Research Report

Psychometric Comparisons of 4 Measures for Assessing Upper-Extremity Function in People With Stroke

Jau-Hong Lin, Miao-Ju Hsu, Ching-Fan Sheu, Tzung-Shian Wu, Ruey-Tay Lin, Chia-Hsin Chen, Ching-Lin Hsieh

Background. Functional limitation of the upper extremities is common in patients with stroke. An upper-extremity measure with sound psychometric properties is indispensable for clinical and research use.

Objective. The purpose of this study was to compare the psychometric properties of 4 clinical measures for assessing upper-extremity motor function in people with stroke: the upper-extremity subscale of the Fugl-Meyer Motor Test (UE-FM), the upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, the Action Research Arm Test (ARAT), and the Wolf Motor Function Test.

Design. This was a prospective, longitudinal study.

Methods. Fifty-three people with stroke were evaluated with the 4 measures at 4 time points (14, 30, 90, and 180 days after stroke). Thirty-five participants completed all of the assessments. The ceiling and floor effects, validity (concurrent validity and predictive validity), and responsiveness of each measure were examined. Interrater reliability and test-retest reliability also were examined.

Results. All measures, except for the UE-FM, had significant floor effects or ceiling effects at one or more time points. The Spearman ρ correlation coefficient for each pair of the 4 measures was $\geq$.81, indicating high concurrent validity. The predictive validity of the 4 measures was satisfactory (Spearman ρ, $\geq$.51). The responsiveness of the 4 measures at 14 to 180 days after stroke was moderate (.52 $\leq$ effect size $\leq$ .79). The 4 measures had good interrater reliability (intraclass correlation coefficient [ICC], $\geq$.92) and test-retest reliability (ICC, $\geq$.97). Only the minimal detectable changes of the UE-FM (8% of the highest possible score) and the ARAT (6%) were satisfactory.

Limitations. The sample size was too small to conduct data analysis according to type or severity of stroke. In addition, the timed component of the Wolf Motor Function Test was not used in this study.

Conclusions. All 4 measures showed sufficient validity, responsiveness, and reliability in participants with stroke. The UE-FM for assessing impairment and the ARAT for assessing disability had satisfactory minimal detectable changes, supporting their utility in clinical settings.

J.-H. Lin, PT, PhD, is Professor, Department of Physical Therapy, College of Health Science, Kaohsiung Medical University, Kaohsiung, Taiwan, and Department of Rehabilitation, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan.

M.-J. Hsu, PhD, is Assistant Professor, Department of Physical Therapy, College of Health Science, Kaohsiung Medical University, and Department of Rehabilitation, Kaohsiung Medical University Hospital.

C.-F. Sheu, PhD, is Professor, Institute of Education, National Cheng Kung University, Tainan, Taiwan.

T.-S. Wu, BS, PT, is a postgraduate student, Department and Graduate Institute of Neurology, College of Medicine, Kaohsiung Medical University.

R.-T. Lin, MD, is Associate Professor, Department of Neurology, Kaohsiung Medical University Hospital.

C.-H. Chen, PhD, is Assistant Professor, Department of Physical Medicine and Rehabilitation, Kaohsiung Medical University Hospital.

C.-L. Hsieh, OT, PhD, is Professor, School of Occupational Therapy, College of Medicine, National Taiwan University, 4F, No. 17, Xuzhou Rd, Taipei 100, Taiwan, and Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan. Address all correspondence to Dr Hsieh at: clhsieh@ntu.edu.tw.

[Lin J-H, Hsu M-J, Sheu C-F, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. *Phys Ther.* 2009;89:840–850.]

Functional limitation of the upper extremities (UEs) is one of the most common disabling deficits after stroke.[1,2] At hospital admission after stroke onset, more than two thirds of all patients have an arm impairment resulting in disability affecting daily living,[2,3] and only one third of all patients with stroke have regained some dexterity at 6 months after stroke.[4] Both impairment and disability are key elements of the assessment of people with stroke. For clinicians and researchers studying and treating UE impairment and resulting disability after stroke, a UE measure with sound psychometric properties (ie, reliability, validity, and responsiveness) is indispensable.

Several measures for assessing UE impairment or disability have been developed.[5–10] However, no single instrument is universally accepted for research or clinical use. Commonly used measures include the UE subscale of the Fugl-Meyer Motor Test (UE-FM),[6] the UE subscale of the Stroke Rehabilitation Assessment of Movement (UE-STREAM),[7] the Action Research Arm Test (ARAT),[8] and the Wolf Motor Function Test (WMFT).[5]

The UE-FM, which is composed of 33 items related to movements of the proximal and distal parts of the UEs, is the measure most frequently used to evaluate UE impairment.[6] Several studies evaluating the psychometric properties of the UE-FM in people with stroke have demonstrated satisfactory reliability, validity, and responsiveness.[6,10–15] However, the ceiling and floor effects of the UE-FM for people with stroke throughout different recovery stages—which are crucial to the determination of whether the measure assesses a limited range of impairment—have rarely been reported.[16]

The STREAM was developed as an outcome measure for assessing the motor impairments and basic mobility of people with stroke.[7] It consists of three 10-item subscales: upper-limb movements (UE-STREAM), lower-limb movements, and mobility.[7] The STREAM, which has good interrater and intrarater reliability and internal consistency, is sensitive to changes in people with stroke.[17–19] However, the predictive validity of the UE-STREAM remains largely unknown, a fact that could limit its utility for predicting health outcomes.

