J Rehabil Med 2005; 37: 75–82

RELIABILITY OF GAIT PERFORMANCE TESTS IN MEN AND WOMEN WITH HEMIPARESIS AFTER STROKE

Ulla-Britt Flansbjer,[1,2] Anna Maria Holmbäck,[3] David Downham,[4] Carolynn Patten[5,6] and Jan Lexell[1,2,7]

From the [1]Department of Rehabilitation, Lund University Hospital, Lund, Sweden, [2]Department of Community Medicine, Lund University, Malmö, Sweden, [3]Department of Physical Therapy, Lund University Hospital, Lund, Sweden, [4]Department of Mathematical Sciences, University of Liverpool, Liverpool, UK, [5]Rehabilitation Research & Development Center/153, VA Palo Alto Health Care System, Palo Alto, CA, USA, [6]Department of Orthopaedic Surgery, Stanford University School of Medicine, Stanford, CA, USA and [7]Department of Health Sciences, Luleå University of Technology, Boden, Sweden

Objective: **To assess the reliability of 6 gait performance tests in individuals with chronic mild to moderate post-stroke hemiparesis.**

Design: **An intra-rater (between occasions) test-retest reliability study.**

Subjects: **Fifty men and women (mean age 58 ± 6.4 years) 6–46 months post-stroke.**

Methods: **The Timed "Up & Go" test, the Comfortable and the Fast Gait Speed tests, the Stair Climbing ascend and descend tests and the 6-Minute Walk test were assessed 7 days apart. Reliability was evaluated with the intraclass correlation coefficient ($ICC_{2,1}$), the Bland & Altman analysis, the standard error of measurement (SEM and SEM%) and the smallest real difference (SRD and SRD%).**

Results: **Test-retest agreements were high ($ICC_{2,1}$ 0.94–0.99) with no discernible systematic differences between the tests. The standard error of measurement (SEM%), representing the smallest change that indicates a real (clinical) improvement for a group of individuals, was small (< 9%). The smallest real difference (SRD%), representing the smallest change that indicates a real (clinical) improvement for a single individual, was also small (13–23%).**

Conclusion: **These commonly used gait performance tests are highly reliable and can be recommended to evaluate improvements in various aspects of gait performance in individuals with chronic mild to moderate hemiparesis after stroke.**

Key words: activities of daily living, cerebrovascular accident, gait, outcome assessment, rehabilitation, reproducibility of results, research design, walking.

J Rehabil Med 2005; 37: 75–82

Correspondence address: Ulla-Britt Flansbjer, Department of Rehabilitation, Lund University Hospital, Orupssjukhuset, SE-221 85, Lund, Sweden.
E-mail: ulla-britt.flansbjer@skane.se

Submitted February 6, 2004; accepted May 21, 2004

INTRODUCTION

Hemiparesis is one of the most common impairments after stroke and contributes significantly to reduced gait performance (1). Although a majority of stroke patients will be able to walk independently (2), many cannot walk with sufficient speed and endurance to enable them to resume all their daily activities (3–5). The retraining of locomotor skills in order to improve gait performance is therefore one of the main components in stroke rehabilitation (6, 7).

To assess gait performance after stroke and changes following interventions, clinically and scientifically robust measurement tools are needed (8). In particular, measurement tools must be reliable, where reliability refers to the consistency of measurements and the relative absence of measurement errors (9). There is consensus that several statistical methods and indices, covering both agreement between measurements, systematic changes in the mean and measurement errors, are required to fully assess the reliability of a measurement tool (10–15). Importantly, a measurement tool can be considered highly reliable, as indicated by the various statistical methods and indices, but may not be sufficiently sensitive to detect a real (clinical) improvement following, for example, an intervention. Results from the reliability analysis can be used to define limits for the smallest change that indicate such improvements, both for a group of patients and for individual patients.

A variety of gait performance tests have been used in stroke patients (16, 17). Several of the tests have been analysed for intra-rater reliability in stroke patients (8, 16, 18–24). Even though the tests were found to be reliable, the statistical analyses were not sufficiently comprehensive and only one or a few of the gait performance tests were evaluated in each study. Moreover, no study has defined limits for the smallest change that indicate a real (clinical) improvement in stroke patients. Further studies of the reliability of gait performance tests in men and women with hemiparesis after stroke are therefore needed.

The overall aim of this study was to assess the reliability of gait performance tests in individuals with post-stroke hemiparesis. A set of statistical methods was used to evaluate comprehensively the intra-rater (between occasions) test-retest reliability of 6 different gait performance tests in 50 men and women with mild to moderate hemiparesis 6 months or more after stroke. Limits were also defined for the smallest change that indicate a real (clinical) improvement following, for example, an intervention both for a group of stroke patients and for individual stroke patients.

