J Rehabil Med 2002; 34: 20–24

INTER-RATER RELIABILITY AND VALIDITY OF THE STROKE REHABILITATION ASSESSMENT OF MOVEMENT (STREAM) INSTRUMENT

Chun-Hou Wang,[1] Ching-Lin Hsieh,[2] May-Hui Dai,[3] Chia-Hui Chen[3] and Yu-Fen Lai[3]

From the [1]School of Rehabilitation, Chung-Shan Medical and Dental College, Taichung, [2]School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei, and [3]Department of Physical Therapy, Chung-Shan Medical and Dental College Hospital, Taichung, Taiwan, ROC

The Stroke Rehabilitation Assessment of Movement (STREAM) instrument is used to measure motor and mobility problems in patients who have experienced a stroke. The purposes of the study were to examine the inter-rater reliability, concurrent and convergent validity of the STREAM instrument in stroke patients. Fifty-four stroke patients participated in the study. For the purpose of inter-rater reliability, the STREAM instrument was administered by two raters on each patient within a 2-day period. Validity was assessed by comparing the patients' scores on the STREAM instrument with those obtained from the other well-established measures. Weighted kappa statistics for inter-rater agreement on scores for individual items ranged from 0.55 to 0.94. The intraclass correlation coefficient for the total score was 0.96 indicating very high inter-rater reliability. The intraclass correlation coefficients were also very high in each of the subscales. The total STREAM score was moderately to highly associated with the score of the Barthel Index and Fugl-Meyer motor assessment scale, rho = 0.67, and 0.95, respectively. The STREAM subscale scores were closely associated with scores of the other well-validated measures. Our results demonstrate that consistent and valid information can be obtained from the STREAM instrument and support its use in the value of the STREAM evaluation of motor and mobility recovery in persons who have experienced a stroke.

Key words: cerebrovascular accident, reproducibility of results, movement.

J Rehabil Med 2002; 34: 20–24

Correspondence address: Dr Ching-Lin Hsieh, 7 Chun-Shan South Road, School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei 100, Taiwan, ROC. E-mail: mike26@ha.mc.ntu.edu.tw

Accepted May 16, 2001

INTRODUCTION

Stroke is a major cause of mortality and disability in many countries (1). Almost all persons who have experienced a stroke develop motor and mobility problems. A systematic assessment of persons who have experienced a stroke including motor and mobility evaluations is important in planning treatment and assessing recovery over time. Although a number of assessment tools are available to measure the recovery of movement following stroke, they have rarely been used in clinical practice because of lengthy administration time, complexity of scoring, and dependence on equipment (2). Assessment methods should be relevant, reliable, valid, sensitive to change in the clinical condition, easy to use and communicable (3).

The Stroke Rehabilitation Assessment of Movement (STREAM) instrument (2, 4) was designed to provide a comprehensive, objective, and quantitative evaluation of the motor functioning of individuals with stroke. It was also designed to be quick and simple to administer. The STREAM instrument consists of 30 items that are equally distributed among 3 subscales: upper-limb movements, lower-limb movements, and basic mobility items (2, 4). The psychometric characteristics of the STREAM instrument have rarely been examined. Daley et al. (2) found that the STREAM instrument showed excellent reliability. However, the modest size of the sample in their study ($n = 20$) may limit the generalization of the results. Furthermore, the validity of the STREAM instrument is not well documented. Further psychometric characteristics testing of the STREAM instrument is needed to determine its utility in both research and clinical settings.

The purposes of this study were to examine the inter-rater reliability of the STREAM instrument on individual item scores, subscale scores, and total score, and to determine the concurrent and convergent validity of the STREAM instrument.

SUBJECTS AND METHODS

Subjects

Consecutive stroke patients admitted to the Physical Medicine and Rehabilitation Department at Chung-Shan Medical and Dental College Hospital in Taichung, Taiwan, ROC, from August 1999 through March 2000, were recruited using the following criteria: (1) diagnosis of cerebral hemorrhage (code 431 from the International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9]) and cerebral infarction (code 434); (2) first onset of stroke; and (3) ability to follow verbal commands. The clinical diagnosis of stroke was confirmed by physicians using the neuroimaging examination (computed tomography or magnetic resonance imaging). Patients with any major comorbid conditions that might interfere with motor function or its assessment (e.g. severe rheumatoid arthritis) were excluded. All subjects gave informed consent prior to inclusion in the study.

