Scand J Rehab Med 27: 27–36, 1995

THE BALANCE SCALE: RELIABILITY ASSESSMENT WITH ELDERLY RESIDENTS AND PATIENTS WITH AN ACUTE STROKE

K. Berg,[1] PhD, PT, S. Wood-Dauphinee,[2] PhD, PT and J. I. Williams,[3] PhD

From the [1] Center for Gerontology and Health Care Research, Brown University, Providence, RI, USA, [2] School of Physical and Occupational Therapy, McGill University, Montréal, Québec and [3] Institute for Clinical Evaluative Sciences, Sunnybrooke Health Sciences Center, Toronto, Ontario, Canada

ABSTRACT. The objective of this study was to assess the reliability of the Balance Scale. Subjects were chosen from a larger group of 113 elderly residents and 70 stroke patients participating in a psychometric study. Elderly residents were examined at baseline, and at 3, 6 and 9 months, and the stroke patients were evaluated at 2, 4, 6 and 12 weeks post onset. The Cronbach's alphas at each evaluation were greater than 0.83 and 0.97 for the elderly residents and stroke patients respectively, showing strong internal consistency. To assess inter-rater reliability, therapists treating 35 stroke patients were asked to administer the Balance Scale within 24 hours of the independent evaluator. Similarly, caregivers at the Residence were asked to test the elderly residents within one week of the independent evaluator. To assess intra-rater reliability, 18 residents and 6 stroke patients were assessed one week apart by the same rater. The agreement between raters was excellent (ICC = 0.98) as was the consistency within the same rater at two points in time (ICC = 0.97). The results support the use of the Balance Scale in these groups.

Key words: measurement, rehabilitation, stroke, geriatrics, balance.

The Balance Scale consists of 14 items that require subjects to maintain positions of varying difficulty and perform specific tasks such as rising from a chair and times stepping. Scoring is based on a subject's ability to perform the 14 items or movements independently and meet certain time or distance requirements. Each item is graded 0–4, giving a total of 56. The scale is portable, easy to administer, requires a ruler and stopwatch as equipment and takes only 10–15 min to complete.

The content of the Balance Scale was developed in three phases, each surveying a different panel of geriatric patients and health professionals (1). The content is consistent with the definition that balance is the ability to maintain an upright posture under a variety of conditions. Individuals must be able to hold positions of varying difficulty and to make appropriate postural adjustments for voluntary movements. The Balance Scale is targeted to frail elderly and patients undergoing rehabilitation, and is intended to serve multiple purposes including: quantitative description of ability, monitoring of patients' progress over time and evaluation of the effectiveness of interventions in clinical practice and research.

A preliminary assessment demonstrated high reliability based on ratings of videotaped performances of 14 patients (1). Consequently, following the preliminary study, a full validation study was undertaken that compared Balance Scale scores with clinicians' judgments, laboratory measures of sway, Tinetti Balance Sub-scale scores (22, 23), use of mobility aids, and the occurrence of falls in older adults. In addition, during the early recovery period post stroke, changes in Balance Scale scores were compared with changes in functional status and motor performance. Results of the validity testing in these groups have been described elsewhere (2, 3).

The purpose of the present study is, therefore, to report on the reliability assessment planned within the full validation study using both elderly individuals and patients with stroke. Reliability is basic to good measurement because without consistent scoring it is virtually impossible to demonstrate that scales or instruments validly measure the intended concept or property. In addition, because errors can occur with each testing, high reliability is required when repeated measurements are used to monitor the clinical status of patients or evaluate the effectiveness of treatments. It is therefore important to provide as much information as possible on the measurement properties of new instruments to help potential users generalize the findings to their own situations and needs.

This study examines the reliability of the Balance Scale when used in conditions that more closely resemble clinical reality, with raters independently scoring patients at different points of time. Testing under less controlled test conditions incorporates more sources of potential error, but the estimates derived from such assessment offer more valuable information to potential users on how to control and interpret results.

Specifically, the objectives were:

1. To assess the internal consistency of the Balance Scale when used with elderly residents and acute stroke patients.
2. To assess the inter-rater reliability between pairs of observers independently rating the same subject.
3. To assess the intra-rater reliability of the same observer rating the same two points in time.

MATERIAL AND METHODS

Subjects and methods for testing internal consistency

The internal consistency of the Balance Scale was examined within two inter-related longitudinal studies using older adults living in a seniors' residence and patients with a recent stroke. The methods and criteria for subject selection of the measurement study have been detailed elsewhere (3). Briefly, the eligibility criteria for the first sub-study were: aged 60 years and older, medically stable, independently mobile with or without a walking aid, and willing to participate in the study. Subjects were evaluated by independent evaluators at baseline, and at 3, 6 and 9 months, and were followed to one year.

