability, as measured by $ICC_{2.3,}$ was 0.99, which seems excellent. The discrepancy between the level of the relative and the absolute reliability is most likely caused by the heterogeneity in this study population (Figure 3, Table 1) which inflates the relative reliability.[16]

The level of the relative absolute test-retest reliability (SRD%), the most comparable measure, observed for the TFHPT in this study (21.3%) is better than what Chen et al.[2] reported (54%) and is at approximately the same level as Ekstrand et al.[3] reported (24%) for the NHPT. Although the SRD% measures reported in the studies by Chen et al. and by Ekstrand et al. were calculated in different ways than the $SRD\%_{2.3}$ reported in this study, the measures used in these three studies are fairly equivalent.[16] Several methodological differences between these three studies could have affected the results.[2, 3] *First*, the results of the TFHPT and NHPT were

measured using different scales, and the use of time for completion of the test in the NHPT should accommodate more variability than the peg count used in the TFHPT. However, the SRD% results should still be comparable between the TFHPT and the NHPT because this relative measure of absolute reliability adjusts for different scales and study populations.[14, 17] *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. Thus, the TFHPT may seem more reliable because of possible random error from day-to-day variation in performance which was not captured in this study.[17, 21] *Third*, the longer time since stroke in this study of the TFHPT compared to the study of NHPT by Chen et al.[2] may have resulted in seemingly better reliability for the TFHPT because of a more stable level of hand function. In the study by Chen et al., a systematic error may have originated in recovery from stroke in the 3-5 days between the test and retest trials because the time since stroke was 3 months or less for a quarter of the study sample.[22]

The implications of the results of this study are that the TFHPT can be used in a clinical situation to detect changes in a patient's hand function. The test procedure should employ an average of three trials on each occasion, and a change of 2.3 pegs or more between two occasions should be considered real improvement/worsening. Furthermore, it seems that the ceiling effects in the TFHPT can be considered acceptable. Only two of the persons assessed for eligibility inserted 25 pegs, and only one of them actually hit the ceiling because according to this individual's best times for completion of the 25 pegs, he/she would have been able to insert more pegs if available. This occurred in a sample where approximately a quarter of the included participants suffered from mild impairment of arm and hand function as judged by the Fugl-Meyer test.[23, 24]

There was a tendency towards improved reliability after the CIMT period, which was due to decreased systematic error and decreased random error. The decreased systematic error can be

observed in the main effects of trial in the ANOVA results.[16] The decreased systematic error is most likely due to a decreased learning effect when the participants had previous experience in the test. The learning effect is indicated by the increases in the mean values over the trials, especially over trials 1-2, and the decreased learning effect is indicated by the less pronounced increase in the post-intervention trials.[14, 17] The lower random error can be observed from the lower $SRD_{3.3}$ results in the post-intervention trials.[16] The cause of the decreased random error is less clear, but it could also be attributed to the decreased systematic error.[17] This is because the magnitude of the learning effect probably differs between individuals, which will show as random error. Furthermore, it is likely that the $SRD_{2.3}$ result of 2.3 pegs for TFHPT could, in reality, be adjusted downwards. A peg test is often used to evaluate a rehabilitation period; because the error is smaller in the post-intervention trials, the "true" SRD may be somewhere between the SRDs of the pre- and post-intervention trials (2.3 vs 1.8).

Four weaknesses of this study should be considered. The sample included a relatively low number of participants with few observations above 20 pegs and participants who were selected because they should benefit from CIMT.[10, 17] These sample qualities may thus, to some degree, hinder the generalization of the results to other groups of people suffering from stroke. In addition, the intended practice trial was included as one of three trials in the analyses which appears to have contributed to systematic error through an increased learning effect, indicated by a large increase in the mean values between trials 1 and 2.[14, 16, 17] Thus, to mitigate the learning effect, a practice trial preceding regular trials is recommended. Moreover, the possible day-to-day variation was not captured in the present study design. The advantage of this approach is that it yields a pure result for measurement error for the instrument in this population; the disadvantage is that the result is less clinically applicable.[17, 21] Finally, in this study, sensitivity to change and validity were not examined. However, the criterion validity for NHPT has mostly shown a moderate to excellent level [4, 5] and the underlying skill

assessed with the TFHPT is most likely the same. A high reliability level is a prerequisite for high validity, and because the reliability of the TFHPT was at the same level as that of the NHPT, the criterion validity should also be similar.[21]