A total of 54 patients met the selection criteria and agreed to participate in the study. Most of the patients excluded were those who could not follow commands (e.g. patients with global aphasia). Three subjects with bilateral paralyses (without marked bilateral motor impairment) were included in this study; their voluntary limb movements on the more involved side were assessed. The median length of time after stroke onset of the subjects was 74 days (ranging from 25 to

 ISSN 1650–1977

Table I. *Clinical characteristics of the stroke patients* ($n = 54$)

Characteristic	Frequencies except specified
Gender (male/female)	30/24
Age (mean year (SD))	60.3 (12.8)
Days after onset (median (range))	74 (25–361)
Diagnosis	
Cerebral hemorrhage	29
Cerebral infarction	25
Side of paresis	
Right	25
Left	26
Bilateral	3
STREAM[a] (mean (SD, skewness))	27 (16.7, 0.35)
FM (median (range))	27 (5–92)
RMI (median (range))	3 (0–13)
BI (median (range))	50 (0–90)

[a] The average score of the two raters on the Stroke Rehabilitation Assessment of Movement (STREAM) instrument.

FM: Fugl-Meyer motor assessment scale (upper extremity, lower extremity and balance); RMI: Rivermead Mobility Index; BI: Barthel Index.

361 days). Further information about the characteristics of the study sample is presented in Table I.

Procedures

The study protocol was divided into two parts. The first part was an inter-rater reliability study. The STREAM instrument was administered by two physical therapists with a random order on the same patient in the same physical environment within a 2-day time period. Most of the patients were evaluated on the same day or within 24 hours. The 2-day period was established to minimize the effect of a possible spontaneous recovery, a confounding variable that could affect the result. Both of the physical therapists voluntarily participated in this part of the study. They were blinded to the results of each other's assessment during the study period.

Prior to the beginning of the study, the raters familiarized themselves with the STREAM instrument and its clinical application. Both raters reviewed the original literature describing the test and received two hours of in-service training on the administration of the evaluation. To improve their efficiency, both raters were asked to use this instrument daily in their clinical practice for at least one week before participating in the study.

The second part of the protocol was a validity study. Assessment of validity requires the use of standard instruments with which the scale is to be compared (5). The standard instrument criteria for evaluating upper and lower extremity motor impairment, mobility and disability were administered during the same period as the reliability study by another physical therapist who was blind to the results of the STREAM instrument. Concurrent validity of the subscales of the STREAM instrument were assessed by comparing the results for the upper-limb movements, lower-limb movements, and basic mobility subscales of the STREAM instrument with those of the upper extremity subscale of the Fugl-Meyer motor assessment scale (FMUE) (6), the lower extremity subscale of the Fugl-Meyer motor assessment scale (FMLE), and the Rivermead mobility index (RMI) (7). Convergent validity of the STREAM instrument was assessed by comparing the total scores of the STREAM instrument with the total scores of the motor function and balance subscales of the Fugl-Meyer assessment scale (FM), and the Barthel index (BI) (8).

Instruments

The STREAM instrument consists of 30 items that are equally distributed among 3 subscales: upper-limb movements, lower-limb movements, and basic mobility items. The items and testing positions of the STREAM instrument are listed in Table II. Voluntary movements of the limbs are scored on a 3-point scale (0: unable to perform the test movement, 1: able to only partially perform the test movement, and 2: able to complete the test movement). Basic mobility items are scored on a 4-point scale similar to that used for scoring limb movements except that a category has been added to allow for independence with the help of a mobility aid. Thus, the maximum raw total STREAM score is 70, with each of the limb subscales scored out of 20 points and the mobility subscale scored out of 30 points.

The STREAM instrument was designed to be quick and simple to administer. The following equipment is required: a chair, a stool with 18-cm height, pillows, and stairs. For further details of the test's standardization and administration the reader should refer to the original articles of Daley et al. (2, 4).

The FM is probably the most widely known scale of motor recovery after stroke. It primarily assesses motor control ability. The scale includes six subgroups: upper and lower extremity motor function, range of motion, pain, sensation and balance (6). The possible scores of upper extremity subscale, lower extremity subscale and balance subscale range from 0–66, 0–34, and 0–14 points, respectively. The FM is reliable and valid (9, 10).

The RMI was developed to measure mobility in patients with head injury or stroke. The RMI is a Guttman scale, which comprises 14 questions and one direct observation, and covers a range of hierarchical activities from turning over in bed to running (7). The highest score is 15, indicates the highest mobility status. The RMI was found to be reliable, valid (7, 11) and sensitive to change over time (11). The BI is a weighted scale of 10 items of basic activities of daily living including feeding, bathing, grooming, dressing, bladder and bowel control, chair/bed transfer, ambulation, and stair climbing (8, 12). The maximal score is 100 indicating that the patient is fully independent in physical functioning. The lowest score is 0, representing a totally dependent bedridden state. Its reliability and validity has been well established (12–14).