In the second sub-study, patients with the diagnosis of acute stroke were recruited in two large acute care hospitals. Patients were eligible if they were: age 40 years and older, medically stable, showing evidence of motor impairment, living in the greater Montreal area and admitted to hospital with an acute stroke of less than 14 days duration. Balance Scale evaluations of the stroke patients were made at 2, 4, 6 and 12 weeks by the independent evaluators.

For descriptive purposes, research assistants assessed the residents on the Barthel Index (19) and the Mini-Mental State Exam (11). The 11 questions on the Mini Mental State Exam (MMSE) address orientation, memory, language, calculation, attention, and spatial ability. Total scores range from 0–30. A score below 18 indicates definite impairment whereas a score between 23 and 18 is suggestive of mild cognitive dysfunction (4, 25). The Barthel Index measures functional levels of independence by assessing 15 items related to self-care and mobility (13–16). Each item is scored by determining whether the patient can perform the requested activity independently, with assistance or supervision, or not at all. The scores for each item are summed and the total can range from zero (complete dependency) to 100 (independence in terms of personal care). The index can also be considered in terms of its two sub-scales. Self-Care (0–53) and Mobility (0–47).

Subjects and methods for testing inter and intra rater reliability

Elderly subjects and stroke patients were chosen for the inter and intra-rater reliability study from the larger pool of patients based on their willingness to undergo additional testing and the availability of a caregiver to administer the Scale within the appropriate time frame. The research assistance also tried to select subjects from both patient groups who represented the range of ability to balance.

To examine inter-rater reliability, the residents were evaluated by two raters within one week. One rating was performed by the independent evaluator at the time of a scheduled follow-up evaluation. Within a week of this evaluation, a second rating was done by a senior matron or a nurse who was familiar with the subject. In the stroke study, current care providers who were nurses or occupational or physical therapists were asked to administer the Balance Scale within 24 hours of the independent evaluator. The evaluations were carried out in both the general hospital and in the rehabilitation facilities.

To assess intra-rater reliability, the Balance Scale was administered twice by the same person, at least one week apart for both elderly residents and patients with stroke. No training was provided for the caregivers but they were able to read through the Scale and ask questions. No effort was made to standardize the location or other environmental factors of the paired tests.

Analysis

Descriptive statistics were used to examine the baseline sociodemographic and clinical characteristics of all the subjects. In addition, the characteristics of the sub-groups in the inter and intra-rater reliability study were compared with the larger groups of elderly residents and stroke patients.

Prior to assessing internal consistency, the magnitude and direction of the Pearson's correlation coefficients of each item with every other item in the Balance Scale were examined in a correlation matrix. The descriptive statistics for the items in the Balance Scale also included the frequency distribution of scores for each of the five response categories, the mean score for each item and the item-to-total correlations.

Internal consistency was tested by Cronbach's Alpha (6) at each evaluation time. The underlying assumption of this statistic is that each item is considered to be measuring the same common concept and thus the sum is likely to give a better estimate than any single item. The more the items covary relative to the sum of their variance, the higher is the Cronbach's Alpha. Because the items should be correlated with each other and the total score to capture the concept of interest, this form of reliability is called internal consistency. Cronbach's Alpha is regarded as high if greater than 0.80. An item-to-total correlation shows the degree of association between each individual item and the total score of the other items in the scale. An item-to-total correlation is considered adequate if it is above 0.4.

To assess the consistency of the findings each aspect of the analysis was repeated at each evaluation point for the two study populations, the elderly residents and the patients with a diagnosis of stroke.

To examine the reproducibility of the Balance Scale, paired ratings for each subject in the inter-rating reliability study were plotted to show the raw scores and extent of agreement in scores for the elderly residents and patients with stroke. A similar plot was made for the subjects in the intra-rater study.

Inter observer and intra observer agreement were quantified with the intraclass correlation coefficient (ICC) (9, 10) which has a range of 0 to 1 (perfect agreement). The ICC estimates the magnitude of true variation between subjects relative to the total variation in scores. The estimates for the variance and derived for the analysis of variance. As the inter-reliability assessment included different pairs of ratings for each subject, the variance was obtained from a one-way analysis of variance. However, when considering the intra-rater reliability, the variance estimates were derived from a two-way analysis of variance, using subjects and time as the factors. Time was included to examine whether the sequence of evaluation systematically influenced the scores. For example, within a given pair, did the second test score tend to be higher than the first?

