J Rehabil Med 2019; 51: 652–659

ORIGINAL REPORT

INTRA- AND INTER-RATER RELIABILITY OF FUGL-MEYER ASSESSMENT OF UPPER EXTREMITY IN STROKE

Edgar D. HERNÁNDEZ, PT, MSc[1], Claudia P. GALEANO, PT[2], Nubia E. BARBOSA, PT[2], Sandra M. FORERO, PT[2], Åsa NORDIN, PT, MSc[3], Katharina S. SUNNERHAGEN, MD, PhD[3] and Margit ALT MURPHY, PT, PhD[3]
From the [1]Universidad Nacional de Colombia, Departamento del Movimiento Corporal Humano, [2]Central Military Hospital of Colombia, Bogota, Colombia, [3]Institute of Neuroscience and Physiology, Clinical Neuroscience, Rehabilitation Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Objective: **The Fugl-Meyer Assessment of Upper Extremity (FMA-UE) is recommended for evaluation of sensorimotor impairment post stroke, but the item-level reliability of the scale is unknown. This study aims to determine intra- and inter-rater reliability of the FMA-UE at item-, subscale- and total score level in patients with early subacute stroke.**
Design: **Intra/inter-rater reliability.**
Subjects: **Sixty consecutively included patients with stroke (mean age 65.9 years) admitted to Central Military Hospital of Colombia, Bogota.**
Methods: **Two physiotherapists scored FMA-UE independently on 2 consecutive days within 10 days post stroke. A rank-based statistical method for paired ordinal data was used to assess the level of agreement, systematic and random disagreements.**
Results: **Systematic disagreements either in position or concentration were detected in 4 items of the shoulder section. The item level intra- and inter-rater agreement was high (79–100%). The 70% agreement was also reached for the subscales and the total score when 1–3-point difference was accepted.**
Conclusion: **The FMA-UE is reliable both within and between raters in patients with stroke in the early subacute phase. A wider international use of FMA-UE will allow comparison of stroke recovery between regions and countries and thereby potentially improve the quality of care and rehabilitation in persons with stroke worldwide.**

Key words: reliability; psychometrics; non-parametric statistics; stroke rehabilitation; upper extremity; motor activity; physical therapy specialty; reproducibility of results.

Accepted Aug 9, 2019; Epub ahead of print Aug 26, 2019

J Rehabil Med 2019; 51: 652–659

Correspondence address: Margit Alt Murphy, Per Dubbsgatan 14, plan 3, 413 45 Gothenburg, Sweden. E-mail: margit.alt-murphy@neuro.gu.se

LAY ABSTRACT

The Fugl-Meyer Assessment of Upper Extremity (FMA-UE) is one of the most used and recommended assessment scales of sensorimotor function in stroke. This study investigated the reliability of the scale when different therapists assessed the patient's performance at the same test session and when the assessment was performed by the same therapist but on 2 different occasions. Sixty individuals with stroke at the Central Military Hospital of Colombia were included. The results showed that the agreement in each individual movement (FMA-UE includes 33 movements/items) was 79% or above. Disagreements in scorings between raters were noted for 4 single items. These disagreements were probably caused by the spontaneous recovery that occurred in the early subacute phase after stroke. The item, subscale and total score level reliabilities were high and the scale can be recommended for use in general, including in Spanish-speaking countries. It is important, however, that standardized testing procedures are followed.

Hemiparesis is one of the most frequent sequelae of stroke, causing significant disability (1, 2). Motor deficits of the upper limb are common and affect approximately 50–70% of patients admitted to the hospital in the acute, subacute phase (3–5) and 40% in the chronic phase (6, 7). Impaired function of the upper limb makes it difficult to carry out basic movements and daily tasks in an efficient way (8). Upper limb sensorimotor impairment after stroke is commonly assessed by using the Fugl-Meyer Assessment for Upper Extremity (FMA-UE). It is considered as gold standard and is the only impairment level measure recommended for stroke trials (9). The FMA-UE is well-established internationally, clinically feasible and shows excellent reliability, validity and responsiveness (10, 11). FMA-UE is widely used to determine the severity of stroke and to quantify recovery (12).