The ARAT is a standardized ordinal scale that was designed to measure UE disability through the assessment of 4 basic movements: primary grasp, grip, pinch, and gross movements of flexion and extension at the elbow and shoulder.[8,20] The ARAT can be completed in 10 minutes.[14,21,22] However, the use of the ARAT is not always feasible because of the requirement for specific materials (eg, a specially designed table).[22] The reliability, validity, and responsiveness of the ARAT for people with stroke have been established.[8,13,23–25] However, the minimal detectable change (MDC)[26] of the ARAT is lacking, limiting the ability of users to determine whether there has been real improvement (beyond random measurement error) between repeated assessments for a patient.

The WMFT was developed to assess UE disability in people with chronic stroke and receiving constraint-induced movement therapy.[5,27] It was reduced to 17 items, including 2 strength (force-generating capacity) measurements and 15 function-based tasks, in the most recent version.[27] The 15 function-based tasks of the WMFT are divided into 2 scales: performance time and functional ability. Studies exploring the psychometric properties of the WMFT have focused largely on people with subacute and chronic stroke and mild hemiparesis, and the results have shown that the WMFT has high interrater and test-retest reliability for both performance time and functional ability.[28,29] The criterion validity is supported by the significant relationship between UE-FM and WMFT scores for the affected limb in people with chronic stroke.[27] Nevertheless, to date, the responsiveness and MDC of the WMFT have not been fully examined. Therefore, further research on the psychometric properties of the WMFT is warranted.

Although the psychometric properties of these 4 measures have been investigated, at least 2 limitations can be noted. First, few studies have compared the psychometric properties of the different UE measures in the same cohort of patients with stroke.[13,24,30–33] Because psychometric properties are sample dependent,[34,35] it is difficult to interpret the results of studies comparing UE measures across samples with different characteristics. Second, the scopes of the psychometric properties of the UE measures examined in previous studies are limited, notable omissions being predictive validity and measurement errors (eg, MDC) within the same raters or among different raters.[16] These limitations make interpretation of the resulting UE measures difficult, particularly in determining whether a difference represents real change or measurement error. Thus, a comprehensive comparison of the psychometric properties of commonly used UE measures is warranted.

This article was published ahead of print on June 25, 2009, at www.ptjournal.org.

Table 1.
Basic Characteristics of Participants With Stroke[a]

Characteristic	Value[b] at Indicated Days After Stroke (No. of Participants)				Value[b] for Participants Who:	
	14 (53)	30 (42)	90 (36)	180 (35)	Completed Follow-up at 180 Days After Stroke (n=35)	Dropped Out (n=18)
Sex, no. of men/women	31/22	26/16	20/16	21/14	21/14	10/8
Age, y	64.3 (11.7)	63.9 (11.2)	64.0 (11.6)	64.0 (12.5)	64.0 (12.5)	64.7 (10.3)
Diagnosis, no. of participants						
Cerebral hemorrhage	19	16	12	12	13	6
Cerebral infarction	34	26	24	23	22	12
Side of stroke, no. of participants						
Right	29	21	17	14	26	3
Left	24	21	19	21	9	15
BI	9.1 (5.6)	13.8 (5.8)	16.8 (4.7)	16.6 (5.0)	9.2 (5.6)	8.9 (5.8)
UE-FM score	32.0 (25.9)	43.9 (25.3)	48.0 (23.7)	46.0 (24.0)	32.4 (26.2)	31.2 (26.0)
UE-STREAM score	10.0 (7.8)	13.1 (7.7)	14.7 (7.1)	14.5 (7.0)	9.8 (7.8)	10.3 (8.2)
ARAT score	18.6 (22.7)	31.4 (24.2)	37.7 (23.4)	36.3 (23.6)	18.5 (22.5)	18.7 (23.7)
WMFT score	33.7 (30.6)	49.0 (29.2)	55.3 (26.9)	53.1 (27.4)	33.3 (30.7)	34.4 (31.4)

[a] BI=Barthel Index, UE-FM=upper-extremity subscale of the Fugl-Meyer Motor Test, UE-STREAM=upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, ARAT=Action Research Arm Test, WMFT=Wolf Motor Function Test.
[b] Data are reported as $\overline{X}$ (SD), unless otherwise indicated.

The purpose of this study was to compare the reliability (test-retest reliability, interrater reliability, and MDC), validity (concurrent validity and predictive validity), and responsiveness of the UE-FM, UE-STREAM, ARAT, and WMFT in people with stroke at different recovery stages.

Method

Participants

The study protocol consisted of 2 stages. In the first stage of the study, the interrater reliability, validity, and responsiveness were tested in people with stroke who were admitted consecutively to the Department of Neurology at Kaohsiung Medical University Hospital from September 1, 2006, to August 31, 2007. A total of 120 people with stroke were contacted through a neurologist and were invited to participate in this stage of the study if they met the following criteria: a diagnosis of cerebral hemorrhage or cerebral infarction, first onset of stroke, stroke onset within 2 weeks before hospital admission, and ability to follow instructions to complete the testing. We excluded people with other major diseases (eg, cancer, dementia, severe rheumatoid arthritis), preexisting disabilities, or another stroke during the follow-up period and people living outside a 30-km radius from the hospital.

In the second stage of the study, test-retest reliability was examined. We recruited another independent sample of people with chronic stroke and undergoing outpatient therapy in the rehabilitation department. These people had had a stroke at least 1 year before recruitment and voluntarily participated in the study. We excluded people with unstable medical conditions and people who were unable to follow instructions.