The reliability estimates and confidence limits for the ICC were performed for all subjects and separately for each longitudinal study. Reliability coefficients of 0.80 and above are generally considered high, but, when making decisions about individuals, more stringent criteria are recommended (17, 20).

RESULTS

Characteristics of the subjects

Table I displays the sociodemographic and clinical characteristics of 113 elderly residents at entry to the study. The residents were predominantly female (83%), English speaking (90.3%) and well-educated with an average of 12.6 years of schooling. Residents had a mean of 3.9 (SD 1.4) conditions for which they took, on average, 3.9 (SD 2.0) medications. The most common conditions were cardiovascular diseases (55.8%), hypertension (52.2%) and rheumatic diseases (43.4%).

Overall, the older adults were quite independent in the basic activities of daily living with a mean Barthel Index score of 98.3 (SD 4.2). The mean Balance Scale score was 46·8 (SD 6.6) and the average Mini-Mental State Examination score was within the normal range for this age group, 27.9 (SD 2.7) out of a possible 30 points. Only 9 residents scored below 24, indicating possible cognitive impairment.

Table II presents the sociodemographic and medical characteristics of the 70 stroke patients at entry to the study. The mean age of the subjects was 71.6 years and the majority (95.7%) lived at home prior to the onset of the stroke. The proportion of males (51.4%) to females was approximately equal, as was the side of impairment. Patients had a mean of 2.6 (SD 1.4) comorbid medical conditions for which they took on average 3.8 (SD 1.9) medications.

As shown in Tables I and II, the 31 elderly residents and the 36 stroke patients who participated in the inter and intra rater reliability study displayed similar characteristics to those of the larger group.

Internal consistency

The inter-item correlation matrices for the baseline evaluation are displayed in Table III. Because all elderly residents could sit unsupported, the relationship of this item to each of the others and the total score could not be tested. As shown in Table III, the average inter-item correlation for the elderly residents

Table I. *Characteristics of the elderly residents at baseline*

	Elderly residents evaluated to assess internal consistency ($n=113$)		Elderly residents in inter- and intra-rater reliability study ($n=31$)	
	Mean	(SD)	Mean	(SD)
Age (years)	83.5	(5.3)	84.4	(5.0)
Education (years)	12.6	(3.9)	12.5	(3.3)
Mean # diagnoses	3.9	(1.4)	3.8	(1.2)
Mean # medications	3.9	(2.0)	4.0	(2.2)
Mental status (MMSE 0–30)	27.9	(2.7)	27.5	(2.9)
Barthel Scores (0–100)	98.3	(4.2)	99.9	(0.4)
Sex	Number	(%)	Number	(%)
Female	93	(82.0)	23	(74.2)
Mobility aids				
None	49	(43.3)	17	(54.8)
Cane outdoors	26	(23.0)	7	(22.6)
Cane	29	(25.7)	7	(22.6)
Walker	9	(8.0)	—	—

Table II. *Characteristics of the stroke patients at baseline*

	Stroke patients evaluated to assess internal consistency ($n=70$)		Stroke patients in inter- and intra-rater reliability study ($n=36$)	
	Mean	(SD)	Mean	(SD)
Age (years)	71.6	(10.1)	72.4	(9.1)
Education (years)	8.6	(3.5)	10.1	(3.3)
# Comorbid conditions	2.6	(1.4)	2.9	(1.6)
# Medications	3.8	(1.9)	3.9	(2.0)
Sex	Number	(%)	Number	(%)
Female	34	(48.6)	18	(50.0)
Side of weakness				
Left	32	(47.7)	21	(58.3)

was fairly low (0.34), but the corresponding Cronbach's alpha was high, indicating that the items are contributing additional information to the total score. The results for the stroke patients show higher inter-item correlations, with none below 0.40.

The analyses were repeated at each evaluation point to assess the consistency of the results. Cronbach's Alpha were above 0.83 at each evaluation of the elderly residents, suggesting that the scale is measuring one underlying concept. In the stroke group, Cronbach's Alpha values were even higher (0.97 to 0.98) than in elderly residents and consistent at each evaluation. All item-to-total correlations were above 0.62. The item-to-total correlations for both stroke patients and elderly residents at baseline are shown in Table IV.

Tandem standing and standing on one leg demonstrated item-to-total correlations below 0.4 on two of the four occasions. This means that in the sample of elderly residents, a subject's performance on these items is not strongly related to scores for the remaining items. However, because these same items worked well for stroke patients and were above 0.4 for the other two evaluation points, it was decided to retain them. In both groups these two items had the lowest mean score, indicating a greater degree of difficulty.