Both the intra- and inter-rater reliability of the FMA-UE, by means of the intraclass correlation coefficient (ICC), have demonstrated to be excellent, with reported values above 0.90, both for the total and subscale level in the chronic and subacute phase (13–17). The ICC and other correlation methods are valid for measuring the strength of association, but are limited for evaluation of agreement between assessments. Clinical scales, such as the FMA-UE, produce ordinal data, in which the ordered categories represent only rank order and not a numerical value (18). Previous reliability studies of FMA-UE have predominantly used statistical analyses appropriate to continuous data rather than non-parametric statistical methods and evaluated the reliability of the FMA-UE

doi: 10.2340/16501977-2590

summed scores in relatively small samples (generally ≤30). A recent study reported, however, weighted kappa values of ≥0.7 for inter-rater reliability of the individual item scores when FMA-UE scored from a video were compared with direct observation in chronic stroke (19). The item-level reliability evaluation is essential, since single items of FMA-UE have been proposed for prediction of motor recovery after stroke (20). The FMA-UE has been translated to several languages and recently into Colombian Spanish (21) following the protocol and manual according to the original English/Swedish version (22). The current study used the translated Colombian Spanish FMA-UE for reliability evaluation. Thus, the aim of this study was to evaluate the intra- and inter-rater reliability of the FMA-UE at item, subscale and total score level in people with early subacute stroke.

METHODS

Population

In total, 60 individuals with stroke were consecutively included during a 17-month period (Fig. 1). The inclusion criteria were: first-ever stroke, admitted to the Central Military Hospital of Colombia 4–9 days post-stroke, National Institutes of Health Stroke Scale (NIHSS) greater than 0 at admission, and age between 18 and 90 years. Exclusion criteria were: other disorders, such as blindness, deafness, amputation of lower or upper limb, cerebellar stroke. Patients who could not cooperate in FMA testing due to impaired cognition or severe medical condition were also excluded. Ethical approval was received from the ethics committee of the Central Military Hospital (Act number 9, 12 June 2013) and a signed informed consent was obtained from all participants or their family member. Data collection was carried out between November 2014 and April 2016. The Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines (23) and the checklist for reliability evaluation from the consensus-based standards for selection of health status measurement instruments (COSMIN) were followed to ensure the methodological quality of the study (24).

The sample size was based on preliminary results from the pilot study with 10 individuals with stroke (21) and previous studies using the rank invariant method for reliability testing at the item level (25). For the planned study design 60 individuals with stroke were considered to be sufficient.

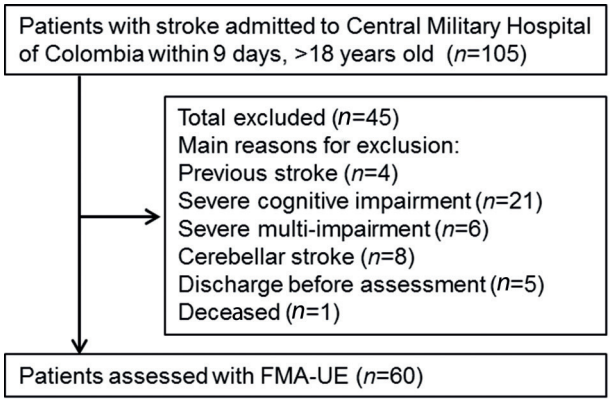

Fig 1. Flowchart for study inclusion.

Fugl-Meyer Assessment of upper extremity

The FMA-UE examines reflex activity, voluntary movements within, partially out and independent of synergies (22). The scale includes 33 items divided into 4 subscales: shoulder/elbow (A, 18 items), wrist (B, 5 items), hand (C, 7 items) and coordination/speed (D, 3 items). Each item is scored on an ordinal 3-point scale, where 2 points are assigned when the movement is performed fully, 1 point when performed partially, and 0 points when the movement cannot be performed. A total score of 66 indicates better sensorimotor function.

Three trained physiotherapists (raters A, B and C) with more than 20 years of clinical experience were randomly assigned into pairs to perform assessments. For practical reasons a fourth rater (also trained and experienced) was involved in assessments of 4 patients. These assessments were not included in the intra-rater analysis. All raters were involved in the translation process of the FMA-UE to Spanish, which also included joint practical training with guidance of experts and data collection for a previous pilot study (21). The patient's performance on the FMA-UE was simultaneously, but independently, scored by one pair of raters on 2 consecutive days. The first assessment was performed between 4 and 9 days post-stroke. During the first assessment one of the raters was acting as test leader (i.e. instructing the patient and scoring) and the other as observer (scoring by observing). These roles were assigned randomly and switched on the second assessment day. The raters did not communicate during the testing session or afterwards regarding the scoring. The scoring protocols of different colours were used for different days, and the completed protocols were stored in sealed envelopes, which were opened at the time of statistical analysis.