In conclusion, our results suggest that the smallest detectable difference between two assessments using a test procedure with an average of three trials conducted by a single tester should be just above two pegs with the TFHPT. Furthermore, to reach an acceptable level of measurement error, the use of the average of multiple trials is crucial. Future research should focus on optimizing the number of trials.

ACKNOWLEDGEMENTS

The authors thank the staff at the clinics for the extra work associated with participating in the RCT. The authors thank Håkan Littbrand for input on the conception of the study and Monika Edström for administrative work with the study.

AUTHOR CONTRIBUTIONS

Study design and data interpretation: SE, FG, BL and MH. Data acquisition: SE, BL and MH. Statistical analysis: SE and FG. Drafting and finalization of the manuscript: SE. Critical revision of the manuscript: FG, BL and MH. Final approval of the submitted manuscript: FG, BL and MH.

COMPETING INTERESTS

None declared.

FUNDING

The Norrbacka-Eugenia Foundation; The Swedish STROKE-Association; The Stroke Foundation of Northen Sweden; and the Centre for Clinical Research Sörmland, Uppsala University.

DATA SHARING

No additional data are available.

PATIENT CONSENT

Written informed consent was obtained from the participants.

ETHICS APPROVAL

The Regional Ethical Review Board in Umeå, reference 09-104M, with additional approval Dnr 2010/314-32M, Dnr 2011-244-32M, and 2012-235-32M.

REFERENCES

1. Santisteban L, Teremetz M, Bleton JP, et al. Upper limb outcome measures used in stroke rehabilitation studies: a systematic literature review. *PLoS One* 2016;11:e0154792.
2. 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;23:435–40.
3. Ekstrand E, Lexell J, Brogardh C. Test-retest reliability and convergent validity of three manual dexterity measures in persons with chronic stroke. *PM R* 2016;8:935–43.
4. 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;33:96–103.
5. Lin KC, Chuang LL, Wu CY, et al. Responsiveness and validity of three dexterous function measures in stroke rehabilitation. *J Rehabil Res Dev* 2010;47:563–71.
6. Carter RE, Lubinsky J, Domholdt E. *Rehabilitation Research: Principles and Applications*. St. Louis, MO: Elsevier-Saunders, 2016:239–244.
7. Mathiowetz V, Weber K, Kashman N, et al. Adult norms for the Nine-Hole Peg Test of Finger Dexterity. *Occup Ther J Res* 1985;5:25–38.
8. Grice KO, Vogel KA, Le V, et al. Adult norms for a commercially available Nine Hole Peg Test for finger dexterity. *Am J Occup Ther* 2003;57:570–3.
9. Heller A, Wade DT, Wood VA, et al. Arm function after stroke: measurement and recovery over the first three months. *J Neurol Neurosurg Psychiatry* 1987;50:714–9.
10. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil* 2005;84:719–23.
11. Fugl-Meyer AR, Jaasko L, Leyman I, et al. The post-stroke hemiplegic patient. 1. A method for evaluating of physical performance. *Scand J Rehabil Med* 1975;7:13-31.

12. Lindmark B, Hamrin E. Evaluation of functional capacity after stroke as a basis for active intervention. Validation of a modified chart for motor capacity assessment. *Scand J Rehabil Med* 1988;20:111–5.

13. Lindmark B, Hamrin E. Evaluation of functional capacity after stroke as a basis for active intervention. Presentation of a modified chart for motor capacity assessment and its reliability. *Scand J Rehabil Med* 1988;20:103–9.

14. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217–38.