The internal consistency in both populations is considered high. The relatively stronger internal consistency and higher inter-item correlations in stroke patients may be partially explained by the greater intersubject variability in performance. The elderly residents has narrower range of scores. For example, 93.8% of the residents received four points for standing unsupported. This uniformity of scores across subjects makes it difficult to obtain a high coefficient for a correlation between two variables.

Inter-rater reliability

In total, 32 individual raters (one nurse, 21 physical therapists, six occupational therapists, two senior matrons and two physical therapy students) were used to rate 35 stroke patients and 28 elderly residents in the inter-rater reliability study. Each patient was evaluated twice, by random pairs of raters. The Balance Scale scores for all the subjects covered the entire range (0–56) of the scale and had a mean of 37.1 (SD 17.2), averaged over both ratings.

Fig. 1 shows the two ratings for each of the 28 elderly residents, arranged in order of the mean Balance Scale scores. The scores ranged from 25 to 55. The paired scores appear consistent with most ratings within a few points of each other. The worst case showed a difference of eight points.

Fig. 2 presents the paired ratings for the 35 stroke patients whose scores encompassed the entire range of the Balance Scale (0–56). As in Fig. 1, cases are arranged on the basis of the mean of the paired ratings. The scores demonstrate generally good agreement, but there are occasional differences in certain subjects.

The intraclass correlation coefficient (ICC) was used to quantify the agreement between the raters. This statistic estimates the true variance between subjects relative to the total observed variance in scores. Variance estimates are obtained from the one-way analysis of variance. Overall, when all subjects were included, the ICC was 0.98 (95% CI lower

Table III. *Inter-item correlations of the Balance Scale at the initial evaluation*

	1. Sit-stand	2. Stand	3. Sit	4. Stand-sit	5. Trans	6. Stand EC	7. Stand FT	8. Arm reach	9. Pick up	10. Twist turn	11. Turn 360°	12. Step stool	13. Tandem
Elderly subjects ($n=113$)													
2 Standing	0.36												
4 Stand to sit	0.69	0.37											
5 Transfer	0.68	0.27		0.64									
6 Stand eyes closed	0.56	0.54		0.54	0.47								
7 Stand feet together	0.45	0.19		0.44	0.36	0.33							
8 Arm reaching	0.41	0.44		0.41	0.28	0.22	0.17						
9 Object Pick up	0.36	0.44		0.28	0.34	0.23	0.15	0.28					
10 Twist turn	0.56	0.32		0.52	0.49	0.62	0.37	0.27	0.42				
11 Turn 360°	0.44	0.20		0.42	0.40	0.47	0.10	0.31	0.49	0.54			
12 Step on stool	0.36	0.15		0.39	0.41	0.28	0.30	0.35	0.38	0.37	0.45		
13 Tandem standing	0.15	0.10		0.22	0.20	0.09	0.25	0.21	0.28	0.18	0.26	0.35	
14 One leg standing	0.24	0.14		0.20	0.22	0.25	0.14	0.24	0.24	0.32	0.37	0.40	0.28
				average r		minimum		maximum		percent below 0.3			
Cronbach's alpha = .83				0.34		0.09		0.69		41%			
Stroke patients ($n=69$)													
2 Standing	0.88												
3 Sit	0.68	0.73											
4 Stand to sit	0.88	0.90	0.66										
5 Transfer	0.90	0.89	0.71	0.83									
6 Stand eyes closed	0.83	0.92	0.69	0.86	0.87								
7 Stand feet together	0.82	0.81	0.58	0.79	0.81	0.78							
8 Arm reaching	0.88	0.87	0.62	0.87	0.88	0.81	0.89						
9 Object pick up	0.76	0.82	0.55	0.83	0.79	0.77	0.80	0.79					
10 Twist turn	0.84	0.88	0.63	0.89	0.86	0.85	0.88	0.92	0.90				
11 Turn 360°	0.74	0.77	0.50	0.74	0.80	0.76	0.81	0.81	0.87	0.86			
12 Step on stool	0.63	0.65	0.40	0.65	0.67	0.60	0.72	0.68	0.77	0.73	0.78		
13 Tandem standing	0.65	0.70	0.46	0.70	0.74	0.71	0.80	0.76	0.77	0.78	0.77	0.73	
14 One leg standing	0.51	0.60	0.40	0.62	0.63	0.60	0.62	0.69	0.72	0.74	0.66	0.70	0.75
				average r		minimum		maximum		percent below 0.3			
Cronbach's alpha = .97				0.75		0.40		0.92		0%			