Other clinical assessments

The initial severity of the stroke was evaluated using the NIHSS at hospital admission (26–28). The minimum score of 0 indicates no impairment, and the maximum score of 42 indicates severe impairment. Stroke severity was classified as mild (0–4), moderate (5–15), or severe (16–24), or very severe (≥25) (26). The disability level was assessed by using the Modified Rankin Scale (0–6) at discharge, on which a lower score indicates less disability (29).

Statistical analysis

Descriptive statistics were calculated for the background data. The floor and ceiling effect for the FMA-UE was defined as present in the patient cohort when more than 15% of patients received the lowest or highest score of the scale (30).

For the intra- and inter-rater reliability, a rank invariant method especially designed for analysis of disagreements in paired ordinal data was used (18, 31, 32) (the software is available at http://avdic.se/svenssonsmetod.html). The systematic disagreement between raters was expressed as relative position (RP), relative concentration (RC) and relative rank variation (RV) (18). RP indicates the extent to which the distribution of scores from an assessment is systematically shifted towards higher or lower categories. RC shows whether the scores are more or less concentrated towards the central categories of the scale compared with the other assessment. RP and RC values can vary from –1 to 1, where 0 means no difference between raters.

Values within –0.1 and 0.1 were considered negligibly small with reference to clinical relevance, while values outside this range were considered as clinically relevant disagreements (33). The RV indicates disagreement caused by individual variability and varies between 0 and 1 and a value <0.1 means that the difference is negligible. Statistically significant disagreement of RP, RC and RV was indicated with a 95% confidence interval (95% CI) that did not include the value zero. Scatterplot and relative operating curve (ROC) were used to visually analyse the systematic disagreements. The degree of agreement was determined by using the percentage of agreement (PA). Agreement ≥70% was considered satisfactory. For the summed scores (subscale and total scores), a minimum disagreement in points to reach at least 70% PA was also calculated.

RESULTS

In total 105 patients were screened, of whom 60 (48% women, mean age 65.9 years) met the inclusion criteria and were assessed with the FMA-UE (Table I). The main reason for exclusion was severe cognitive impairment that hindered cooperation during the assessment ($n=21$) (Fig. 1). Among the included patients, 93% had ischaemic stroke and 7% haemorrhagic stroke. The FMA-UE scores of the entire group ranged from 4 to 66 points. Out of 60 patients 25% scored ≤48 and 25% ≥65. There was no floor effect observed, since all patients received some points on the first occasion. However, 13 patients (21.7%) received a full score of 66 points on the first occasion, which indicates a ceiling effect.

Table I. Demographic and clinical characteristics ($n=60$)

Characteristics	
Age, years, mean (SD)	65.9 (17.3)
Sex, male/female, %	52/48
Ischaemic/haemorrhagic stroke, %	93/7
Right/left hemiparesis, %	55/45
Thrombolysis, *n*	8
Hospitalization, days, mean (SD)	12 (10)
Modified Rankin Scale, median (Q1–Q3)	2 (1–4)
0 Without symptoms	3
1 Without significant disability	22
2 Mild disability	10
3 Moderate disability	5
4 Moderately severe disability	16
5 Severe disability	4
NIHSS Scale, median (Q1–Q3)	5 (3–10)
Mild 0–5	25
Moderate 6–14	20
Severe 15–24	2
Very severe ≥25	0
Patients without NIHSS scorings	13
Discharged from hospital, *n*	
Home	56
Homecare	1
Intermediate care	1
Died in hospital	2
Fugl Meyer Assessment of upper extremity	
FMA-UE, 1st occasion, median (Q1–Q3)	58 (48–65)
FMA-UE, 2nd occasion, median (Q1–Q3)	59.5 (45–66)

FMA-UE: Fugl-Meyer Assessment Upper Extremity; SD: standard deviation; NIHSS: National Institutes of Health Stroke Scale.

Intra-rater reliability

At the item level, statistically significant systematic disagreement of relative position (RP) was noted for shoulder flexion 0–90° (A.III.) and normal reflex activity (A.V., Table II). All these disagreements were positive, which indicate that a higher category was systematically more frequently used for these items on the second occasion. A negative RC value was noted for one of the raters for elbow extension and forearm pronation within extensor synergy, which means that a more central scoring was more often used on the first occasion compared with the second within the same rater. This disagreement showed the same tendency, as seen in RP values, indicating that a higher score was more frequently used on the second test occasion compared with the first for these items. A shift towards higher score was also seen in the total score A–D. Individual disagreements, measured as RV, were all close to zero across all raters. Scatterplots showing paired intra-rater and inter-rater assessments of the total score A–D along with ROC are presented in Fig. 2. A curved ROC indicates disagreement in position and an S-shaped curve indicates that the raters concentrate their assessments differently on the scale categories. Exact RP and RC values along with 95% CI are displayed in Tables SI–II[1].