 ISSN 1650–1977
DOI 10.1080/16501970410017215

SUBJECTS AND METHODS

Participants

A sample of 50 community-dwelling subjects (38 men, 12 women) was selected from the Comprehensive Integrated Rehab Unit database in the Department of Rehabilitation, Lund University Hospital. The ages for the men were (mean, (SD), range) 59 (7), 46–72 years, and for the women 58 (5), 50–66 years. The times from stroke onset until the first test session were 16 months (± 5, 6–46) for the men and 18 months (± 5, 6–33) for the women. Clinical characteristics of the 50 subjects are presented in Table I. All subjects met the following inclusion criteria: (i) hemiparesis resulting from an ischaemic or haemorrhagic stroke; (ii) a minimum of 6 months and a maximum of 48 months post-stroke; (iii) ability to walk at least 300 metres with or without a unilateral assistive device; (iv) ability to understand both verbal and written information; (v) medically stable with no other diseases that significantly influenced gait performance; and (vi) discharged from interdisciplinary rehabilitation services. All 50 subjects were contacted by telephone, received written information and thereafter gave their informed consent. Prior to the first test session, all subjects completed a questionnaire, which provided demographic and medical information. All subjects were checked by the responsible physician (JL). The Ethics Research Committee of Lund University, Lund, Sweden approved the study.

Pre-test assessments

To characterize the group, each subject was interviewed and scored with the Functional Independence Measure motor domain (FIM; Swedish version of FIM^{SM}) (25) prior to the first test session. In addition, the rehabilitation admission and discharge FIM motor scores were retrieved from the database. The occurrence of spasticity in the lesional leg was assessed with the Modified Ashworth scale (MAS) (26) before each of the 2 test sessions. The MAS is a 6-point rating scale, ranging from 0 (no increase in tone, both low and normal tone) to 5 (the limb is rigid in flexion or extension). The subjects were tested in a supine position with shoes and ankle-foot orthosis removed.

Gait performance tests

Each subject underwent the following 6 gait performance tests: the Timed "Up & Go" test (TUG), the Comfortable and the Fast Gait Speed tests (CGS and FGS), the Stair Climbing ascend and descend tests (SCas and SCde) and the 6-Minute Walk test (6MW). These tests were first applied in healthy elderly people and then adopted for stroke patients. Each test followed the original description. The tests were explained succinctly to each subject. No verbal encouragement was given during the tests. Throughout each session, subjects wore comfortable shoes. The use of orthosis and assistive device has varied in previous studies (19, 24, 27). Dean et al. (27) found no significant between-group effects for walking speed measured using preferred assistive device. Therefore, subjects in the present study were allowed to use, if needed, their ankle-foot orthosis and their assistive device: 7 subjects used their ankle-foot orthosis, 12 subjects used their assistive device during the 6MW, 10 during the CGS and FGS, and 4 subjects during the TUG. A digital stopwatch with an accuracy of one decimal figure in units of 1 second was used to measure time. Subjects were offered refreshments (water or apple juice) during the test sessions.

Table I. *Clinical characteristics of the subjects*

	Men (n = 38)		Women (n = 12)	
	n	%	*n*	%
Type of stroke				
Ischaemic	28	74	9	75
Haemorrhagic	10	26	3	25
Hemiparetic side				
Weakness in right side	18	47	2	16
Weakness in left side	20	53	10	84
Use of assistive device				
No walking aid	25	66	7	58
Walking aid	8	21	3	25
Ankle-foot orthosis and walking aid	5	13	2	17
Self reported walking ability				
<1000 metres	6	16	0	0
1000–3000 metres	17	45	9	75
>3000 metres	15	39	3	25

The TUG (23) is a modified version of the "Get-Up and Go" test (28). The TUG (23) was developed primarily to evaluate basic functional mobility in frail elderly persons. For the TUG, the subjects sat in a chair (seat height 44 cm, depth 45 cm, width 49 cm, armrest height 64 cm) placed at the end of a marked 3-metre walkway. Subjects were instructed to sit with their back against the chair, and on the word "go", stand up, walk at a comfortable speed ("like fetching something in your kitchen") past the 3-metre mark, turn around, walk back and sit down in the chair. Each subject did 1 trial to become familiar with the test. After a 1-minute rest, the TUG was performed twice separated by a 1-minute rest. The time from the start until the subject sat down in the chair with back support was measured and the mean of the 2 tests was recorded.

Gait speed timed over short distances (mostly 5–10 metres) has been used frequently as a determinant of mobility in both healthy elderly individuals (29, 30) and stroke patients (8, 19, 21, 22, 31). For the CGS and FGS, subjects were tested in a corridor and the walkway was marked on the linoleum floor with tape in different colours approximately 15 cm from one wall. The total marked distance was 14 metres and the subjects were timed over the middle 10 metres. Standing behind the first mark, the subjects were instructed to walk to the last mark and were informed that they would be timed for part of the walkway. For the CGS, the subjects were told to walk at a self-selected comfortable pace ("like walking in the park"). For the FGS, the subjects were told to walk as fast and safely as possible without running ("like hurrying to reach the bus"). Subjects started with the CGS 3 times in succession and with 30 seconds between each trial. After a further 30 seconds rest they continued with the FGS, also 3 times in succession, with 30 seconds between each trial. The time (in seconds) taken to walk 10 metres was recorded for each trial. The mean times for the 3 trials of CGS and FGS were then determined and used to calculate the 2 velocities (metres/second).

Stair climbing is a part of many measurement tools, e.g. FIM, and is used to evaluate mobility. For the SCas and SCde, subjects were tested in an isolated part of the hospital. The flight of stairs had 12 steps with rails on both sides. The steps were 135 cm wide, 15 cm high and 30 cm deep with a black rubber strip around the edge. Subjects were instructed to walk as fast and safely as possible, and preferably in a step-through pattern. Before the start, each subject decided whether or not to use the handrail: 36 subjects used the handrail during the SCas and 39 subjects during the SCde. The subjects climbed up 1 flight of stairs first (SCas) and stopped. After a 30-second rest, the subject climbed down again (SCde) on the same command. After another 30-second rest, subjects completed a second trial up and down the flight of stairs. The time (in seconds) from when the first foot left the ground until the second foot touched the ground on the last step was measured for the SCas and SCde separately. The means of the 2 trials for SCas and SCde were recorded.