Table IV. *Corrected item-to-total correlations for elderly residents and stroke patients at baseline*

Scale item	Elderly residents ($n = 113$)	Stroke patients ($n = 69$)
Sit to stand	0.64	0.90
Standing	0.41	0.93
Sitting	—	0.67
Stand to sit	0.63	0.91
Transfer	0.60	0.92
Stand eyes closed	0.56	0.89
Stand feet together	0.40	0.89
Arm reaching	0.46	0.92
Object pick up	0.53	0.88
Twisting	0.62	0.95
Turn 360°	0.60	0.86
Stepping	0.59	0.74
Tandem standing	0.38	0.80
One leg standing	0.44	0.70
Cronbach's alpha	0.83	0.97

bound 0.97) indicating excellent agreement. When the analysis was restricted to elderly residents and then repeated for stroke patients, the respective ICCs were 0.92 (95% CI lower bound 0.85) and 0.98 (95% CI lower bound 0.96).

Intra-rater reliability

To assess intra-rater reliability, seven raters (five physical therapists, one occupational therapist, and one nurse), evaluated 24 stable subjects (18 elderly residents and six stroke patients) twice, one week apart. The range of Balance Scale scores was from 4-56 with an average of 46.0 (SD 11.0) over both ratings. The paired ratings showed high levels of agreement with 71% within two points of each other.

The ICC was used to assess the level of agreement using information from a two-way analysis of variance with subjects and time as factors. The ICC for all subjects was 0.97 (95% CI 0.93 −0.99); whereas elderly residents showed an ICC of 0.91 (0.80 −0.96) and stroke patients an ICC of 0.99 (CI 0.94 −0.99).

DISCUSSION

The high levels of reliability demonstrated in this study support the potential of the Balance Scale to serve multiple purposes in research and clinical practice. The stringent requirements are especially desirable when measures are used to evaluate the effectiveness of interventions and monitor the status of patients over time because clinically meaningful differences may be small and errors may occur at each testing. In research, any excess measuring errors will adversely influence the sample size, cost of the study and the power to detect a true treatment effect. In clinical practice, reliability coefficients as high as 0.94 are recommended when results of repeated tests will be used to make decisions about individuals (17, 20).

The high estimates of inter-rater (ICC 0.98) and intra-rater (ICC 0.97) reliability are more surprising because they include errors from a variety of potential sources that occur in real life. Raters may differ in how closely they follow the written instructions. Caregivers may rate higher than independent evaluators because they know the patient did the task well on a previous occasion. Patients may alter their performance

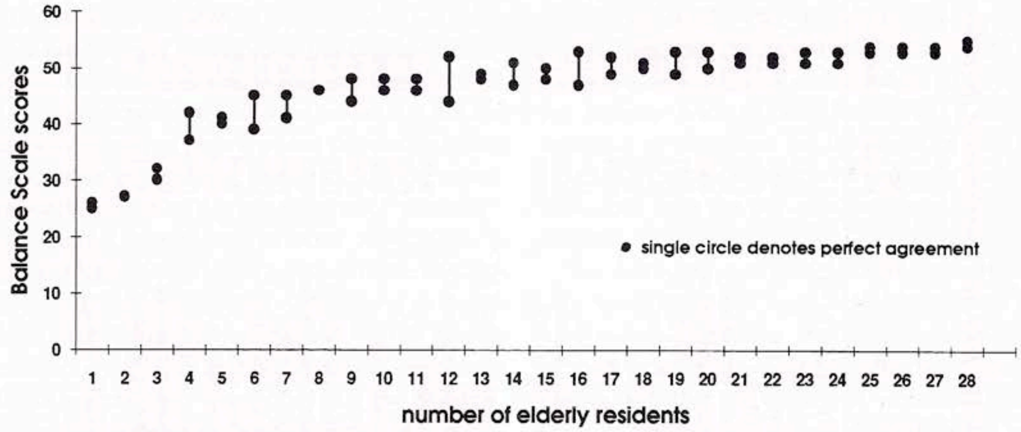

Fig. 1. Paired Balance Scale scores for elderly residents in the inter-rater reliability study ($n = 28$).