The PA between test occasion 1 and 2 within each rater was above 79% for all tested items (Tables II and III). For the reflex activity (A.I.), full agreement was reached. Full agreement at least in one rater was also noted for following items: hand to lumbar spine, mass flexion and extension of the hand, cylinder and spherical grasp. The PA was, as expected, lower for the subscale A (48–59%), B, C and D (63–89%), and for the total score A–D (33–46%), than for single items, since the sum-scores include larger number of categories. A 70% PA was reached for subscale B, C and D when a 1-point difference between test occasions was accepted. Two- and 3-point difference was needed to reach 70% PA in all 3 raters for the subscale A and the total score A–D, respectively.

Inter-rater reliability

A statistically significant systematic disagreement in RC was noted for the forearm pronation (A.II.), which means that the rater with a role of leader was systematically using a more central score compared with the rater who acted as observer (Table II). All other observed systematic disagreements were negligible or not statistically significant. Individual disagreements,

[1]https://doi.org/10.2340/16501977-2590

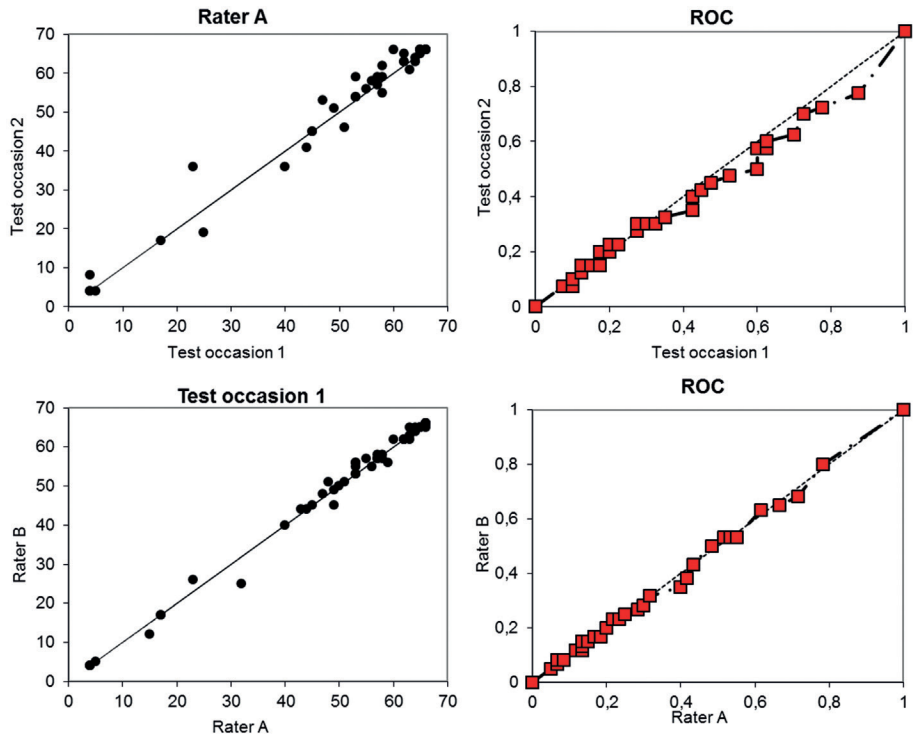

Fig. 2. Scatterplots and relative operating curves (ROC) showing intra- and inter-rater reliability of Fugl-Meyer Assessment of Upper Extremity (FMA-UE) total score (0–66). A ROC curve under the reference line indicates that a higher score was more likely used by the rater on the second test occasion.