A total of 53 people who met the selection criteria participated in the study. However, 11 individuals who either had a recurrent stroke during hospitalization or declined to participate were not evaluated further. Another 6 individuals were lost to follow-up at 90 days after stroke, and another individual was lost to follow-up between 90 and 180 days after stroke. A total of 35 participants completed all of the assessments. Table 1 shows the basic characteristics of the participants at different time points (14, 30, 90, and 180 days) and the participants who completed follow-up at 180 days, as well as the participants who did not complete follow-up at 180 days.

Written informed consent was obtained from each individual before participation in this study.

Procedure

For validity and responsiveness, the 4 UE measures and the Barthel Index (BI) were administered by a physical therapist (therapist A) to the participants at 14, 30, 90, and 180 days after stroke. The BI score at 180 days after stroke was used as the criterion for examining the predictive validity of the 4 measures administered at 14 days after stroke.

For interrater reliability, 2 specially trained physical therapists (therapists A and B) individually administered the 4 measures to the participants at 14 days after stroke. Both therapists individually administered the 4 measures within a 2-day period to minimize the effects of a possible spontaneous recovery. The sequence of testing was random and counterbalanced for both therapists. For test-retest reliability, therapist B administered the 4 measures twice, 1 week apart, to an independent sample of participants with chronic stroke.

During the testing periods, participants could rest as much as they wanted. Participants' demographic details and major comorbidity data were collected from medical records.

Measures

The FM consists of the 33-item upper-extremity subscale (UE-FM) and the 17-item lower-extremity subscale.[6] The items of the FM are mainly scored on a 3-point scale, from 0 to 2. The total score on the UE-FM ranges from 0 to 66.

The STREAM consists of three 10-item subscales: upper-limb movements (UE-STREAM), lower-limb movements, and mobility.[7] Extremity movements are scored on a 3-point scale, from 0 to 2. The total score on the UE-STREAM ranges from 0 to 20.

The ARAT[8] comprises 19 items in 4 categories: grasp, grip, pinch, and gross movements. Each item is graded on a 4-point scale, from 0 to 3. The total score on the ARAT ranges from 0 to 57.

The 15 function-based tasks of the WMFT[5,28] are divided into 2 scales: performance time and functional ability. In this study, only the functional ability scale was used to assess UE movement components required for daily tasks. We did not include the timed component of the WMFT because most of the participants in our pilot test took longer than 2 minutes to complete the required tasks, a finding that obviously would cause a floor effect. The quality of movement is scored with a 6-point scale, with scores ranging from 0 (not attempted) to 5 (normal movement). The total score on the WMFT ranges from 0 to 75.

The BI is a measure of function in basic activities of daily living (ADL).[36] Scores on the BI range from 0 to 20. The reliability, validity, and responsiveness of the BI in subjects with stroke are well established.[37,38] The key features and detailed items of the 4 measures are summarized in the Appendix.

Data Analysis

Distribution. The score distributions of the 4 measures were examined for floor and ceiling effects. The floor effect is the percentage of the sample scoring the minimum possible points, reflecting the extent to which scores cluster at the bottom of the scale range. The ceiling effect represents the opposite extreme. Floor or ceiling effects exceeding 20% of the sample size were considered substantial.[39]

Validity. Concurrent validity was established by examining the interrelationships of each pair of the 4 measures at 4 time points with the Spearman ρ correlation coefficient. A ρ value of between 0 and .25 was considered to represent a low association, a value of between .25 and .5 represented a fair association, a value of between .5 and .75 represented a moderate association, and a value of greater than .75 represented a high association.[40]

Predictive validity was assessed by examining the linear associations between the scores on the 4 measures at 14 days after stroke and the BI score at 180 days after stroke with the Spearman ρ correlation coefficient. We used the aforementioned criteria for interpretation.[40]

Responsiveness. Two approaches were used to examine the responsiveness of each measure during 3 periods: 14 to 30, 14 to 90, and 14 to 180 days after stroke. First, the effect size was defined as the observed mean change scores divided by the standard deviation of the baseline score. According to the criteria of Cohen,[41] effect sizes of greater than .8 are large, sizes of .5 to .8 are moderate, and sizes of .2 to .5 are small. Second, we used the Wilcoxon matched-pairs signed rank test to determine the statistical significance of the change scores.

Reliability. The interrater reliability and test-retest reliability of the 4 measures were analyzed with the intraclass correlation coefficient (ICC). We used the fixed-effect model of the ICC[42] to examine the degree of agreement between repeated measurements by the 2 raters for the same participant. In addition, we used the random-effect model of the ICC to determine the level of agreement between test-retest assessments. Intraclass correlation coefficients of $\geq$.80 indicate high agreement.[43]

We quantified random measurement errors with the standard error of measurement (SEM) as follows: (standard deviation of all test-retest scores) $\times \sqrt{(1-\text{ICC})}$. The MDC

Table 2.
Floor and Ceiling Effects of the 4 Measures at Different Recovery Stages[a]

Days After Stroke (No. of Participants)	UE-FM		UE-STREAM		ARAT		WMFT	
	Floor Effect	Ceiling Effect	Floor Effect	Ceiling Effect	Floor Effect	Ceiling Effect	Floor Effect	Ceiling Effect
14 (53)	9.4	5.7	20.8	18.9	41.5	9.4	26.4	17.0
30 (42)	1.9	17.0	5.7	26.4	17.0	20.8	9.4	24.5
90 (36)	5.7	17.0	1.9	30.2	11.3	20.8	3.8	28.3
180 (35)	1.9	17.0	1.9	32.1	11.3	22.6	7.5	24.5

[a] Data are reported as percentages of participants. UE-FM=upper-extremity subscale of the Fugl-Meyer Motor Test, UE-STREAM=upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, ARAT=Action Research Arm Test, WMFT=Wolf Motor Function Test.