The 6MW is commonly used to assess patients with cardiovascular or cardiorespiratory problems (32, 33) and is regarded as a submaximal test of aerobic capacity. It is adapted from the "12-minutes-walk-test" which, in turn, was adapted from the "12-minutes run-test" (34). For the 6MW, subjects were tested in a 2.2-metre wide corridor with a linoleum floor in a quiet part of the hospital. The subjects were instructed to walk 30 metres between 2 marks on the floor. After passing either mark, they were told to turn and walk back. Subjects were also instructed to cover as much ground as possible ("to walk as far as possible during 6 minutes"). They were allowed to rest and then to continue walking; only 1 subject had to rest during the test. The subjects were informed when 3 minutes of the test remained. As it has been shown that verbal commands can influence the distance walked (35), no verbal encouragement was given during the test. The 6MW was done once and the number of 30-metre lengths was counted. One wall was also marked every metre so that the distance walked could be measured to the nearest metre.

Procedure

The subjects were tested on 2 occasions, at the same time of the day and 7 days apart; for 2 subjects, the interval was 10 and 13 days, and for

3 subjects the test sessions were not at the same time of day. All subjects were provided transport free of charge to and from the test site. The same physiotherapist (U-BF) supervised all tests. Each test session lasted approximately 1.5 hours.

At the first test session the individual was again informed about the purpose and disposition of the study. Following the pre-test assessments, the tests were performed in the following order: the TUG, the CGS and the FGS, the SCas and SCde and finally the 6MW. Subjects rested on a chair for 5 minutes, before the first walking test (TUG), the SCas and the 6MW, respectively.

After each completed test session, subjects could ask questions and could be helped with stretching. After the first test session, the subjects received information about the second test session but were not informed about their results. A written summary and oral information about the test results were given after completion of the second test session.

Data and statistical analysis

The 6 recorded variables from the gait performance tests, obtained from the 2 test sessions, were used in the analysis. The difference between men and women for each of the 6 variables was assessed with the two-sided *t*-test. The relationship between the 6 variables in each of the 2 test sessions was addressed using Pearson's correlation coefficient. A significance level greater than 0.05 represented non-significance. All calculations were performed using the SPSS 11.0 Software for Windows (SPSS Inc., Chicago, Ill., USA).

Agreement between measurements was analysed by the intraclass correlation coefficients, $ICC_{1,1}$ and $ICC_{2,1}$ (36). As they gave essentially the same results, we only used the $ICC_{2,1}$ since that also provided the basis for the calculations of the standard error of the measurement (SEM). If BMS represents the variability between subjects, WMS the variability in the measurements within subjects, JMS the variability between test sessions, EMS the residual mean square and n the number of subjects, then for 2 test sessions

$$ICC_{2,1} = (BMS - EMS)/(BMS + EMS + 2(JMS - EMS)/n) \quad (1)$$

For $ICC_{2,1}$ a two-way ANOVA was used. The 95% confidence interval (95% CI) for $ICC_{2,1}$ was obtained from the ANOVA tables.

Systematic changes in the mean were assessed with the "Bland & Altman analyses" (11). The "Bland & Altman analyses" included the following calculations:

$$\bar{d} = \text{the mean difference between the 2 test sessions (test 2 minus test 1)} \quad (2)$$

$$SD_{diff} = \text{the standard deviation of the differences between the 2 test sessions} \quad (3)$$

$$\text{standard error (SE) of } \bar{d} = SD_{diff}/\sqrt{n} \quad (4)$$

$$95\% \text{ confidence intervals of } \bar{d} \text{ (95\% CI)} = \bar{d} \pm 2.01 \times SE \quad (5)$$

Here, SD_{diff} was used to calculate SE, which in turn was used to calculate the 95% CI for the mean of the differences. The value 2.01 in equation (5) was obtained from the t-table with 49 (n–1) degrees of freedom (df). If zero is included within the 95% CI, it is inferred that there is no significant systematic bias in the data. The "Bland & Altman analyses" also included the formation of graphs (so called "Bland & Altman graphs"), with the difference between test session 2 and test session 1 (2 minus 1) plotted against the mean of the 2 test sessions for each subject. These graphs can be used to visualize systematic variations around the zero line, to illustrate heteroscedasticity (10) – which occurs when the difference between test-retest measurements generally increase as the mean value of the measurements increase – and to identify outliers. For each of the 6 variables the possibility of heteroscedasticity was addressed by forming the Pearson's correlation coefficient of the absolute differences between test sessions 2 and 1 and the mean of the 2 test sessions for each subject (37). An outlier was considered to be present when the difference between the 2 test sessions was outside 2 standard deviations (SD).