Table II. Intra-rater agreement within each rater (A, B and C) and inter-rater agreement between all raters during test occasions 1 and 2 for Fugl-Meyer Assessment Upper Extremity (FMA-UE) subscale A items and sums

	Intra-rater agreement (PA %)			Inter-rater agreement (PA %)	
	Rater A	Rater B	Rater C	Test occasion 1	Test occasion 2
A. UPPER EXTREMITY					
I. Reflex activity					
Flexors and extensors	100	100	100	100	100
II. Within synergies					
Shoulder retraction	85	94	83	92	92
Shoulder elevation	87	94	89	93	98
Shoulder abduction	85	94	92	97	90
Shoulder external rotation	87	94	94	95	95
Elbow flexion	95	97	97	92	93
Forearm supination	79	82	89	90	88
Adduction/internal rotation	95	85	89	95	100
Elbow extension	97	**85 (RC)**[a]	94	93	100
Forearm pronation	95	**82 (RC)**[a]	89	**90 (RC)**[a]	97
SUM A II, range 0–18p	63	66	64	80	82
SUM A II, 1 point difference	85	74	76	90	90
III. Mixed synergies					
Hand to lumbar spine	97	100	92	95	98
Shoulder flexion 0–90°	**82 (RP)**[a]	85	97	95	97
Pronation–supination	87	97	92	97	97
SUM A III, range 0–6p	77	82	86	90	95
IV. Little or no synergies					
Shoulder abduction 0–90°	87	88	89	93	93
Shoulder flexion 90–180°	82	88	92	90	92
Pronation–supination	87	79	92	97	93
SUM A IV, range 0–6	**72 (RC)**[a]	68	89	83	88
V. Normal reflex activity					
Biceps, triceps, finger flexors	**90 (RP)**[a]	82	89	97	98
SUM A, range 0–36p	48	55	59	68	78
SUM A. 1-point difference	73	63	71	88	82
SUM A. 2-points difference	75	71	79	96	96

[a]Statistically significant disagreement (absolute value of RP/RC≥ 0.1 and 95% CI does not include 0) marked in bold.
PA: percentage of agreement; RP: relative position; RC: relative concentration.

Table III. Intra-rater agreement within each rater (A,B and C) and inter-rater agreement between all raters during test occasion 1 and 2 for Fugl-Meyer Assessment Upper Extremity (FMA-UE) subscale B, C and D items, sums and the total score A–D

	Intra-rater agreement (PA %)			Inter-rater agreement (PA %)	
	Rater A	Rater B	Rater C	Test occasion 1	Test occasion 2
B. WRIST					
Stability at 15° dorsiflexion	90	94	89	98	98
Repeated wrist flexion	92	91	89	97	97
Stability at 15° dorsiflexion	85	88	83	95	95
Repeated wrist flexion	87	88	89	97	97
Circumduction	90	91	97	93	93
SUM B, range 0–10p	66	74	72	89	89
SUM B, 1 point	88	87	87	97	97
C. HAND					
Mass flexion	97	100	100	100	100
Mass extension	97	100	100	100	100
Hook grasp	90	94	89	100	100
Thumb adduction	92	97	94	97	100
Opposition/pincer grasp	85	91	92	93	100
Cylinder grasp	92	100	94	97	98
Spherical grasp	92	100	97	98	100
SUM C, range 0–14p	75	89	74	93	98
SUM C, 1 point	95	97	92	98	100
D. COORDINATION/SPEED					
Tremor	87	94	92	93	93
Dysmetria	82	88	83	90	92
Time	85	88	97	98	98
SUM D, range 0–6p	67	71	78	85	85
TOTAL A–D, range 0–66p	**33 (RC)**[a]	45	46	67	75
TOTAL A–D, 1 point difference	60	61	61	80	83
TOTAL A–D, 2 points difference	70	66	68	87	88
TOTAL A–D, 3 points difference	78	79	82	97	93

[a]Statistically significant disagreement (absolute value of RP/RC≥ 0.1 and 95% CI does not include 0) marked in bold.
PA: percentage of agreement, RP: relative position; RC: relative concentration.

measured as RV, were all close to zero. Exact RP and RC values along with 95% CI are displayed in Tables SIII–IV[1].

The PA was above 90% for all items between the raters (Tables II and III). Full agreement (100%) was observed for reflex activity (A.I.), adduction/internal rotation and elbow extension (A.II), stability of wrist (B), mass extension and flexion of the hand, hook grasp, thumb adduction, pincer grasp and spherical grasp (C) at least in 1 of the test occasions. At the subscale and total score level the PA varied between 67% and 93% on the first test occasion and between 75% and 98% on the second test occasion, which indicates improved agreement on the second test occasion. An 80% PA was reached for the subscale A and total score A–D when a 1-point difference between test occasions was accepted.