Table 3.
Concurrent Validity of the 4 Measures at Different Recovery Stages[a]

Days After Stroke (No. of Participants)	UE-FM vs UE-STREAM	UE-FM vs ARAT	UE-FM vs WMFT	UE-STREAM vs ARAT	UE-STREAM vs WMFT	ARAT vs WMFT
14 (53)	.96	.90	.93	.89	.95	.92
30 (42)	.94	.90	.96	.94	.91	.97
90 (36)	.93	.82	.85	.94	.84	.81
180 (35)	.94	.92	.94	.87	.90	.92

[a] Data are reported as Spearman ρ correlations. UE-FM=upper-extremity subscale of the Fugl-Meyer Motor Test, UE-STREAM=upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, ARAT=Action Research Arm Test, WMFT=Wolf Motor Function Test.

Table 4.
Responsiveness of the 4 Measures at Different Recovery Stages[a]

Days After Stroke (No. of Participants)	Effect Size[b]				Wilcoxon Test[c]			
	UE-FM	UE-STREAM	ARAT	WMFT	UE-FM	UE-STREAM	ARAT	WMFT
14–30 (42)	.37	.33	.49	.44	5.0	4.2	4.5	4.4
14–90 (36)	.48	.51	.70	.58	4.3	4.2	4.4	4.2
14–180 (35)	.52	.60	.79	.64	4.4	4.5	4.5	4.5

[a] UE-FM=upper-extremity subscale of the Fugl-Meyer Motor Test, UE-STREAM=upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, ARAT=Action Research Arm Test, WMFT=Wolf Motor Function Test.
[b] An effect size of .2–.5 is small; an effect size of .5–.8 is moderate.
[c] $P<.001$ for all values.

($1.96 \times \text{SEM} \times \sqrt{2}$)[44] was used as a threshold to determine whether the change score for an individual subject was real at the 95% confidence level. The MDC of a measure was considered satisfactory when the MDC was less than 10% of the highest possible score on the measure.[45]

Role of the Funding Source

This study was supported by a research grant from the National Science Council (NSC95-2314-B-037-068).

Results

The participants had a wide range of disability, and their sum scores on the BI were scattered throughout the full range of scores (0–20). Comparing the characteristics of the 35 participants who completed follow-up with those of the 18 participants who did not complete follow-up, we found no significant differences in BI, UE-FM, UE-STREAM, ARAT, or WMFT motor scores at 14 days after stroke.

Table 2 shows that the UE-STREAM, ARAT, and WMFT had significant floor effects (≥21% of the participants) at 14 days after stroke and notable ceiling effects (≥21% of the participants) at 30, 90, and 180 days after stroke. The UE-FM was the only measure that did not exhibit obvious floor or ceiling effects at any of the 4 time points.

The correlations for each pair of the 4 measures at the 4 time points were high (ρ=.81–.97) (Tab. 3). The scores on the 4 measures at 14 days

Table 5.
Interrater Reliability for 30 Participants With Acute Stroke and Test-Retest Reliability for 30 Participants With Chronic Stroke[a]

Measure (Possible Score Range)	Interrater Reliability			Test-Retest Reliability		
	ICC (95% CI)	MDC	MDC%	ICC (95% CI)	MDC	MDC%
UE-FM (0–66)	.96 (.92–.98)	12.9	20	.99 (.99–1.00)	5.2	8
UE-STREAM (0–20)	.96 (.92–.98)	3.9	20	.99 (.97–.99)	2.3	12
ARAT (0–57)	.95 (.90–.98)	13.1	23	.99 (.99–1.00)	3.5	6
WMFT (0–75)	.92 (.85–.97)	20.2	27	.97 (.94–.99)	12.0	16

[a] ICC=intraclass correlation coefficient, 95% CI=95% confidence interval, MDC=minimal detectable change, MDC%=MDC/highest possible score of a measure, UE-FM=upper-extremity subscale of the Fugl-Meyer Motor Test, UE-STREAM=upper-extremity subscale of the Stroke Rehabilitation Assessment of Movement, ARAT=Action Research Arm Test, WMFT=Wolf Motor Function Test.

after stroke were moderately correlated with those on the BI at 180 days after stroke (ρ=.51–.59).

Table 4 shows that the 4 measures had moderate responsiveness (effect sizes of ≥.52) in detecting changes from 14 to 180 days after stroke and generally low to moderate responsiveness during the other periods (at 14–30 days after stroke, effect sizes ranged from .33 to .49; at 14–90 days after stroke, effect sizes ranged from .48 to .70), as determined with the benchmarks of Cohen.[41] The changes in the 4 measures were all significant (P<.001).

A total of 30 participants were included in the interrater reliability analysis because 23 of the original 53 participants in the sample were either unable or unwilling to be retested within 48 hours. This group consisted of 13 women and 17 men, with a mean age of 61.7 years (SD=11.3 years). Table 5 shows that interrater reliability for the 4 measures was high (ICC, ≥.92; lower limit of 95% confidence interval, ≥.85). However, the MDCs of the 4 measures were ≥10% of their corresponding highest scores.