15. Koenker R. A note on studentizing a test for heteroscedasticity. *J Econom* 1981;17:107–12.

16. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231–40.

17. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30:1–15.

18. Bland JM, Altman DG. Measurement error. *BMJ* 1996;313:744.

19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.

20. Huang SL, Hsieh CL, Lin JH, et al. Optimal scoring methods of hand-strength tests in patients with stroke. *Int J Rehabil Res* 2011;34:178–80.

21. Sim J, Wright C. *Research in Health Care: Concepts, Designs and Methods*. Cheltenham, England: Nelson Thornes Ltd 2002: 123–33.

22. Kwakkel G, Kollen B, Twisk J. Impact of time on improvement of outcome after stroke. *Stroke* 2006;37:2348–53.

23. Duncan PW, Goldstein LB, Horner RD, et al. Similar motor recovery of upper and lower extremities after stroke. *Stroke* 1994;25:1181–8.

24. 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;98:456–62.

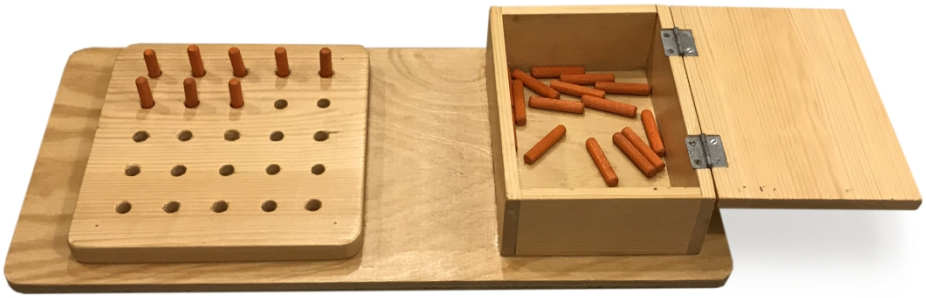
FIGURE LEGENDS

Figure 1. The twenty-five-hole peg test.

Figure 2. Flowchart of the recruitment process in the study. *Constraint-induced movement therapy.

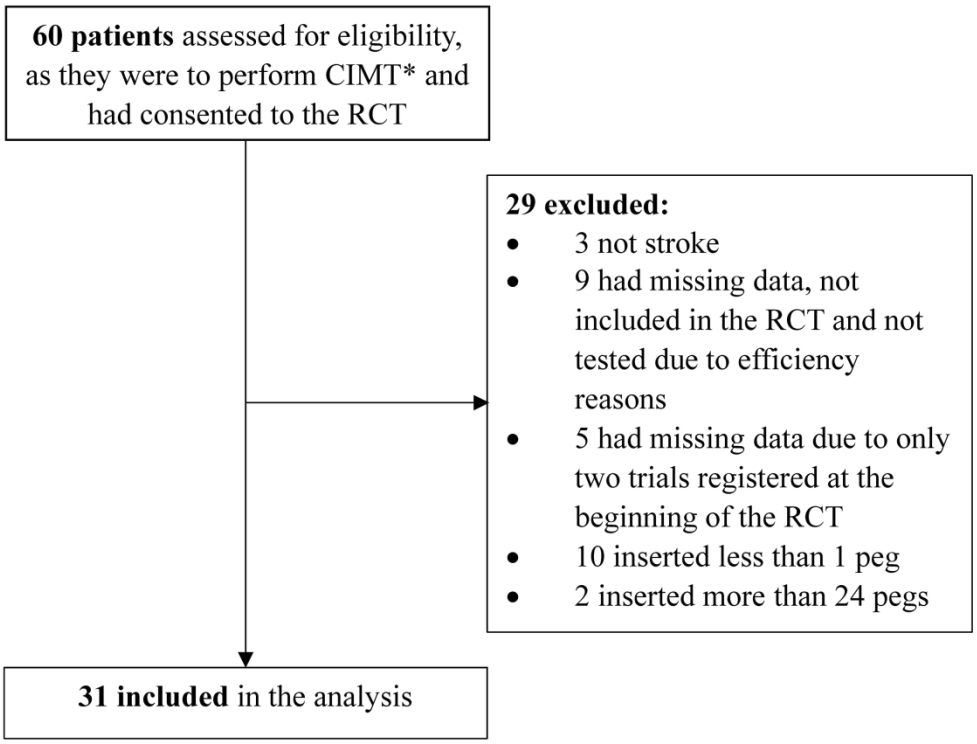
Figure 3. Bland-Altman plots of numbers of pegs from pre-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