DISCUSSION

This study demonstrated that the FMA-UE is a reliable clinical instrument for the evaluation of upper extremity motor function early after stroke. Only one item (forearm pronation within synergies) in the inter-rater reliability and 4 items (elbow extension, forearm pronation within synergies, shoulder flexion to 90°, normal reflex activity) out of 33 in the intra-rater reliability testing showed statistically significant systematic disagreements, either in relative position or in concentration. A systematic shift towards higher scores on the second test occasion within the same rater was observed for some items and for the total score. In addition, the intra- and inter-rater agreement was high (79–100%) for all single items, which confirms that the use of single items from the FMA-UE might be warranted. The 70% intra-rater agreement was also reached for the subscale C, but a 1-point difference was needed for the subscale B and D and a 3-point for the subscale A and the total score A–D. Inter-rater agreement was above 80% for subscales B, C and D, and only 1-point difference was needed for subscale A and the total score A–D to reach this level of agreement.

This study is the first to investigate the item-level intra- and inter-rater reliability of the FMA-UE in a relatively large sample of patients early after stroke. Previous studies have to a large extent evaluated reliability in relatively small samples and used statistical methods, such as ICC, which are less suitable for ordinal data (33). However, a recent study used weighted kappa statistics and reported high item-level reliability when the scorings of the FMA-UE were made from the video (19). Weighted kappa is a commonly used measure of agreement, but it still fails to identify the systematic disagreements and ignores the rank invariant properties of ordinal data. It also assumes that the raters have equal skill level, which

means that systematic disagreements are ignored (33, 34). In addition, the weighted kappa value depends on the choice of weights and is sensitive to the number of categories, which means that the value increases when the number of categories decreases (33).

The current study used a statistical method particularly designed for ordinal data. This method makes it possible to measure type and extent of observed inter- and intra-rater disagreements in terms of systematic and non-systematic disagreements (32). The disagreements caused by random individual variability between- and within-raters were negligibly small and non-significant in the current study. This means that the extent of uncertainty in interpretation of the FMA-UE scale categories among raters was small. This small individual variability or low level of noise can be considered as a sign of good quality of the FMA-UE scale properties and/or that the heterogeneity among raters was small. In this study all raters had several years' clinical experience and underwent formal training on FMA assessment organized by the experts in the field prior to data collection (21). All raters were also taking part of the translation process of the scale, which might have positively influenced the results (21). The fact that each item of the FMA-UE is scored on only 3 levels can also increase the possibility to reach high reliability, which can be considered as a strength of the scale. The more approximate scoring might consecutively reduce the sensitivity of the scale to change, but the use of total score will, in turn, reduce this risk.

The systematic disagreements observed were few and occurred mostly in the intra-rater testing performed over 2 consecutive days. These disagreements were predominantly caused by a systematic shift of the scores towards a higher score on the second test occasions. Since a spontaneous recovery can be expected to occur during the first 10 days post stroke, it is likely that the observed disagreements were caused mainly by spontaneous recovery along with a possible learning effect. However, extra attention should be paid in terms of standardization and training for these specific items that showed systematic disagreements (e.g. shoulder flexion 0–90°, elbow extension and pronation within synergies as well as normal reflex activity). This is an example of how a rank invariant method, as used in the current study, can be used as guidance for identifying the problematic items of a scale.

In addition to high agreement observed at the item level, good intra- and inter-rater reliability was seen at the subscale and total score level. For the inter-rater reliability, a 70% agreement of the FMA-UE total score was reached when only one point difference between raters' scores was accepted. Similarly, for the intra-rater testing, a 3-point difference in the FMA-UE total score was needed to reach the 70% agreement within raters. These numbers can be used as guidance for estimating the expected variance in scorings of the total score between several trained and experienced assessors in the early subacute stage after stroke. The suggested minimum clinical difference for the FMA-UE is 3.6 points (35) and minimal clinically important difference is 9–10 points (36) in the subacute stage after stroke. In the current study, good agreement for the FMA-UE total score was reached with lower values for all paired comparisons. This might be accounted for joint training with experts and identification of problem areas during the translation and cultural adaptation process of the scale prior data collection (21).