An independent sample of 30 participants with chronic stroke participated in the test-retest reliability study. This group consisted of 14 women and 16 men, with a mean age of 56.6 years (SD=11.6 years) and a mean of 693.2 days (SD=56.6 days) from stroke onset to admission. The test-retest reliability of the 4 measures was high (ICC, ≥.97; lower limit of 95% confidence interval, ≥.94). Only the MDCs of the UE-FM and the ARAT were below 10% of their corresponding highest scores.

Discussion and Conclusions

The present study is the first to concurrently and systematically compare the psychometric properties of the UE-FM, UE-STREAM, ARAT, and WMFT in a sample of people with stroke. In addition, we evaluated participants at 4 specific time points up to 180 days after stroke to assess how appropriate these measures are for use. Our findings provide an empirical foundation on which clinicians and researchers may base the selection of UE motor measures for people at different recovery stages after stroke.

The distribution of UE motor measures at different recovery stages after stroke has rarely been reported.[16,24] However, such information is important in determining whether a measure assesses only a restricted functional range in people with stroke. Hsueh and Hsieh[24] reported that the ARAT showed notable floor effects in 48 inpatients receiving rehabilitation after stroke. More recently, the UE-STREAM showed notable ceiling effects at admission and discharge in inpatients undergoing rehabilitation.[16] In the present study, all of the UE motor measures tested, except for the UE-FM, showed notable floor effects at 14 days after stroke and ceiling effects at 30, 90, and 180 days after stroke. The smaller floor effect seen with the UE-FM may have been the result of 9 participants scoring points on some of the flexor and extensor reflex items at 14 days after stroke, even though they had no active movement. The ceiling effects of 3 measures at 30, 90, and 180 days after stroke may have been attributable to the loss of participants with severe impairments in the follow-up evaluations. However, compared with the individuals who dropped out, the participants who completed follow-up did not show significant differences in motor scores on the 4 measures. These results indicate that the UE-FM assesses a wider spectrum of UE motor function and is more discriminative for people with very poor or very good motor function than the other 3 measures at different recovery stages.

The validity of a measure is of critical importance because it represents whether the measure assesses what it intends to measure. De Weerdt and Harrison[14] reported that the relationship between scores on the UE-FM and the ARAT was extremely high at

2 (Pearson r, .91) and 8 (Pearson r, .94) weeks after stroke onset. Hsieh et al[23] demonstrated that the score on the ARAT was closely correlated (Pearson r, $>$.87) with scores on the other well-validated measures for evaluating UE impairment and disability. Wang et al[18] found that the UE-STREAM was closely associated with the UE-FM (Spearman ρ, .87) in people at 25 to 361 days after stroke onset. In the present study, the association between each pair of measures was extremely high (ρ=.81–.97). Our results are generally in accordance with the findings of previous studies.[14,18,19,23,29,46] These observations provide strong evidence of the concurrent validity of the 4 measures for assessing UE motor function in people with stroke and demonstrate that for the UE, impairment scores (on the UE-FM or the UE-STREAM) are closely correlated with the level of disability (on the ARAT or the WMFT).

The early prediction of a patient's functional status is important for patient care.[47] Chae et al[48] found that the FM score was a good predictor of disability after rehabilitation for stroke, as measured with the Functional Independence Measure. Ahmed et al[19] reported that the STREAM score during the first week after stroke was able to predict the BI score after 3 months. Nevertheless, those 2 previous studies did not address the potential impact of UE subscale scores on ADL. In the present study, the BI was used as the criterion for investigating predictive validity. However, a patient with severely impaired UE function might still be able to score high on the BI by performing ADL with compensatory strategies. Such a situation would compromise the relationship between ADL function and UE function. Our finding of a moderate relationship between the BI score and the scores on the 4 UE measures sufficiently supports the predictive validity of the 4 UE measures. Our findings further confirm the validity and clinical utility of the 4 measures.

Responsiveness is important for any measurement tool designed to evaluate change over time.[49] To our knowledge, few studies have compared the responsiveness of the 4 measures.[16,19,30,33] A previous study indicated that the UE-FM and the ARAT were equally sensitive to changes during inpatient rehabilitation for acute stroke.[30] Another study showed that the ARAT was more responsive to improvements in UE function than the UE-FM in people with chronic stroke.[33] More recently, both the UE-FM and the UE-STREAM showed appropriate responsiveness (effect sizes, .34 and .38, respectively) in detecting changes during hospital rehabilitation.[16] In the present study, the ARAT showed the highest responsiveness among the 4 measures at different recovery stages. Furthermore, these measures showed low responsiveness at the early stage of recovery (14–30 days after stroke), according to the criteria of Cohen, but they showed moderate responsiveness in detecting changes during the other periods (14–90 and 14–180 days after stroke). All changes in the 4 measures at each stage were significant. In other words, our findings suggest that the 4 measures are able to detect small changes in subjects.

The test-retest agreement of the 4 measures was very high (ICC, ≥.97). This result is consistent with those reported in previous studies.[6,10,11,17,28,50,51] Our observations suggest that the 4 measures are highly reliable in monitoring changes in patients' UE motor function when used by trained raters.