Table II. *The items and testing positions of the Stroke Rehabilitation Assessment of Movement instrument*[a]

Testing position	Test movements (subscale)
Supine	Scapular protraction (U), elbow extension (U), bending hip and knee (L), rolling (M), bridging (M), supine to sitting (M)
Sitting	Scapular elevation (U), raising hand to touch top of head (U), hand to sacrum (U), raising arm to fullest elevation (U), supination and pronation (U), making a fist (U), fingers total extension (U), opposition (U), hip flexion (L), knee extension (L), knee flexion (L), dorsiflexion (L), plantarflexion (L), knee flexion and dorsiflexion (L), sitting to standing (M)
Standing	Standing for 20 counts (M), hip abduction (L), knee flexion (L), dorsiflexion (L), placing affected foot onto first step (M), 3 steps backward (M), 3 steps to affected side (M), 10 m walk (M), walking down 3 stairs (M)

[a] For further details of the test's standardization and administration the reader should refer to the original articles of Daley et al. (2, 4). U = upper extremity subscale; L = lower extremity subscale; M = basic mobility subscale.

	0	0.5	
	3	0.5	569
Moderate to good agreement (0.5–0.75)	6	0.6	223344
	1	0.6	7
	6	0.7	112334
	5	0.7	55889
Excellent agreement (>0.75)	3	0.8	012
	2	0.8	67
	4	0.9	0344

Fig. 1. Stem-and-leaf plot of quadratically weighted kappa statistics showing inter-rater agreement on scoring of the 30 items of the Stroke Rehabilitation Assessment of Movement (STREAM) instrument. Numbers to the left of each box represent the number of test items; numbers in each box (stem) represent the first decimal place of the kappa statistic; numbers to the right of the box (leaves) represent the second decimal place of the kappa statistic. Therefore, for the three items in the second row, the values would be 0.55, 0.56 and 0.59.

Statistical analysis

The inter-rater agreement on individual items of the STREAM instrument was analyzed using the quadratic weighted kappa statistic. The weighted kappa score measures the agreement among raters adjusted for the amount of agreement expected by chance and the magnitude of disagreements (15). A kappa value of more than 0.75 indicates excellent agreement, 0.4–0.75 indicates fair to good agreement, and less than 0.4 indicates poor agreement (16).

The inter-rater reliability of the subscale scores and the total score of the STREAM instrument were analyzed using the intra-class correlation coefficient (ICC) statistic. The ICC was employed to examine the degree of agreement between repeated measurements taken by the two raters on the same patient. The ICC expresses measurement error and agreement as the relation between true variance and observed variance. The ICC not only can provide estimates of both association and agreement but also can be used with more than two sets of data (e.g. raters) (17). A two-way ANOVA was employed to compute the variances needed to estimate the inter-rater reliability ICC values. The fixed effect of ICC Model 3 (18) was used to compute the ICC value for inter-rater reliability. An ICC value of more than 0.80 indicates high reliability (19). The 95% confidence interval was calculated for each ICC to take sampling variation into account. Paired *t*-test was performed on the mean difference between scores obtained on the two STREAM measurements to determine the presence of a systematic bias.

The degree of validity was assessed using Spearman's rank-order correlation coefficient. We examined the relationship between the total score and subscales of the STREAM instrument with those of other well-established instruments. The average score of the two raters of the STREAM instrument was used.

RESULTS

Figure 1 is a stem-and-leaf plot summarizing the distribution of weighted kappa statistics for the inter-rater agreement on scores for individual items. Weighted kappa statistics for each of the 30 items ranged from 0.55 to 0.94 indicating moderate-to-excellent agreement.

The ICC for the total score was 0.96 (95% confidence interval: 0.94–0.98, F = 54.2, $p < 0.0001$) indicating very high inter-rater reliability. ICCs were also very high in each of the subscales (Table III).

Paired *t*-test showed a systematic bias on two subscales (voluntary movement of lower limb, basic mobility) and hence on the total score. One rater scored systematically higher than the other rater (mean difference = 2.7 ± 4.5, $p < 0.001$). Figure 2 shows the relationship between the patients' scores on the STREAM instrument as rated by the two raters.

Correlation analyses showed that the subscales of STREAM instrument results were closely associated with the FMUE, FMLE and RMI measurements, Spearman's rho = 0.87, 0.78, and 0.83 respectively; ($p < 0.001$). The total score of the STREAM instrument was moderately to highly associated with the score of the BI and FM, rho = 0.67, and 0.95, respectively; $p < 0.001$.