Figure 4. Bland-Altman plots of numbers of pegs from post-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.



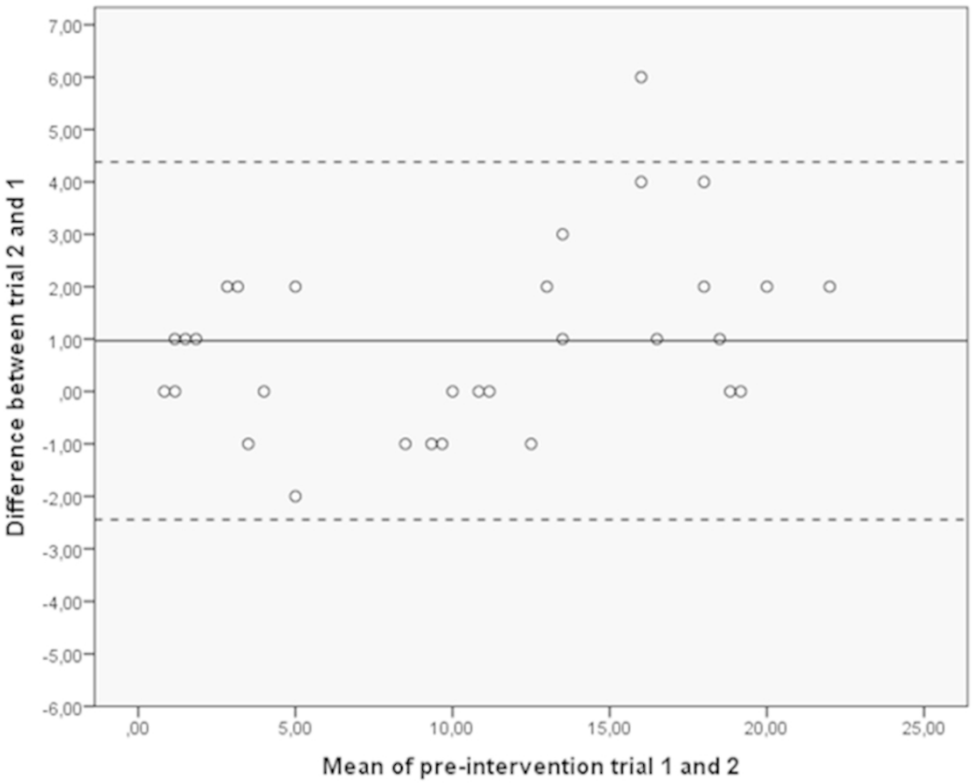
The twenty-five-hole peg test.

208x140mm (300 x 300 DPI)