The values for MDCs are useful for clinicians in determining whether an individual patient has achieved real changes.[26] We found that only the MDCs of the UE-FM and the ARAT were below 10% of their corresponding highest scores, indicating a satisfactory level of measurement error. Our findings suggest that changes of more than 6 points, 3 points, 4 points, and 12 points in the total scores on the UE-FM (highest possible score: 66), UE-STREAM (20), ARAT (57), and WMFT (75), respectively, for each patient assessed by an individual rater are not likely to be attributable to chance variation or measurement error and can be interpreted by clinicians as a real change with 95% confidence. For example, a previous single-case study reported an improvement in the UE-FM score of about 9.5 points in a patient with chronic stroke after modified constraint-induced movement therapy[52]; because the improvement (9.5 points) in the patient exceeded the MDC (5.2 points) of the UE-FM, the treatment effect could be well justified.

Researchers usually report the mean change, P value, or effect size of a study group after intervention. However, even if the mean changes within a study group are significant, the number of participants in the study group whose changes achieve the MDC is still unknown. Thus, reporting the proportion of participants who have achieved improvement beyond the MDC helps translate research findings into clinical practice.

We found that the ICCs for the interrater reliability of the 4 measures were high (≥.92). Similar results were reported in previous studies.[11,18,27,50] The MDCs for the interrater reliability of the UE-FM, UE-STREAM, ARAT, and WMFT were 12.9, 3.9, 13.1, and 20.2, respectively. As expected, the MDCs obtained from different raters were higher than those obtained from a single rater. The MDC can help clinicians and researchers judge whether changes after therapy are actually real changes in UE motor function when the assessments are administered by different raters.

One of the limitations of the present study was that the sample size was too small to conduct data analysis according to type or severity of stroke. Further studies with a large sample of subjects with characteristics different from those enrolled in the present study are necessary to analyze the effects of the type of stroke or the level of severity on the psychometric properties of measures. The ages of the individuals in our sample also were slightly lower than the average age of stroke onset in Taiwan. The reason for this difference may be that we did not recruit patients who had more-severe impairments and, therefore, could not follow instructions to complete the tests. Such subjects are more likely to be older. In addition, our decision not to use the timed component of the WMFT further limited the scope of the present study.

In summary, the UE-FM, UE-STREAM, ARAT, and WMFT showed acceptable levels of reliability, validity, and responsiveness. As a measure of arm impairment, the UE-FM showed more-acceptable levels of measurement error than the UE-STREAM in our participants. For assessing UE disability, the ARAT showed more-acceptable levels of measurement error than the WMFT.

Dr J.-H. Lin and Dr Hsu provided concept/idea/research design. Dr J.-H. Lin, Dr Hsu, and Dr Sheu provided writing. Dr J.-H. Lin and Mr Wu provided data collection. Dr J.-H. Lin, Dr Hsu, Dr Sheu, and Mr Wu provided data analysis. Mr Wu provided project management. Dr R.-T. Lin, Dr Chen, and Dr Hsieh provided participants. Dr Hsu, Dr Sheu, Dr R.-T. Lin, Dr Chen, and Dr Hsieh provided consultation (including review of manuscript before submission).

All study procedures received ethics approval from the Institutional Review Board at Kaohsiung Medical University Chung-Ho Memorial Hospital. This study was supported by a research grant from the National Science Council (NSC95-2314-B-037-068).

This article was received September 15, 2008, and was accepted April 16, 2009.

DOI: 10.2522/ptj.20080285

References

1 Broeks JG, Lankhorst GJ, Rumping K, Prevo AJ. The long-term outcome of arm function after stroke: results of a follow-up study. *Disabil Rehabil*. 1999;21:357–364.

2 Nakayama H, Jorgensen HS, Raaschou HO, Olsen TS. Compensation in recovery of upper extremity function after stroke: the Copenhagen Stroke Study. *Arch Phys Med Rehabil*. 1994;75:852–857.

3 Jorgensen HS, Nakayama H, Pedersen PM, et al. Epidemiology of stroke-related disability. *Clin Geriatr Med*. 1999;15:785–799.

4 Dobkin BH. Clinical practice: rehabilitation after stroke. *N Engl J Med*. 2005;352:1677–1684.

5 Wolf SL, Lecraw DE, Barton LA, Jann BB. Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Exp Neurol*. 1989;104:125–132.

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

7 Daley K, Mayo N, Danys I, et al. The Stroke Rehabilitation Assessment of Movement (STREAM): refining and validating the content. *Physiother Can*. 1997;49:269–278.

8 Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res*. 1981;4:483–492.

9 Sabari JS, Lim AL, Velozo CA, et al. Assessing arm and hand function after stroke: a validity test of the hierarchical scoring system used in the Motor Assessment Scale for stroke. *Arch Phys Med Rehabil*. 2005;86:1609–1615.

10 Gowland C, Stratford PW, Ward M, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. *Stroke*. 1993;24:58–63.

11 Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. *Phys Ther*. 1983;63:1606–1610.

12 Berglund K, Fugl-Meyer AR. Upper extremity function in hemiplegia: a cross-validation study of two assessment methods. *Scand J Rehabil Med*. 1986;18:155–157.

13 Platz T, Pinkowski C, van Wijck F, et al. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. *Clin Rehabil*. 2005;19:404–411.

14 De Weerdt WJG, Harrison MA. Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnström-Fugl-Meyer test and the Action Research Arm test. *Physiother Can*. 1985;37:65–70.

15 Wood-Dauphinée SL, Williams JI, Shapiro SH. Examining outcome measures in a clinical study of stroke. *Stroke*. 1990;21:731–739.

16 Hsueh IP, Hsu MJ, Sheu CF, et al. Psychometric comparisons of 2 versions of the Fugl-Meyer Motor Scale and 2 versions of the Stroke Rehabilitation Assessment of Movement. *Neurorehabil Neural Repair*. 2008;22:737–744.