DISCUSSION

The psychometric characteristics of the STREAM instrument have rarely been examined. The first objective of this study was to determine the inter-rater reliability of the STREAM instrument. We also investigated the convergent validity of the STREAM instrument by examining the relation between performance on the STREAM instrument and performance on other well-validated measurements.

Item reliability analysis revealed that inter-rater agreement ranged from a weighted kappa of 0.55 to 0.94, indicating moderate-to-excellent agreement. The subscales and total score on the STREAM instrument also had a high inter-rater reliability. These data support that the STREAM instrument is reliable at the individual item, subscale, and over-all scale levels when performed by different raters. These results are similar to the findings of Daley et al. (2). However, the high inter-rater

Table III. *Inter-rater reliability analysis of the Stroke Rehabilitation Assessment of Movement (STREAM) instrument*

	Rater A Mean (SD)	Rater B Mean (SD)	Mean difference	Paired *t* value (*p*)	ICC (95% CI)
UE[a]	5.4 (5.9)	5.8 (6.2)	0.33	1.31 (0.194)	0.95 (0.92–0.97)
LE[b]	5.7 (4.7)	7.1 (5.8)	1.46	4.98 (<0.001)	0.92 (0.86–0.95)
Mobility[c]	14.6 (7.7)	15.5 (8.0)	0.87	2.06 (0.044)	0.92 (0.87–0.95)
Total score	25.7 (16.1)	28.4 (17.7)	2.67	4.31 (<0.001)	0.96 (0.94–0.98)

[a] The subscale scores of voluntary movement of upper limbs of the STREAM instrument.
[b] The subscale scores of voluntary movement of lower limbs of the STREAM instrument.
[c] The subscale scores of basic mobility of the STREAM instrument.
ICC, Intraclass correlation coefficient; CI, Confidence interval.

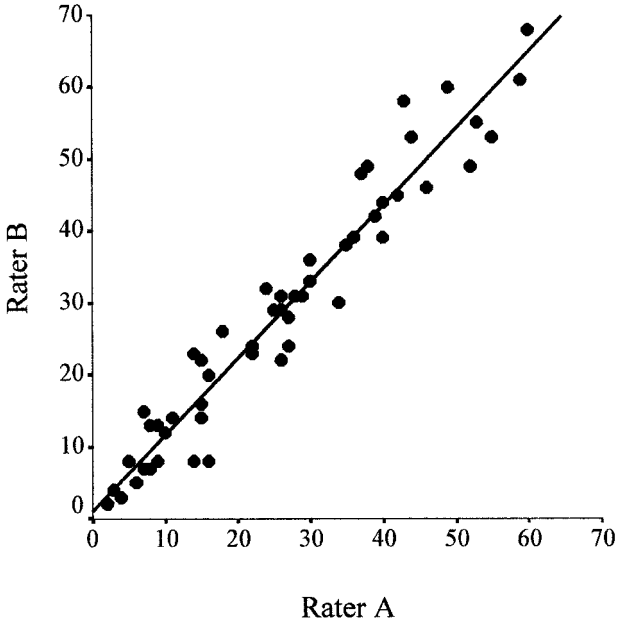

Fig. 2. The relationship between the patients' scores on the STREAM instrument as rated by the two raters.

reliability in the present study might have been due to use of a small number of well-trained and experienced therapists. Untrained raters and raters with less experience, either with stroke patients or with administration of the STREAM instrument, may not achieve a similar degree of consistency.

A systematic bias was found for two subscales (voluntary movement of lower limb, basic mobility) and for the total score (Table III). However, the magnitude of the mean difference in scores between the two raters was small (mean difference = 2.7 out of 70 for the total score). Therefore, this statistically significant difference may not be clinically significant. Increasing sample size for reliability studies yields more precise estimation of ICCs but increases the likelihood of disclosing systematic biases that are not of clinical significance (20). The estimation of ICCs takes into account the systematic bias and the random error (21). Our results showed that ICCs were very high in spite of the presence of systematic bias between the raters.

This study provides support for the concurrent validity of the STREAM instrument. High associations were found between the performance on the upper limb subscale of the STREAM instrument and the FMUE, the lower limb subscale of the STREAM instrument and the FMLE, and the mobility subscale of the STREAM instrument and the RMI. The high associations between the subscales of the STREAM instrument and other well-validated instruments imply a similarity of constructs. These results support the utility of the STREAM instrument to assess voluntary movement and basic mobility in persons who have experienced a stroke.

