Umeå University

This is an accepted version of a paper published in *Disability and Rehabilitation*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Johansson, G., Häger, C. (2012)

"Measurement properties of the motor evaluation scale for upper extremity in stroke patients (MESUPES)"

Disability and Rehabilitation, 34(4): 288-294

URL: <http://dx.doi.org/10.3109/09638288.2011.606343>

Access to the published version may require subscription.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-52025>

<http://diva-portal.org>

RESEARCH PAPER

Measurement Properties of the Motor Evaluation Scale for Upper Extremity in Stroke Patients

GUDRUN JOHANSSON¹ & CHARLOTTE HÄGER¹

¹Department of Community Medicine and Rehabilitation; Physiotherapy, Umeå University, Umeå, Sweden

Correspondence: C Häger, Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Umeå, SE-901 87 Umeå, Sweden.

Tel: +4690786 9275. Fax: +4690786 9267.

E-mail: charlotte.hager@physiother.umu.se

Keywords

Stroke; motor evaluation scale for upper extremity in stroke; measures; arms; movement

Abstract

Purpose. To investigate interrater reliability of the Motor Evaluation Scale for Upper Extremity in Stroke patients (MESUPES), to provide estimates of the minimal detectable change (MDC) of the MESUPES, and to investigate concurrent validity in relation to the arm scores of the Modified Motor Assessment Scale (M MAS).

Methods. Forty-two stroke patients (mean age 56 ± 12 years) were independently assessed within a 48 hours window by two raters in different pairs (total raters = 4).

Results. Weighted kappa analysis indicated good to very good agreement at item level (range 0.63-0.96). The relative and absolute reliability of the total score of MESUPES (maximum 58) was high according to the intraclass correlation coefficients (ICC=0.98) and the standard error of measurement (SEM=2.68). The MDC for three levels of confidence was calculated: A score change of 8, 7 and 5 is necessary for a MDC to have confidence of 95%, 90% and 80% respectively of a genuine change. Correlation between the MESUPES and M MAS was high ($r_s=0.87$).

Conclusions. The MESUPES shows high interrater reliability, and our study provides useful estimates of MDC for different levels of certainty. Additional research to confirm concurrent validity and to examine other psychometric properties of the MESUPES such as sensitivity is needed.

Introduction

Limitations in arm and hand function are a major problem after stroke and cause difficulties in patients' daily lives [1]. Recent research has demonstrated that the adult central nervous system retains a much higher capacity for plasticity and reorganization than earlier believed [2]. Therefore, an important goal of stroke rehabilitation is to substantially increase the functional use of the affected arm while minimizing compensatory strategies and avoiding learned disuse [3]. In order to evaluate rehabilitative interventions, therapists need accurate and reliable measurements which also capture the movement quality. Although achieving good quality of movement is proclaimed as an important goal in modern stroke rehabilitation there are few appropriate measures to evaluate this in the upper extremity [4-6]. The existing measurements usually have a quantitative approach (for example measures time to perform a movement, the number of possible repetitions of a movement, and active range of motion) without identifying whether the movement is similar to premorbid movements or whether the action occurs through compensatory movements. In addition, there is a lack of standardization of passive range of motion and movement quality of active range especially when there is a low level of arm function [7].

Consensus building has been considered essential for the development of reliable measures of quality of movement [5]. This approach has been used in the development of the new assessment tool, the Motor Evaluation Scale of Upper Extremity in Stroke Patients (MESUPES) [8-10]. This is a scale that purportedly measures quality of movement of the hemiparetic upper extremity and it was translated by Van de Winckel et al from its' original Italian version into English in collaboration with the original authors [6]. The MESUPES was then also investigated for internal

construct validity and unidimensionality by using Rasch analysis on 398 patients from Belgium, Germany and Switzerland. Rasch analysis made it possible to transform ordinal data of MESUPES to equal interval measures. The MESUPES was divided into two sub scales: the MESUPES-arm test and the MESUPES-hand test. The MESUPES showed no different item functioning across gender, age, country of residence, side of lesion, and type of stroke or time since onset of the stroke [6]. Further, 56 patients were randomly selected for a study of reliability. According to Van de Winckel et al [6], the MESUPES met the statistical requirement for validity and relative reliability. However, absolute reliability was not addressed in their study despite that it is nowadays recommended to include measures of both relative and absolute reliability when reporting psychometric properties of new scales [11, 12]. For a clinician who wants to interpret changes in a patient's condition, whether this be improvement or deterioration over time, it is more useful to be able to interpret this in light of the reliability of an instrument in actual units (absolute reliability) instead of comparison with correlation coefficients (relative reliability). Such an approach allows clinicians to indentify the smallest amount of change in score in MESUPES that likely reflects a true change in an individual patient rather than variation in measurement. Establishing the psychometric properties of a new instrument such as MESUPES is more validly achieved also if investigated by independent researchers, rather than those behind its construction. Further, when instruments are used in countries with a different language than that of the original scale, it is important to use proper methods of cross-cultural adaptation and translation, in order to guarantee their validity and reliability.