The current study aimed to include a representative cohort of patients with stroke who would be candidates for sensorimotor assessment using either upper or lower extremity FMA. This, however, resulted in that approximately 22% of the included patients' received maximum score on the FMA-UE. This selection bias might have influenced the study results by reducing the variability of the possible scores in these patients, and making it more likely to achieve a high intra- and inter-rater reliability. This study included patients from the Central Military Hospital, which is free of cost for those serving in the armed forces. Due to the previous civil war, traumatic injuries have dominated, but stroke is increasing. This is the cause of the relatively low recruitment rate of 3.5/month. However, the cohort contains almost half women; the mean age is representative for the middle-income countries and has a variation in sociodemographic background. A possible methodological limitation of the study could be that the time between test occasions 1 and 2 was only one day, which might have caused a recall bias for the raters. To minimize this bias a longer time between assessments could have been used, but then again it would instead increase the likelihood of spontaneous recovery to occur. Here, a video-based assessment could have been a possible solution, although scoring from video is not common in clinical praxis and the results would only be valid for video-based assessments. In the current study the assessments were done by 3 physiotherapists, which is relevant considering clinical praxis. In clinical acute settings it is common that several therapists perform the testing. In these situations continuing training becomes even more important.

Standardized testing protocols and training procedures for FMA-UE are crucial in order to reach sufficient accuracy in scorings and agreement between assessors at different time-points (14, 37). A recent study showed, however, that the FMA-UE was modified in 12 out of 79 studies (11). This heterogeneity limits the ability to pool data from different studies and

synthesize evidence (11). Increased effort is, however, needed to improve uniform assessment of FME-UE across different clinical and research sites across the world. This can be achieved by increased awareness of modifications made to the original scale, but also improved access to the scoring protocols, manuals and training materials used in different countries and languages. The original FMA-UE scale approved by the Axel Fugl-Meyer, the official translations and an instruction video on how to perform the testing are freely available for non-profit use (www.neurophys.gu.se/rehabmed) to any clinical or research setting around the world. All these efforts are important to achieve a more standardized use of the scale.

In conclusion, the FMA-UE showed excellent inter- and intra-rater reliability in the assessment of sensorimotor function in the acute/subacute phase after stroke. Systematic disagreements were detected only in 4 items of the shoulder section. The agreement was excellent at the item level and satisfactory at the subscale and total score level. The findings from the current study, confirming the reliability of the single items of the FMA-UE, might be used as guidance in future studies on stroke recovery. In addition to recommendation of use of the scale in Colombian patient populations, it can be recommended as a reliable clinical assessment tool for use in other clinical and research settings. Wider international use of the FMA-UE has the potential to improve physiotherapists' evaluations of motor impairments in patients with stroke, and to enable comparisons of stroke populations between different countries.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to the Central Military Hospital of Colombia, Universidad Nacional de Colombia and the patients who participated in the study. This publication is in memory of Nancy Stella Landinez Parra who was one of the initiators of the study, and took active part in the study planning and data collection.

Funding: This project was funded by the Central Military Hospital through Research Project Number 2013059, registered with the Research Unit; the grant for strengthening established partnerships 2017 at University of Gothenburg; the Swedish state under the agreement between the Swedish government and the country councils, the ALF-agreement (ALFGBG-775561, ALFGBG-718711); Swedish Research Council (VR2017-00946).

The authors have no conflicts of interest to declare.