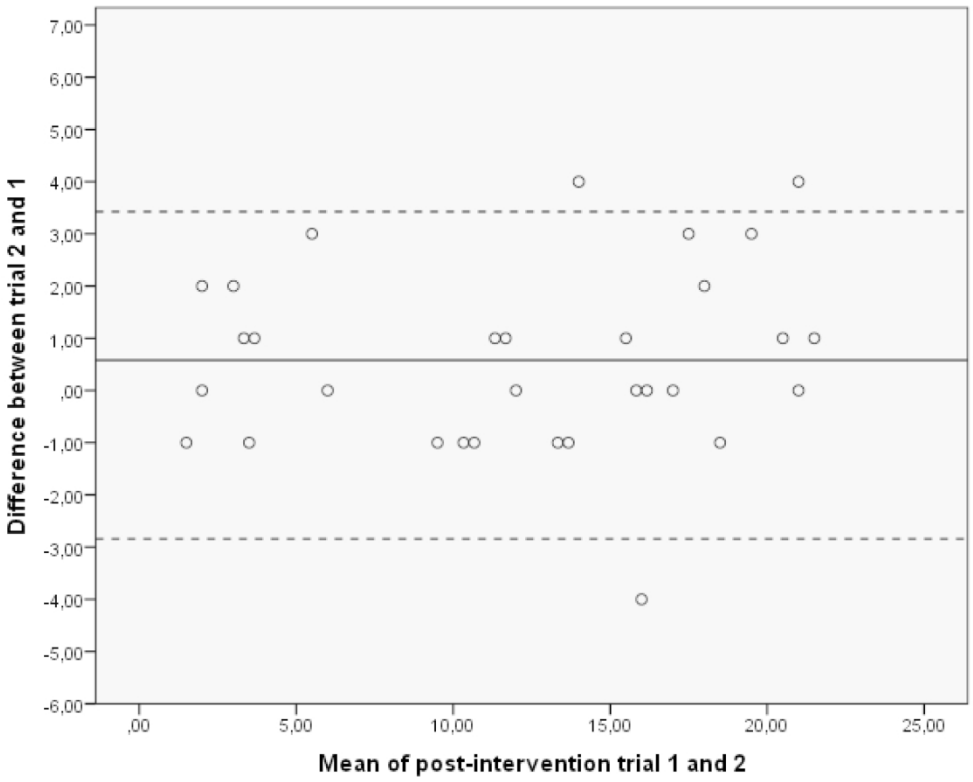
Flowchart of the recruitment process in the study. *Constraint-induced movement therapy.

121x92mm (600 x 600 DPI)



Bland-Altman plots of numbers of pegs from pre-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

140x112mm (600 x 600 DPI)



Bland-Altman plots of numbers of pegs from post-intervention trials. The mean of trials 1 and 2 was plotted against the difference of trials 2 and 1 for each subject. The centre line displays the mean difference for the group between trials 2 and 1. The upper and lower confidence limits were calculated as the mean difference ± SD of the mean difference × 1.96.

140x112mm (600 x 600 DPI)

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	(Reliability), 1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4-5
	4	Study objectives and hypotheses	5-6
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	10, registered, refnr: ISRCTN24868616
Participants	**6**	Eligibility criteria	6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	Setting: 6 Dates: 11 Exact locations of the clinics not included.
	9	Whether participants formed a consecutive, random or convenience series	6
Test methods	**10a**	Index test, in sufficient detail to allow replication	Not applicable
	10b	Reference standard, in sufficient detail to allow replication	n.a.
	11	Rationale for choosing the reference standard (if alternatives exist)	n.a.
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	n.a.
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	n.a.
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	n.a.
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	Reliability measures 8-10
	15	How indeterminate index test or reference standard results were handled	n.a.
	16	How missing data on the index test and reference standard were handled	n.a.
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	n.a.
	18	Intended sample size and how it was determined	6
RESULTS			
Participants	**19**	Flow of participants, using a diagram	11
	20	Baseline demographic and clinical characteristics of participants	11
	21a	Distribution of severity of disease in those with the target condition	Table 1, figure 3 and 4.
	21b	Distribution of alternative diagnoses in those without the target condition	n.a.
	22	Time interval and any clinical interventions between index test and reference standard	7
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Plots instead, figure 3 and 4.
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Page 12, table 3 and 4
	25	Any adverse events from performing the index test or the reference standard	n.a.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	17

	27	Implications for practice, including the intended use and clinical role of the index test	16
OTHER INFORMATION			
	28	Registration number and name of registry	ISRCTN registery; referene number: ISRCTN24868616
	29	Where the full study protocol can be accessed	At the registery (above), but not detailed.
	30	Sources of funding and other support; role of funders	18-19

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.