In the absence of a gold standard, validity can be established by assessing convergent validity (22). We compared the total scores of the STREAM with scores of the FM, and the BI to assess the convergent validity of the STREAM. The results showed the scores of the patients on the STREAM instrument were moderately to highly associated with those of the BI and FM. These findings further support the validity of the STREAM instrument.

Although an objective, quick, simple and quantitative evaluation tool is required for clinical assessment of recovery following a stroke, most of the assessment tools cannot fit this demand. In addition to having good reliability and validity data, the STREAM instrument has some other important features including ease of administration and completion within 15 minutes. The STREAM instrument may therefore fit very well into the routine clinical assessment process.

Any measurement tool requires extensive examination for the purposes of understanding its particular strengths and limitations (23). Without such analysis, there can be no confidence that it performs in the ways that its developers and users intended. Further research is needed to compare the performance of raters of the STREAM instrument from different disciplines with varying levels of experience. Studies of the STREAM instrument to examine its predictive validity and sensitivity to change and also in other patient groups and age ranges are needed to further establish the clinical utility of the STREAM instrument.

ACKNOWLEDGEMENTS

We are grateful to Dr Mayo and her coworkers for their construction of the Stroke Rehabilitation Assessment of Movement instrument, and agreed the authors to use the STREAM instrument. We also would like to thank Professor Kenneth J. Ottenbacher for his helpful comments on manuscript of this paper.

REFERENCES

1. Sanford J, Moreland J, Swanson LR, Stratford PW, Gowland C. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. Phys Ther 1993; 73: 447–454.
2. Daley K, Mayo N, Wood-Dauphinee S. Reliability of scores on the stroke rehabilitation assessment of movement (STREAM) measure. Phys Ther 1999; 79: 8–23.
3. Halsaa KE, Sødring KM, Bjelland E, Finsurd K, Bautz-Holter E. Inter-rater reliability of the Sødring motor evaluation of stroke patients (SMES). Scand J Rehabil Med 1999; 31: 240–243.
4. Daley K, Mayo N, Danys I, Cabot R, Wood-Dauphinee S. The stroke rehabilitation assessment of movement (STREAM): refining and validating the content. Physio Can 1997; 49: 269–278.
5. Sheikh K. Disability scales: assessment of reliability. Arch Phys Med Rehabil 1986; 67: 245–249.
6. Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient, I: a method for evaluation of physical performance. Scand J Rehabil Med 1975; 7: 13–31.
7. Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead motor assessment. Int Disabil Stud 1991; 13: 50–54.
8. Mahoney F, Barthel D. Functional evaluation: the Barthel Index. Md Med J 1965; 14: 61–65.
9. Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Phys Ther 1983; 63: 1606–1610.
10. Berglund D, Fugl-Meyer A. Upper extremity function in hemi-

plegia: a cross-validation study of two assessment methods. Scand J Rehabil Med 1986; 18: 155–157.
11. Hsieh CL, Hsueh IP, Mao HF. Validity and responsiveness of the Rivermead Mobility Index in stroke patients. Scand J Rehabil Med 2000; 32: 140–142.
12. Granger CV, Devis LS, Peters MC, Sherwood CC, Barrett JE. Stroke rehabilitation: analysis of repeated Barthel Index measures. Arch Phys Med Rehabil 1979; 60: 14–17.
13. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press; 1992.
14. Collin C, Wade DT, Davies S, Hoene V. The Barthel ADL index: a reliability study. Int Disabil Stud 1988; 10: 61–63.
15. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213–220.
16. McCluggage WG, Bharucha H, Caughley LM, Date A, Hamilton PW, Thornton CM, et al. Interobserver variation in the reporting of cervical colposcopic biopsy specimens: comparison of grading systems. J Clin Pathol 1996; 49: 833–835.
17. Ottenbacher K, Stull GA. The analysis and interpretation of medical comparison studies in rehabilitation research. Am J Phys Med Rehabil 1993; 72: 266–271.
18. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–428.
19. Richman J, Makrides L, Prince B. Research methodology and applied statistics. Part 3: measurement procedures in research. Physio Can 1980; 32: 253–257.
20. Desrosiers J, Bravo G, Hebert R, Dubuc N. Reliability of the revised functional autonomy measurement system (SMAF) for epidemiological research. Age Aging 1995; 24: 402–406.
21. Armstrong BK, White E, Saracci R. Principles of exposure measurement in epidemiology. Oxford: Oxford University Press; 1992.
22. Sharrack B, Hughes RAC, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999; 122: 141–159.
23. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. Arch Phys Med Rehabil 1993; 74: 531–536.