Measurement errors were evaluated by the standard error of measurement, SEM, and the SEM%. The SEM was calculated using the square root of the within-subjects error variance:

$$SEM = \sqrt{WMS} \quad (6)$$

The SEM%, the within subject standard deviation as a percentage of the mean, was defined by:

$$SEM\% = (SEM/mean) \times 100 \quad (7)$$

where mean is the mean for all the observations from test sessions 1 and 2. The SEM% is independent of the units of measurement (SEM% is very similar to the coefficient of variation, CV%, which is defined by the standard deviation divided by the mean multiplied by 100). The SEM% represents the limit for the smallest change that indicates a real (clinical) improvement for a group of individuals following, for example, an intervention. In other words, a measurement following an intervention should be outside the range of measurement error to indicate a real improvement for a group.

To define the smallest change that indicates a real (clinical) improvement or a deterioration for a single individual, we used the smallest real difference, SRD, introduced by Beckerman et al. (38). The SRD is algebraically similar to the Limits of Agreement, LOA, described by Bland and Altman (11). The LOA has been used in previous reliability studies (13, 15) and gives materially the same results. The SRD was defined by:

$$SRD = 1.96 \times SEM \times \sqrt{2} \quad (8)$$

The value 1.96 was used when a value from the t-distribution would have been preferred, in this case 2.01. Provided the sample size is sufficiently large, $n > 30$ say, it makes no practical difference whether the value from the t-table or from the normal table is used. Beckerman et al. (38) suggested the calculation of an "error band" around the mean difference of the 2 measurements, $\bar{d}$; the 95% SRD was defined by:

$$95\% \text{ SRD} = \bar{d} \pm SRD \quad (9)$$

To allow the SRD to be independent of the units of measurement, and thereby used to determine a relative difference after an intervention or to detect a relative deterioration over time, the SRD can be expressed as a percentage value, the SRD%, which was defined by

$$SRD\% = (SRD/mean) \times 100 \quad (10)$$

where mean is the mean for all observations from test sessions 1 and 2.

RESULTS

Pre-test assessments

The mean rehabilitation admission FIM motor score varied considerably (mean 63, SD 17.9, range 22–88). The improvement in FIM motor score from admission to discharge from rehabilitation was on average 17 "steps", and improved a further 5 "steps" until the start of the study. Most of the subjects had a low or no increased muscle tone; 18 subjects were scored 0 on both occasions on the MAS and only 7 subjects more than 3 points on each occasion. There were only small differences between the 2 test-sessions; the mean MAS for test-session 1 was 1.56 (SD 2.1, range 0–8) and 1.64 (SD 2.0, range 0–7) for test-session 2.

Gait performance tests

There were no significant differences between the sexes for any of the gait performance tests; throughout the analyses and

Table II. *Summary of the 6 gait performance tests*

	Test session 1			Test session 2		
Test	Mean	SD	Range	Mean	SD	Range
Timed Up & Go (seconds)	14.3	5.2	7.5–25.7	13.7	5.3	6.7–27.7
Gait Speed (metres/second)						
Comfortable	0.89	0.3	0.4–1.4	0.94	0.3	0.4–1.5
Fast	1.3	0.5	0.5–2.2	1.4	0.4	0.5–2.1
Stair Climbing (seconds)						
Ascend	10.6	4.9	5.0–27.5	10.3	4.7	5.5–25.6
Descend	11.0	6.3	4.7–30.8	10.6	5.8	4.4–27.5
6-Minute Walk (metres)	384	132	122–606	398	136	122–648

Table III. *Relationship between the 6 gait performance tests within each test session*

	Test session 2					
Test session 1	TUG	CGS	FGS	SCas	SCde	6MW
TUG	–	−0.86	−0.91	0.86	0.90	−0.92
CGS	−0.84	–	0.92	−0.81	−0.82	0.89
FGS	−0.91	0.88	–	−0.84	−0.87	0.95
SCas	0.88	−0.80	−0.85	–	0.93	−0.83
SCde	0.90	−0.77	−0.83	0.91	–	−0.86
6MW	−0.89	0.84	0.94	−0.82	−0.80	–

All correlation coefficients were significant ($p < 0.001$).
TUG = Timed "Up & Go"; CGS = Comfortable Gait Speed; FGS = Fast Gait Speed; SCas = Stair Climbing ascend; SCde = Stair Climbing descend; 6MW = 6-Minute Walk.

presentations, the results from men and women are therefore combined. The means, standard deviations and the ranges of values for the gait performance tests from the 2 test sessions are presented in Table II. The differences between the means of the 2 tests were smaller than 6% for all of the 6 gait performance tests.

There was a highly significant ($p < 0.001$) correlation between all the gait performance tests within each test session; the absolute values of the correlation coefficients were in the range 0.77–0.95, median 0.86 (Table III). Ten of the 30 absolute correlation coefficients were 0.90 and above, 9 were in the range 0.85–0.89 and the remaining 11 were in the range 0.77–0.84.

Reliability analysis

Using the criteria of Fleiss (39), all tests showed excellent agreement; the values of $ICC_{2,1}$ ranged from 0.94 to 0.99 (Table IV). The 95% confidence intervals for $ICC_{2,1}$ were narrow and ranging from 0.90 to 0.99.

All $\bar{d}$ values were close to zero and the widths of the 95% CI for $\bar{d}$ were narrow (Table IV). The value of $\bar{d}$ for all 6 tests indicated that the performance at the second test session was generally better than at the first. In 4 of the 6 tests (TUG, CGS, SCde and 6MW), zero was not included in the 95% CI of $\bar{d}$ implying a significantly ($p < 0.05$) better performance in the second test sessions.