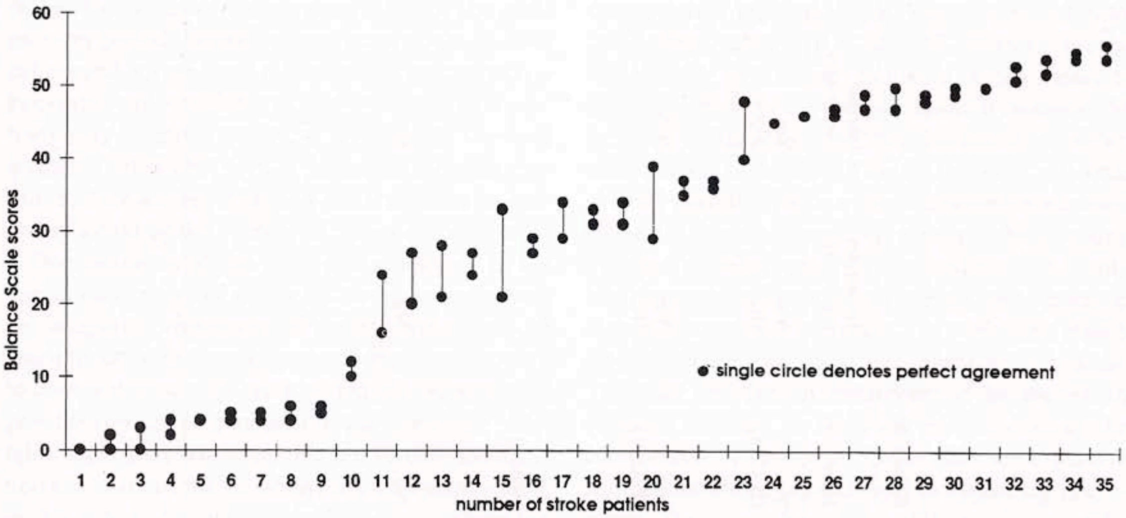

Fig. 2. Paired Balance Scale scores for stroke patients in inter-rater reliability study ($n = 35$).

because of fear, fatigue, cognitive impairment or lack of motivation. Subject may also feel better and more secure with someone they know as opposed to an independent evaluator. Additionally, noise, visual distractions, unsuitable furnishings and other environmental conditions may contribute to measurement error in a performance-based instrument.

The generalizability of the results is strengthened by the varied clinical characteristics of the subjects, the diversity and lack of training of the raters, and the lack of control of the test conditions. The majority of caregivers participating in the study were physical therapists, 66% and 71% of the raters in the inter and intra-rater reliability studies, respectively. Some were student therapists and others had been working more than 30 years. The other professionals were either nurses or occupational therapists. In addition, at the home for the elderly, two paraprofessionals called senior matrons participated in the study. The raters received no formal training in the administration of the Balance Scale but they were asked to read through all the items and ask questions as necessary.

In addition to the demonstrated inter and intra-rater agreement, the Cronbach's alpha estimates were high in both elderly residents (>.83) and stroke patients (>.97), indicating strong internal consistency. Internal consistency is not essential for a measuring instrument but it does facilitate the interpretation of the test results. The primary advantage of having multiple homogeneous items in the Balance Scale is that it provides a basis for a more consistent estimate of the ability of individuals to balance. The overlapping of the confidence intervals for the ICC and Cronbach's alpha for each specific group, is indicative of the shared assumptions of the two types of reliability coefficients (5).

Although still considered in the high range, estimates of reliability specific to the elderly residents were consistently lower than those for the patients with stroke. The explanation for this finding may relate to the impact of a narrower range of scores on the calculation of both Cronbach's alpha and the ICC. In the present study, the full range of scores (0–56) was represented in the group of stroke patients whereas the elderly residents showed less variability in their scores (24–56).

When designing a reliability study, it is advisable to include subjects representing the range of ability of the target population to improve the generalizability of the results. However, it is possible to have high overall reliability across a wide range of scores and still demonstrate some discrepancies that warrant clinical concern. It is therefore important to examine the descriptive information on mean scores and absolute differences in addition to summary stastistics. For example, when the paired ratings were ranked by their mean score and then graphically displayed on Fig. 2, scores or the stroke patients in the middle ranges of ability showed greater absolute differences. This finding may reflect an inherent inconsistency in

Table V. *Scoring methods and reliability assessments of clinical measures of balance*