REFERENCES

1. Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation. Lancet 2011; 377: 1693–1702.
2. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. Lancet Neurology 2009; 8: 741–754.
3. Nakayama H, Jorgensen HS, Raaschou HO, Olsen TS. Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. Arch Phys Med Rehabil 1994; 75: 394–398.
4. Persson HC, Parziali M, Danielsson A, Sunnerhagen KS. Outcome and upper extremity function within 72 hours after first occasion of stroke in an unselected population at a stroke unit. A part of the SALGOT study. BMC Neurology 2012; 12: 162.
5. Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. Stroke 2001; 32: 1279–1284.
6. 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.
7. Parker VM, Wade DT, Langton Hewer R. Loss of arm function after stroke: measurement, frequency, and recovery. Int Rehabil Med 1986; 8: 69–73.
8. Bernal MYP. [Alterations in motor function of upper limbs in hemiplegia - physiotherapy intervention models.] Movimiento Científico 2009; 3: 101–108. (In Spanish).
9. Kwakkel G, Lannin NA, Borschmann K, English C, Ali M, Churilov L, et al. Standardized measurement of sensorimotor recovery in stroke trials: consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. Int J Stroke 2017; 12: 451–461.
10. Alt Murphy M, Resteghini C, Feys P, Lamers I. An overview of systematic reviews on upper extremity outcome measures after stroke. BMC Neurol 2015; 15: 29.
11. Duncan Millar J, van Wijck F, Pollock A, Ali M. Outcome measures in post-stroke arm rehabilitation trials: do existing measures capture outcomes that are important to stroke survivors, carers, and clinicians? Clin Rehabil 2019; 33: 737–749.
12. Crow JL, Kwakkel G, Bussmann JB, Goos JA, Harmeling-van der Wel BC, Early prediction of functional outcome after stroke I. Are the hierarchical properties of the Fugl-Meyer assessment scale the same in acute stroke and chronic stroke? Phys Ther 2014; 94: 977–986.
13. Lin JH, Hsu MJ, Sheu CF, Wu TS, Lin RT, Chen CH, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys Ther 2009; 89: 840–850.
14. See J, Dodakian L, Chou C, Chan V, McKenzie A, Reinkensmeyer DJ, et al. A standardized approach to the Fugl-Meyer assessment and its implications for clinical trials. Neurorehabil Neural Repair 2013; 27: 732–741.
15. Page SJ, Levine P, Hade E. Psychometric properties and administration of the wrist/hand subscales of the Fugl-Meyer Assessment in minimally impaired upper extremity hemiparesis in stroke. Arch Phys Med Rehabil 2012; 93: 2373–2376 e2375.
16. Michaelsen SM, Rocha AS, Knabben RJ, Rodrigues LP, Fernandes CG. Translation, adaptation and inter-rater reliability of the administration manual for the Fugl-Meyer assessment. Rev Bras Fisioter 2011; 15: 80–88.
17. Lundquist CB, Maribo T. The Fugl-Meyer assessment of the upper extremity: reliability, responsiveness and validity of the Danish version. Disabil Rehabil 2017; 39: 934–939.
18. Svensson E, Schillberg B, Kling AM, Nystrom B. Reliability of the balanced inventory for spinal disorders, a questionnaire for evaluation of outcomes in patients with various spinal disorders. J Spinal Disord Tech 2012; 25: 196–204.
19. Amano S, Umeji A, Uchita A, Hashimoto Y, Takebayashi T, Kanata Y, et al. Reliability of remote evaluation for the Fugl-Meyer assessment and the action research arm test in hemiparetic patients after stroke. Top Stroke Rehabil 2018; 25: 432–437.
20. Nijland RH, van Wegen EE, Harmeling-van der Wel BC,

Kwakkel G, Investigators E. Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: early prediction of functional outcome after stroke: the EPOS cohort study. Stroke 2010; 41: 745–750.
21. Barbosa NE, Forero SM, Galeano CP, Hernandez ED, Landinez NS, Sunnerhagen KS, et al. Translation and cultural validation of clinical observational scales – the Fugl-Meyer assessment for post stroke sensorimotor function in Colombian Spanish. Disabil Rehabil 2018; 41: 2317–2323.
22. Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med 1975; 7: 13–31.
23. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007; 147: W163–194.
24. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010; 10: 22.
25. Nordin A, Alt Murphy M, Danielsson A. Intra-rater and inter-rater reliability at the item level of the Action Research Arm Test for patients with stroke. J Rehabil Med 2014; 46: 738–745.
26. Dancer S, Brown AJ, Yanase LR. National Institutes of Health Stroke Scale reliable and valid in plain English. J Neurosci Nurs 2009; 41: 2–5.
27. Dancer S, Brown AJ, Yanase LR. National Institutes of Health Stroke Scale in plain English is reliable for novice nurse users with minimal training. J Emerg Nurs 2017; 43: 221–227.
28. Brott T, Adams HP, Jr., Olinger CP, Marler JR, Barsan WG, Biller J, et al. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 1989; 20: 864–870.
29. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1988; 19: 604–607.
30. Ashford S, Slade M, Malaprade F, Turner-Stokes L. Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review. J Rehabil Med 2008; 40: 787–795.
31. Avdic A, Svensson E. Svenssons method (Version 1.1) Örebro 2010. Interactive software supporting Svenssons method. [Cited 2018 Nov 26] availible from: http://avdic.se/svenssonsmetod.html 2015-08-14.
32. Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. Stat Med 1994; 13: 2437–2453.
33. Svensson E. Different ranking approaches defining association and agreement measures of paired ordinal data. Stat Med 2012; 31: 3104–3117.
34. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001; 33: 47–48.
35. 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.
36. Arya KN, Verma R, Garg RK. Estimating the minimal clinically important difference of an upper extremity recovery measure in subacute stroke patients. Top Stroke Rehabil 2011; 18 Suppl 1: 599–610.
37. Sullivan KJ, Tilson JK, Cen SY, Rose DK, Hershberg J, Correa A, et al. Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. Stroke 2011; 42: 427–432.