From the "Bland & Altman graphs" (Fig. 1), the systematic variation around the zero line for TUG, CGS, SCde and 6MW was revealed. From the graphs, there were indications of a larger variability for higher test values, i.e. heteroscedasticity. The Pearson's correlation coefficient of the absolute differences between test sessions 2 and 1 and the mean of the 2 test sessions for each subject ranged from 0.26 to 0.67, and was significant ($p < 0.05$) for 5 of the 6 tests: CGS, FGS, SCas, SCde and 6MW. A few (2–4) outliers were identified in each of the 6 graphs. When the correlation was recalculated after the exclusion of these outliers, there was no significant relationship between the absolute difference and the mean of the 2 test sessions for any of the tests.

The SEM gives the measurement errors in absolute values (Table IV). The SEM% is independent of the units of measurement and therefore more easily interpreted. The values of

Table IV. *Reliability of the 6 gait performance tests*

Test	$ICC_{2,1}$	95% CI for ICC	$\bar{d}$	95% CI for $\bar{d}$	SEM	SEM%	95% SRD	SRD%
Timed Up & Go (seconds)	0.96	0.93–0.98	−0.58	−1.01 to −0.15	1.14	8.2	−3.75–2.59	23
Gait Speed (metres/seconds)								
Comfortable	0.94	0.90–0.97	0.05	0.02–0.08	0.07	7.9	−0.15–0.25	22
Fast	0.97	0.95–0.98	0.01	−0.02–0.04	0.08	5.7	−0.21–0.22	16
Stair Climbing (seconds)								
Ascend	0.98	0.97–0.99	−0.23	−0.50–0.03	0.67	6.5	−2.10–1.64	18
Descend	0.98	0.96–0.99	−0.41	−0.76 to −0.06	0.90	8.4	−2.92–2.10	23
6-Minute Walk (metres)	0.99	0.98–0.99	14	8–21	18.6	4.8	−37.3–66.0	13

$ICC_{2,1}$ = intraclass correlation coefficient; CI = confidence interval; SEM = standard error of measurement; SRD = smallest real difference.

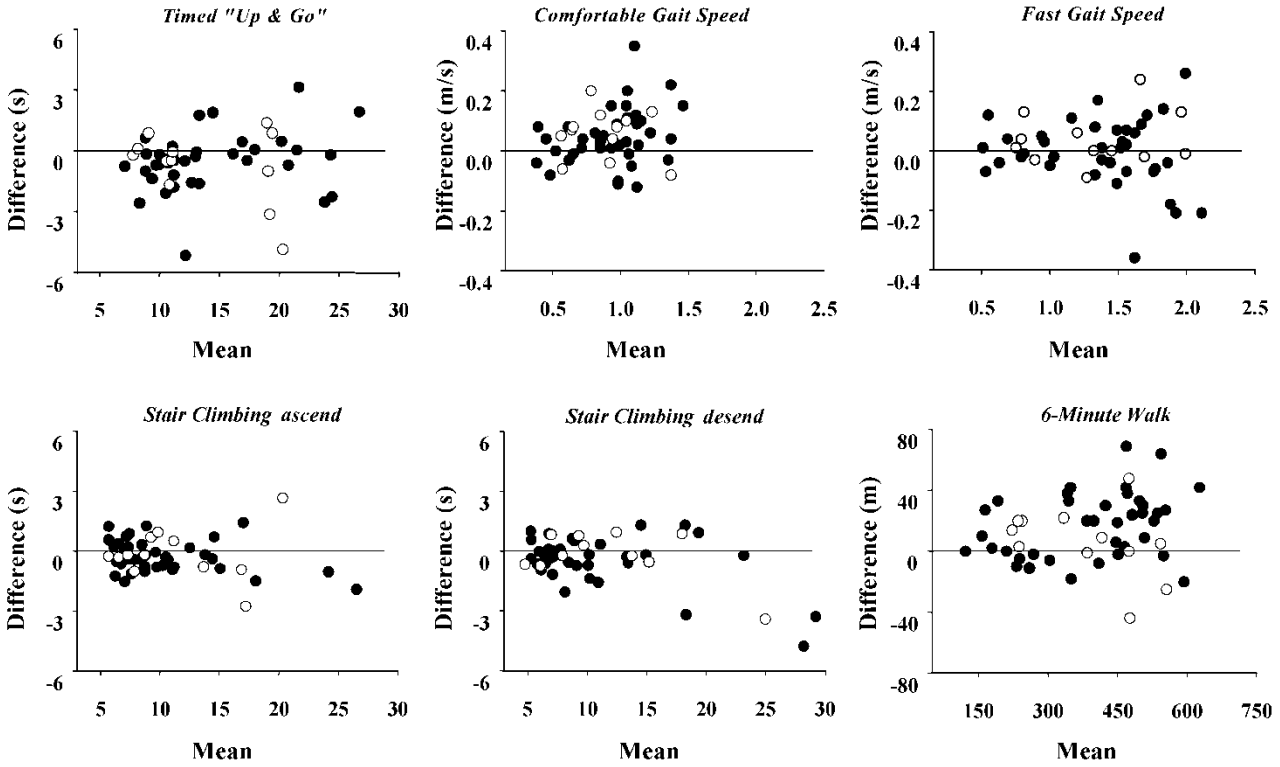

Fig. 1. The differences between test sessions 2 and 1 (test 2 minus test 1) plotted against the means of the 2 test sessions for the 6 gait performance tests for men (filled circles) and women (open circles). From these "Bland & Altman graphs", the systematic variation around the zero line for TUG, CGS, SCde and 6MW was revealed together with the evidence of a larger variability for higher test values, i.e. heteroscedasticity.