	Scoring (range of scores)	Method of assessment	Subjects	Reliability		
				Inter-rater	Intra-rater test-retest	Internal consistency
Balance • Berg et al. 1989 • Berg et al. 1992 • Present study	14 items (0–56)	Response categories based on time, distance and level of supervision in performing task	• Geriatric subjects (in-patients to community dwelling) • Stroke patients	Yes (ICC.98)	Yes (ICC.97)	Yes (alpha. 85–.98)
Functional Reach • Duncan et al. 1990 • Duncan et al. 1992	Single item	Average of 3 tests of reaching while standing. Yardstick attached to wall at shoulder height.	• Community-dwelling elderly	Yes (ICC.98)	Yes (ICC.92)	N/A
Tinetti Balance Sub-Scale of Mobility Score • Tinetti 1986 • Tinetti et al. 1986 • Tinetti et al. 1988	13 items (0–24)	Responsive categories based on subjective observations	• Residents of intermediate care facilities • Community-dwelling elderly	Yes (sparse)	No	No
CTSIT • Shumway-Cook & Horak 1986 • Horak 1987	No total score	Observations of postural sway while standing 30 sec in six sensory conditions	• Not actually tested but recommended for rehabilitation patients	No	No	No
Balance Coding Gabell and Simons 1982	Alphanumeric profile (maximum 6 ABX)	6 hierarchical standing tasks and 3 dichotomously graded tasks	• Geriatric subjects (in-patients to community-dwelling)	No	No	No

the performances of some patients during the early recovery period. Stroke patients must adjust to their deficits and relearn basic motor skills, a process that is associated with variability in performance. This variability may be more noticeable for patients who can attempt all items but vary in the degree to which they can meet the scoring criteria. A better estimate of their ability would be the average of two or more tests.

One limitation of the study is that patients in the low to mid ranges of ability following a stroke were not adequately represented in the intra-rater study. It was difficult to find stroke patients who were unlikely to change during a one week period. Discrepancies of possible concern to clinicians in the inter-rater reliability study were few in number but further investigation can examine the variability in performance levels in this group of patients and, if necessary, assess the number of tests that must be averaged to give a reliable result (10).

Another limitation is that the design of the study did not permit an estimate of error variance for specific types of raters. Nor do we know how much training may be required for non-professionals because most of the raters were physical or occupational therapists. Nonetheless, raters of different professions and varying degrees of experience participates in the study, and any variation in their scoring was included in the error mean square.

Table V compares the scoring methods and reliability assessments of existing measures of balance. Specific estimates of reliability are unavailable for the Balance coding Scale (12) or the CTSIT (18, 21) The Tinetti Balance Sub-scale (22–24) has not reported the intra-rater reliability or the internal consistency. In addition, little information is provided on the circumstances of inter-rater reliability except to say that two observers agreed on 85% to 90% of the items (22, 23). The measure of functional reach (7, 8) has demonstrated excellent inter-rater (ICC 0.98) and test-retest (ICC 0.92) reliability in the clinical setting for subjects who stand independently. A single item test, however, is unlikely to be a sufficiently comprehensive measure of a concept such as balance.

Overall, the Balance Scale has undergone extensive testing of its reliability in realistic clinical conditions. Paired tests were made at different times of the day, often in two separate locations with different furnishings and noise levels or other distractions. Despite the possible sources of variation, the Balance Scale demonstrated high reliability. When combined with the findings related to validity (2, 3) there is now considerable information available on the measurement properties of the Balance Scale. It continues to perform well relative to other measures. A direct comparison with the Tinetti Balance sub-scale showed that the Balance Scale discriminated more efficiently among groups of elderly subjects using different mobility aids (2). In addition, the only other measure to demonstrate the ability to monitor change in status was the measure of Functional Reach (26), which is a component item of the Balance Scale. The final test for an instrument is its use as an outcome measure in a clinical trial evaluating the effectiveness of treatments. Such trials are expensive and cannot afford poor choices in measuring instruments. The current knowledge of the properties of the Balance Scale suggests it is ready for this test.

Conclusion

The Balance Scale has demonstrated high reliability when tested in a variety of clinical and home settings by raters who were given little specific training in the administration of the test. The results provide information for potential users on how well they can generalize the findings to their patients. Overall, the Balance Scale has undergone considerable testing and its measurement properties support the use of the instrument in clinical practice and research.

ACKNOWLEDGEMENT

This research was funded by the National Health Research and Development Programme, Health and Welfare Canada. The authors gratefully acknowledge the comments of Donna Lamping, PhD, J. A. Hanley, PhD and David Gayton, MD, PhD. Many thanks also to all the health care professionals, the patients and their families who gave of their time to participate in the study.