17 Daley K, Mayo N, Wood-Dauphinée S. Reliability of scores on the Stroke Rehabilitation Assessment of Movement (STREAM) measure. *Phys Ther*. 1999;79:8–19.

18 Wang CH, Hsieh CL, Dai MH, et al. Inter-rater reliability and validity of the Stroke Rehabilitation Assessment of Movement (STREAM) instrument. *J Rehabil Med*. 2002;34:20–24.

19 Ahmed S, Mayo NE, Higgins J, et al. The Stroke Rehabilitation Assessment of Movement (STREAM): a comparison with other measures used to evaluate effects of stroke and rehabilitation. *Phys Ther*. 2003;83:617–630.

20 Carroll D. A quantitative test of upper extremity function. *J Chronic Dis*. 1965;18:479–491.

21 van der Lee JH, Roorda LD, Beckerman H, et al. Improving the Action Research Arm test: a unidimensional hierarchical scale. *Clin Rehabil*. 2002;16:646–653.

22 Hsueh IP, Lee MM, Hsieh CL. The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? *Clin Rehabil*. 2002;16:382–388.

23 Hsieh CL, Hsueh IP, Chiang FM, Lin PH. Inter-rater reliability and validity of the Action Research arm test in stroke patients. *Age Ageing*. 1998;27:107–113.

24 Hsueh IP, Hsieh CL. Responsiveness of two upper extremity function instruments for stroke inpatients receiving rehabilitation. *Clin Rehabil*. 2002;16:617–624.

25 Lang CE, Wagner JM, Dromerick AW, Edwards DF. Measurement of upper-extremity function early after stroke: properties of the Action Research Arm Test. *Arch Phys Med Rehabil*. 2006;87:1605–1610.

26 Schuck P, Zwingmann C. The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. *Int J Rehabil Res*. 2003;26:85–91.

27 Wolf SL, Catlin PA, Ellis M, et al. Assessing Wolf Motor Function Test as outcome measure for research in patients after stroke. *Stroke*. 2001;32:1635–1639.

28 Morris DM, Uswatte G, Crago JE, et al. The reliability of the Wolf Motor Function Test for assessing upper extremity function after stroke. *Arch Phys Med Rehabil*. 2001;82:750–755.

29 Wolf SL, Thompson PA, Morris DM, et al. The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. *Neurorehabil Neural Repair*. 2005;19:194–205.

30 Rabadi MH, Rabadi FM. Comparison of the Action Research Arm Test and the Fugl-Meyer Assessment as measures of upper-extremity motor weakness after stroke. *Arch Phys Med Rehabil*. 2006;87:962–966.

31 Barreca SR, Stratford PW, Masters LM, et al. Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. *Phys Ther*. 2006;86:245–253.

32 Barreca SR, Stratford PW, Lambert CL, et al. Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. *Arch Phys Med Rehabil*. 2005;86:1616–1622.

33 van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. *J Rehabil Med*. 2001;33:110–113.

34 Gliner JA, Morgan GA, Harmon RJ. Measurement reliability. *J Am Acad Child Adolesc Psychiatry*. 2001;40:486–488.

35 Hobart J. Rating scales for neurologists. *J Neurol Neurosurg Psychiatry*. 2003;74(suppl 4):iv22–iv26.

36 Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability? *Int Disabil Stud*. 1988;10:64–67.

37 Hsueh IP, Lin JH, Jeng JS, Hsieh CL. Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. *J Neurol Neurosurg Psychiatry*. 2002;73:188–190.

38 Hsueh IP, Lee MM, Hsieh CL. Psychometric characteristics of the Barthel activities of daily living index in stroke patients. *J Formos Med Assoc*. 2001;100:526–532.

39 Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. *Qual Life Res*. 1997;6:561–571.

40 Colton T. *Statistics in Medicine*. Boston, MA: Little, Brown & Co; 1974.

41 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

42 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.

43 Bushnell CD, Johnston DC, Goldstein LB. Retrospective assessment of initial stroke severity: comparison of the NIH Stroke Scale and the Canadian Neurological Scale. *Stroke*. 2001;32:656–660.

44 Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. *J Rheumatol*. 1993;20:561–565.

45 Smidt N, van der Windt DA, Assendelft WJ, et al. Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis. *Arch Phys Med Rehabil*. 2002;83:1145–1150.

46 Filiatrault J, Arsenault AB, Dutil E, Bourbonnais D. Motor function and activities of daily living assessments: a study of three tests for persons with hemiplegia. *Am J Occup Ther*. 1991;45:806–810.

47 Shelton FD, Volpe BT, Reding M. Motor impairment as a predictor of functional recovery and guide to rehabilitation treatment after stroke. *Neurorehabil Neural Repair*. 2001;15:229–237.

48 Chae J, Johnston M, Kim H, Zorowitz R. Admission motor impairment as a predictor of physical disability after stroke rehabilitation. *Am J Phys Med Rehabil*. 1995;74:218–223.

49 Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987;40:171–178.

50 Van der Lee JH, De Groot V, Beckerman H, et al. The intra- and interrater reliability of the Action Research Arm test: a practical test of upper extremity function in patients with stroke. *Arch Phys Med Rehabil*. 2001;82:14–19.

51 Lin JH, Hsueh IP, Sheu CF, Hsieh CL. Psychometric properties of the sensory scale of the Fugl-Meyer Assessment in stroke patients. *Clin Rehabil*. 2004;18:391–397.