SEM% were all low ranging from 4.8% for the 6MW to 8.4% for the SCde (Table IV).

The 95% SRD (Table IV) indicates the range of measurement errors: a value outside this range indicates real (clinical) change. The SRD%, which represents the difference in relative terms, ranged from 13% for the 6MW to 23% for the TUG and the SCde (Table IV). For CGS, FGS and 6MW, the relative difference means that an improvement is represented by increased values, whereas the relative difference for TUG, SCas and SCde means that an improvement is represented by decreased values.

DISCUSSION

In this study the reliability of 6 commonly used gait performance tests were evaluated. We found that these gait performance tests were highly reliable. The analysis indicated that reasonably small improvements are sufficient to detect real changes for a group of stroke patients or individual stroke patients.

Walking has been identified as one of the most important components of Activities and Participation in the International Classification of Functioning, Disability and Health, ICF, Core Set for Stroke (3). In stroke rehabilitation, a major aim is therefore to optimise recovery of locomotor skills and gait performance (1), in order to enable participation in everyday activities (4). To evaluate gait performance after stroke and changes following an intervention, we need reliable measurement tools. The assessment of reliability is a broad concept that encompasses several parts. Over the last decade the analyses have developed from simply using correlation coefficients to more comprehensive sets of statistical methods. Even though there is no consensus as to which statistical methods to use, it is recommended that the assessments should include the analysis of agreement between measurements, systematic changes in the mean and measurement errors (10–15) Several different statistical indices cover these parts, and those applied here are the most commonly used. Recently, the notion of reliability has been expanded (40) and data from the analyses have been used to define limits for the smallest change that indicate a real (clinical) improvement both for a group of stroke patients and for individual stroke patients. Thus, by applying a comprehensive set of statistical methods the reliability of measurement tools can be fully evaluated.

A variety of tests have been described for the assessment of gait performance after stroke (1, 2, 16, 17, 19, 24). However, many tests are fairly extensive, time-consuming or require sophisticated laboratory equipment. The 6 gait performance tests evaluated in this study were selected as they are easy to administer and are meaningful to the patients. These tests also cover various aspects of gait performance, such as velocity, endurance and the complexity of gait, to provide a comprehensive picture of walking capacity after stroke (1).

Three main factors are likely to influence the reliability of the gait performance tests in this study: the subjects tested, the sample size and the test protocol. Even though the 50 men and women had all recovered well from their stroke, they were still restricted by their hemiparesis. For example, for two-thirds of the subjects the 6MW varied from 20% to 80% of the expected value for healthy age-matched people (41). The reliability of these gait performance tests is therefore primarily representative of fairly active post-stroke individuals. Further studies are needed to establish the reliability of these gait performance tests in individuals across a wider spectrum of post-stroke disability and ages. Previous reliability studies of chronic stroke patients, i.e. more than 6 months post-stroke, have been fairly small and very few studies have included more than 25 patients. It has been recommended that the sample size of test-retest reliability studies should be at least 30, and preferably 50 (14, 42). As a general principle, the larger the sample size the more dependable are the estimates of the change in measurement errors and the more compelling is the argument for extrapolating the measurement tool to a given population. Several sources of errors in the test protocol have to be recognised and their effects reduced to optimize reliability. Great care was taken here to standardize the tests, and a test protocol was followed carefully: for example, the same time interval between the tests, the same commands and the same environment. Thus, with all conditions as stable as possible, any variability between the 2 test sessions in this study is taken to represent the variability in the measurement parameters.

In accordance with previous studies (22, 24, 43), we found a high correlation between the gait performance. This is not surprising for some tests because they measure the same or very similar aspects of gait performance. However, the high correlations also mean that tests that measure different aspects of gait performance are related: for example, both CGS and FGS were related to 6MW, which indicates that gait velocity is closely related to gait endurance in these subjects.

The intraclass correlation coefficient (ICC) has become the most commonly used method to evaluate reliability. Even though no clear definition of acceptable ICC "cut-off" values for practical use has been presented, it has been suggested that the ICC should exceed 0.75 to indicate excellent reliability (39). The ICC values for the 6 gait performance tests in this study were well above this "cut-off" value. Surprisingly few studies have actually used the ICC to evaluate reliability of gait performance tests in chronic stroke patients. Eng et al. (43) and Green et al. (20) assessed the test-retest reliability (7 days apart) of CGS in 25 and 22 chronic stroke patients and reported ICC values of 0.95 and above. Baer et al. (44) evaluated the test-retest reliability (2 days apart) for a set of gait performance tests in 26 chronic stroke patients; they reported ICC values of 0.98 or above for the various subscales but no individual ICC values for each gait performance test. Some studies have reported similar ICC values for other gait performance tests but have not described their test design (the number of days apart, the sample size, etc.). Others have not specified which ICC they used, whereas some have only used the Pearson correlation coefficient to evaluate reliability. Although the different forms of ICC and the Pearson r, often take similar values (12), we cannot make detailed comparisons between our data and data from these studies.

It is now being appreciated that using only the ICC for the evaluation of reliability can lead to fallacious conclusions. ICC assesses the agreement between repeated measurements and thereby only the variance between subjects. Comprehensive evaluations of reliability should include assessments of the variability in the measurements within subjects.