REFERENCES

1. Berg, K., Wood-Dauphinee, S., Williams, J. I. & Gayton, D.: Measuring balance in the elderly: preliminary development of an instrument. Physiotherapy Canada *41:* 304, 1989.
2. Berg, K., Maki, B., Williams, J. I., Holliday, P. & Wood-Dauphinee, S.: A comparison of clinical and laboratory measures of postural balance in an elderly population. Arch Phys Med Rehabil *73:* 1073, 1992.
3. Berg, K., Wood-Dauphinee, S., Williams, J. I. & Maki, B.: Measuring balance in the elderly: validation of an instrument. Can J Pub Health *83,* Suppl 2: S71, 1992.
4. Bleecker, M. L., Bolla-Wilson, K., Kawas, C. & Agnew,

J.: Age-specific norms for the Mini-Mental Status Exam. Neurology **38:** 1565, 1988.

5. Bravo, G. & Potvin, L.: Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward an integration of two traditions. J Clin Epidemiol *44:* 381, 1991.
6. Cronbach, L. J.: Coefficient alpha and the internal structure of test. Psychometrika *16:* 297, 1951.
7. Duncan, P.W., Weiner, D.K., Chandler, J. & Studenski, S.: Functional reach: A new clinical measure of balance. J Gerontol *45:* M192, 1990.
8. Duncan, P. W., Studenski, S., Chandler, J. & Prescott, B.: Functional Reach: Predictive validity in a sample of elderly male veterans. J Gerontol *47:* M93, 1992.
9. Ebel, R. L.: Estimation of the reliability of ratings. Psychometrika *16:* 407, 1951.
10. Fleiss, J. L.: The design and analysis of clinical experiments. Toronto, John Wiley and Sons, 1986, 1–32.
11. Folstein, M. F., Folstein, S. E. & McHugh, P. R.: Mini mental state: a practical method for grading the cognitive state of patients for the clinician. J Psychiat Res *12:* 189, 1975.
12. Gabell, A. & Simons, M.B: Balance coding. Physiotherapy 68: 286, 1982.
13. Granger, C. V. & Greer, D. S.: Functional status measurement and medical rehabilitation outcomes. Arch Phys Med Rehabil *57:* 103, 1976.
14. Granger, C. V., Sherwood, C. C. & Greer, D. S.: Functional status measures in comprehensive stroke care program. Arch Phys Med Rehabil *58:* 555, 1977.
15. Granger, C. V., Albrecht, G. L. & Hamilton, B. B.: Outcome of comprehensive medical rehabilitation: measurement by Pulse Profile and the Barthel Index. Arch Phys Med Rehabil *60:* 145, 1979.
16. Granger, C. V., Deurs, L. S., Peters, N. C. et al.: Stroke rehabiliation: Analysis of repeated Barthel Index measures. Arch Phys Med Rehabil *60:* 14, 1979.
17. Helmstadter, G. C.: Principles of Psychological Measurement. Appleton-Century-Crofts, New York, 1964.
18. Horak, F. B.: Clinical measurement of postural control in adults. Phys Ther *67:* 1881, 1987.
19. Lichenstein, M. J., Burger, C., Shields, S. L. & Shiavi, R. G.: Comparison of biomechanics platform measures of balance and videotaped measures of gait with a clinical mobility scale in elderly women. J Gerontol *45:* M49, 1990.
20. Mahoney, F. I. & Barthel, D. W.: Functional evaluation: The Barthel Index. Md State Med J *14:* 61, 1965.
21. Nunnally, J. C.: Psychometric theory. Second ed. McGraw-Hill, New York, 1978.
22. Shumway-Cook, A. & Horak, F B.: Assessing the influence of sensory interaction on balance. Suggestion from the field. Phys Ther *66:* 1548, 1986.
23. Tinetti, M. E.: Performance-oriented assessment of mobility problems in elderly patients. J Am Geriatr Soc *34:* 119, 1986.
24. Tinetti, M. E., Williams, T. F. & Mayewski, R.: A fall risk index for elderly patients based on number of chronic disabilities. Am J Med *80:* 429, 1986.
25. Tinetti, M. E.: Factors associated with serious injury during falls by ambulatory nursing home residents. J Am Geriatr Soc *35:* 644, 1987.
26. Tinetti, M. E., Speechley, M. & Ginter, S. F.: Risk factors for falls among elderly persons living in the community. N Engl J Med *319:* 1701, 1988.
27. Tombaugh, T. N. & McIntyre, N. J.: The Mini-Mental State Examination: A comprehensive review. J Am Geriatr Soc *40:* 922, 1992.
28. Weiner, D. K., Bongoneri, D., Studenski, S., Duncan, P. W. & Kochersberger, G.: Does functional reach improve with rehabilitation? Arch Phys Med Rehabil *74:*796–800, 1993.

Address for offprints:

Katherine Berg
Center for Gerontology and Health Care Research
Brown University
Box G B234
Providence, RI 02912
USA