52 Page SJ, Sisto SA, Levine P. Modified constraint-induced therapy in chronic stroke. *Am J Phys Med Rehabil*. 2002;81:870–875.

Appendix.

Summary of Key Features and Detailed Items of the Upper-Extremity Subscale of the Fugl-Meyer Motor Test (UE-FM), Upper-Extremity Subscale of the Stroke Rehabilitation Assessment of Movement (UE-STREAM), Action Research Arm Test (ARAT), and Wolf Motor Function Test (WMFT)[a]

Parameter	UE-FM	UE-STREAM	ARAT	WMFT
No. of items	33	10	19	15
Scale	Ordinal 3-point	Ordinal 3-point	Ordinal 4-point	Ordinal 6-point
Score range	0–66	0–20	0–57	0–75
Time required to administer (min)	12–15	5–8	8–10	10–12
Measure	Impairment	Impairment	Functional ability	Functional ability
			Grasp	
	1. Shoulder retraction	1. Scapular protraction (supine)	1. Block, wood, 10-cm cube (if score=3, total=18, go to "Grip")	1. Forearm to table (side): participant attempts to place forearm on table by abduction at shoulder.
	2. Shoulder elevation	2. Scapular elevation (sitting)	2. Pick up 10-cm block of wood or 2.5-cm cube (if score=0, total=0, go to "Grip")	2. Forearm to box (side): participant attempts to place forearm on box by abduction at shoulder.
	3. Shoulder abduction	3. Raising arm to highest elevation (sitting)	3. Pick up 2.5-cm block of wood or 5-cm cube	3. Extend elbow (side): participant attempts to reach across table by extending elbow (to side).
	4. Shoulder abduction to 90°	4. Raising hand to touch top of head (sitting)	4. Block, wood, 7.5-cm cube	4. Extend elbow (to side), with weight: participant attempts to push sandbag against outer wrist joint across table by extending elbow.
	5. Shoulder adduction/internal rotation	5. Elbow extension (supine)	5. Ball (cricket), 7.5-cm diameter	5. Hand to table (front): participant attempts to place involved hand on table.
	6. Shoulder external rotation	6. Forearm supination/ pronation (elbow at 90°)	6. Stone, 10×2.5×1 cm	6. Hand to box (front): participant attempts to place hand on box.
			Grip	
	7. Shoulder flexion 0°–90°	7. Hand to sacrum (sitting)	1. Pour water from glass to glass (if score=3, total=12, go to "Pinch")	7. Reach and retrieve (front): participant attempts to pull 0.45-kg (1-lb) weight across table by using elbow flexion and cupped wrist.
	8. Shoulder flexion 90°–180°	8. Making a fist (sitting)	2. Tube, 2.25 cm (if score=0, total=0, go to "Pinch")	8. Lift can (front): participant attempts to lift can and bring it close to lips with cylindrical grasp.
	9. Elbow flexion	9. Finger total extension (sitting)	3. Tube, 1×16 cm	9. Lift pencil (front): participant attempts to pick up pencil by using 3-jaw chuck grasp.
	10. Elbow extension	10. Opposition (sitting)	4. Washer (3.5-cm diameter) over bolt	10. Pick up paper clip (front): participant attempts to pick up paper clip by using pincer grasp.
			Pinch	
	11. Forearm supination		1. Ball bearing, 6 mm, third finger and thumb (if score=3, total=18, go to "Gross Movement")	11. Stack checkers (front): participant attempts to stack checkers onto center checker.
	12. Forearm pronation		2. Marble, 1.5 cm, index finger and thumb (if score=0, total=0, go to "Gross Movement")	12. Flip cards (front): participant attempts to flip each card over by using pincer grasp.

(Continued)

Appendix.
Continued

Parameter	UE-FM	UE-STREAM	ARAT	WMFT
	13. Forearm supination/pronation (elbow at 0°)		3. Ball bearing, third finger and thumb	13. Turning key in lock (front): participant turns key fully to left and right using pincer grasp, while maintaining contact.
	14. Forearm supination/pronation (elbow at 90°)		4. Ball bearing, first finger and thumb	14. Fold towel (front): participant grasps towel, folds it lengthwise, and then uses tested hand to fold towel in half again.
	15. Hand to lumbar spine		5. Marble, third finger and thumb	15. Lift basket (standing): participant picks up basket by grasping handles and placing it on bedside table.
	16. Wrist flexion/extension (elbow at 0°)		6. Marble, second finger and thumb	
			Gross Movement	
	17. Wrist flexion/extension (elbow at 90°)		1. Place hand behind head (if score=3, total=9, finish)	
	18. Wrist extension against resistance (elbow at 0°)		2. Place hand on top of head (if score=0, total=0, finish)	
	19. Wrist extension against resistance (elbow at 90°)		3. Hand to mouth	
	20. Wrist circumduction			
	21. Finger flexion			
	22. Finger extension			
	23. Extension of MCP joints, flexion of PIP or DIP joints			
	24. Grasp: adduct thumb			
	25. Grasp: oppose thumb			
	26. Grasp cylinder			
	27. Grasp tennis ball			
	28. Finger-to-nose speed			
	29. Finger-to-nose tremor			
	30. Finger-to-nose dysmetria			
	31. Finger flexion reflex			
	32. Biceps reflex			
	33. Triceps reflex			

[a] MCP=metacarpophalangeal, PIP=proximal interphalangeal, DIP=distal interphalangeal.