The "Bland & Altman analyses" showed that the performance was generally better during the second test session. For 4 of the 6 tests – TUG, CGS, SCde and 6MW – the difference was significant, suggesting a learning effect in these tests. However, the mean differences between the 2 sessions ($\bar{d}$) were close to zero and the confidence intervals were narrow (c.f. Table IV), indicating that this learning effect was small. In the future, possible learning effects should be taken into account and their implications accommodated.

The "Bland & Altman graphs" displayed, and the statistical analyses revealed, heteroscedasticity for 5 of the 6 tests: the higher the test value the larger was the variability between the test sessions. Subjects who walked with a higher velocity (CGS and FGS) had a larger variability from test session 1 to 2, and, similarly, subjects who covered most ground during the 6MW had a larger variability. On the contrary, subjects who needed longer time to complete the stair climbing test (SCas and SCde) had the larger variability between test sessions. A few outliers in each test explained this heteroscedasticity; when these outliers were excluded from the calculations, no heteroscedasticity was present. We should be aware that heteroscedasticity can occur due to the floor-and-ceiling effect of a measurement tool and that larger values by nature can give rise to larger absolute variability.

Several indices have been suggested for the evaluation of measurement errors; in the present study, we used the SEM and the SEM%. The SEM gives the measurement errors in absolute values, whereas the SEM% is independent of the units of measurement, and therefore more easily interpreted. The SEM% represents the limit for the smallest change that indicates a real (clinical) improvement for a group of individuals following, for example, an intervention. All SEM% values in this study were below 10%. This implies that these tests are sensitive and can be used to detect small, clinically relevant, changes in mild to moderately affected chronic stroke patients. Green et al. (20) evaluated the test-retest reliability of gait speed over 10 metres in chronic stroke patients in a similar way; 22 men and women were tested 1 year post-stroke and small measurement errors were also reported.

The data can also be used to determine whether a method is sufficiently sensitive to detect a real (clinical) change for a single individual. In this study we calculated the smallest real difference (SRD) (38), which has been introduced as a method linking reproducibility to responsiveness. The SRD% is inde-

pendent of the unit of measurement and, like the SEM%, more easily interpreted. For the 6 gait performance tests evaluated in this study, the size of the relative change (SRD%) should exceed 13% (6MW) up to 23% (TUG and SCde) to indicate a real (clinical) change. For the data in Table II, the relative improvement (SRD%) needed to detect such a change for any subject in our study can be calculated. This approach accommodates heteroscedasticity to some extent. For example, in the 6MW, the average subject covered 391 metres and has to walk a further 51 metres to indicate a real (clinical) improvement; the equivalent distances for the slowest and the fastest subjects are 16 metres and 84 metres, respectively. From a clinical standpoint, the SRD% values presented here (c.f. Table IV) seem most reasonable, and confirm that these gait performance tests are useful to detect real (clinical) changes in chronic stroke patients.

Beckerman et al. (38) stated that there is an essential difference between a "clinically relevant change" and the "SRD": "SRD is a clinimetric property of a measurement tool, whereas 'clinically relevant change' is an arbitrarily chosen amount of change indicating which change clinicians and researchers minimally judge as important". An interesting area for future research is to explore the clinimetric property of a measurement tool and how that corresponds to what we as clinicians judge as "clinically relevant". Such research will help us define the optimal outcome measure for gait performance in stroke rehabilitation.

CONCLUSION

All 6 gait performance tests in this study showed: (i) high agreement between the test-retest measurements; (ii) no substantial systematic changes in the mean and small measurement errors; and (iii) sufficient sensitivity to enable the detection of real (clinical) changes in measurement score. Taken together, these tests are all highly reliable and can be recommended in clinical practice as well as research to evaluate various aspects of gait performance and changes over time in chronic stroke patients. Based on the analyses, the FGS and 6 MW are considered having the best reliability. As all the tests are highly related, the choice of test depends on what aspect of gait performance that is evaluated.

ACKNOWLEDGEMENTS

This study was supported by grants from the Swedish Stroke Association, Magn. Bergvall Foundation, the Swedish Association of Neurologically Disabled (NHR), the Swedish Society of Medicine, Gun and Bertil Stohne Foundation, the Crafoord Foundation, the Council for Medical Health Care Research in South Sweden, Lund University Hospital and Region Skåne.

REFERENCES

1. Richards CL, Malouin F, Dean C. Gait in stroke: assessment and rehabilitation. Clin Geriatr Med 1999; 15: 833–855.
2. Jorgensen HS, Nakayama H, Raaschou HO, Olsen TS. Recovery of walking function in stroke patients: the Copenhagen Stroke Study. Arch Phys Med Rehabil 1995; 76: 27–32.
3. Geyh S, Cieza A, Dickson H, Frommelt P, Zaliha O, Ring H, et al. ICF Core Set for stroke. J Rehabil Med 2004; Suppl. 44: 135–141.
4. Parker CJ, Gladman JR, Drummond AE. The role of leisure in stroke rehabilitation. Disabil Rehabil 1997; 19: 1–5.
5. Wade DT, Wood VA, Heller A, Maggs J, Langton Hewer R. Walking after stroke. Measurement and recovery over the first 3 months. Scand J Rehabil Med 1987; 19: 25–30.
6. Shepherd RB. Exercise and training to optimize functional motor performance in stroke: driving neural reorganization? Neural Plast 2001; 8: 121–129.
7. Mauritz KH. Gait training in hemiplegia. Eur J Neurol 2002; 9 (suppl 1): 23–29; dicussion 53–61.
8. Holden MK, Gill KM, Magliozzi MR, Nathan J, Piehl-Baker L. Clinical gait assessment in the neurologically impaired. Reliability and meaningfulness. Phys Ther 1984; 64: 35–40.
9. Rothstein JM. Measurement and clinical practice: Theory and application In: Rothstein JM, ed. Measurment in physical therapy. New York: Churchill Livingstone; 1985, p. 1–46.
10. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998; 26: 217–238.
11. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8: 135–160.
12. Holmback AM, Porter MM, Downham D, Lexell J. Reliability of isokinetic ankle dorsiflexor strength measurements in healthy young men and women. Scand J Rehabil Med 1999; 31: 229–239.
13. Holmback AM, Porter MM, Downham D, Lexell J. Ankle dorsiflexor muscle performance in healthy young men and women: reliability of eccentric peak torque and work measurements. J Rehabil Med 2001; 33: 90–96.
14. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000; 30: 1–15.
15. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998; 12: 187–199.
16. Salbach NM, Mayo NE, Higgins J, Ahmed S, Finch LE, Richards CL. Responsiveness and predictability of gait speed and other disability measures in acute stroke. Arch Phys Med Rehabil 2001; 82: 1204–1212.
17. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press; 1992.
18. Collen FM, Wade DT, Bradshaw CM. Mobility after stroke: reliability of measures of impairment and disability. Int Disabil Stud 1990; 12: 6–9.
19. Evans MD, Goldie PA, Hill KD. Systematic and random error in repeated measurements of temporal and distance parameters of gait after stroke. Arch Phys Med Rehabil 1997; 78: 725–729.
20. Green J, Forster A, Young J. Reliability of gait speed measured by a timed walking test in patients one year after stroke. Clin Rehabil 2002; 16: 306–314.
21. Hill KD, Goldie PA, Baker PA, Greenwood KM. Retest reliability of the temporal and distance characteristics of hemiplegic gait using a footswitch system. Arch Phys Med Rehabil 1994; 75: 577–583.
22. Maeda A, Yuasa T, Nakamura K, Higuchi S, Motohashi Y. Physical performance tests after stroke: reliability and validity. Am J Phys Med Rehabil 2000; 79: 519–525.
23. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991; 39: 142–148.
24. Rossier P, Wade DT. Validity and reliability comparison of 4 mobility measures in patients presenting with neurologic impairment. Arch Phys Med Rehabil 2001; 82: 9–13.
25. Guide for the uniformed data set for medical rehabilitation FIM^{SM}. (Swedish Version 5.0). In: Buffalo NY 14214: State University of New York at Buffalo; 1996.
26. Blackburn M, van Vliet P, Mockett SP. Reliability of measurements obtained with the modified Ashworth scale in the lower extremities of people with stroke. Phys Ther 2002; 82: 25–34.
27. Dean CM, Richards CL, Malouin F. Task-related circuit training improves performance of locomotor tasks in chronic stroke: a

randomized, controlled pilot trial. Arch Phys Med Rehabil 2000; 81: 409–417.
28. Mathias S, Nayak US, Isaacs B. Balance in elderly patients: the "get-up and go" test. Arch Phys Med Rehabil 1986; 67: 387–389.
29. Bohannon RW. Comfortable and maximum walking speed of adults aged 20–79 years: reference values and determinants. Age Ageing 1997; 26: 15–19.
30. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994; 49: M85–94.
31. Goldie PA, Matyas TA, Evans OM. Deficit and change in gait velocity during rehabilitation after stroke. Arch Phys Med Rehabil 1996; 77: 1074–1082.
32. Bittner V, Weiner DH, Yusuf S, Rogers WJ, McIntyre KM, Bangdiwala SI, et al. Prediction of mortality and morbidity with a 6-minute walk test in patients with left ventricular dysfunction. SOLVD Investigators. JAMA 1993; 270: 1702–1707.
33. Butland RJ, Pang J, Gross ER, Woodcock AA, Geddes DM. Two-, six-, and 12-minute walking tests in respiratory disease. Br Med J (Clin Res Ed) 1982; 284: 1607–1608.
34. Cooper KH. A means of assessing maximal oxygen intake. Correlation between field and treadmill testing. JAMA 1968; 203: 201–204.
35. Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL, et al. Effect of encouragement on walking test performance. Thorax 1984; 39: 818–822.
36. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–428.
37. Bland M. An introduction to medical statistics, 3rd edn. Oxford: Oxford University Press; 2000.
38. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10: 571–578.
39. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.
40. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12: 349–362.
41. Enright PL, Sherrill DL. Reference equations for the six-minute walk in healthy adults. Am J Respir Crit Care Med 1998; 58: 1384–1387.
42. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987; 6: 441–448.
43. Eng JJ, Chu KS, Dawson AS, Kim CM, Hepburn KE. Functional walk tests in individuals with stroke: relation to perceived exertion and myocardial exertion. Stroke 2002; 33: 756–761.
44. Baer HR, Wolf SL. Modified emory functional ambulation profile: an outcome measure for the rehabilitation of poststroke gait dysfunction. Stroke 2001; 32: 973–979.