To our knowledge, neither the absolute reliability nor has the concurrent validity of MESUPES have previously been investigated. Establishment of absolute reliability is

necessary in order to provide useful measures of minimal detectable change (MDC) [13]. The MDC represents the limit for the smallest change that indicates a true change in an individual patient, expressed in either actual units or in percent. Concurrent validity is studied when a new measure and a criterion measure (gold standard) are taken at approximately the same time [14]. There is no obvious gold standard for assessment measurement of the upper extremity in stroke. However, among the most common stroke assessment instruments are the Fugl-Meyer Assessment (FMA) [15] and the Motor Assessment Scale (MAS) [16]. Since the MESUPES is not based on synergy stages as FMA, it seemed more pertinent to compare it with the Modified MAS (M MAS) [17], a development of the MAS which has its theoretical background in motor relearning. Furthermore, the arm part of the M MAS is less time-consuming than the FMA and therefore less defatigable for the patients. The latter seemed important as we wanted the gold standard instrument to be taken along with the MESUPES during the first test occasion. In contrast to the MAS, the M MAS excludes the assessment of tone, which has been considered unreliable [18], and includes assessments of both arms. The aims of this study were thus to 1) investigate aspects of interrater reliability of MESUPES, 2) provide useful estimates of MDC for the MESUPES, and 3) examine the concurrent validity of the MESUPES in relation to the M MAS.

Methods

Participants

Adults with a unilateral hemiparesis following an ischemic or hemorrhagic stroke were recruited consecutively from the University Hospital of Northern Sweden either while inpatients or as outpatients. Exclusion criteria based on information obtained from the patient's medical records were other neurological illnesses, orthopaedic or rheumatic impairments or contractures of upper extremity, neglect,

apraxia or inability to co-operate or to understand the instructions. A total of 55 people were eligible for the study. Of these, eight declined to take part in the study and a further five had to be withdrawn as they suffered subsequent medical complications during the study. Accordingly, 42 patients with a mean age of 56 ± 12 years participated. Of these, 64 % were male. Most of the patients were in the chronic phase following stroke, with the median length of time after stroke onset being 7 months (range 3-15.3 months). The scores of the upper extremity part of the M MAS demonstrated mild to severe deficits for the affected arm while the sub score for the non affected arm indicated no clinical deficits. All patients provided scores on the Euroqol 5 Dimensions self report questionnaire [19]. These scores and the patients' other characteristics are shown in Table I. The study was approved by the Regional Ethical Review Board, Umeå, Sweden, with written consent to participate in the study provided by all participants.

- *Insert table I about here* -

Raters

Four physiotherapists with longstanding experience in neurological rehabilitation (mean 13.3 years; SD 6.4; range 8-22 years) acted as raters. They all worked in a rehabilitation centre and had had no previous practice with the MESUPES. Before the initiation of the study, all raters read the manual of the scale and underwent a 2 h training session for practical demonstration of the scale and its' scoring methods. Each therapist also individually practiced assessment with the scale on one or two stroke patients, followed by a consensus meeting where scores were matched and discrepancies discussed. For the study by Van de Winckel [6], the raters were trained for 1 h.

Instrument

The MESUPES assessment tool is an ordinal scale that rewards only qualitatively normal movements in the paretic arm. The scale items reflect the domains of body functions/structures and activity according to the International Classification of Functioning, Disability and Health [20]. The MESUPES includes eight arm items with six response categories (scores 0-5), and nine hand items with three response categories (scores 0-2). The items are briefly described in Table II. The maximum score for the MESUPES is 58. For the arm function (MESUPES-arm test), the tasks are performed in three consecutive phases. In the first phase (scores 0-1), the therapist moves the patient's arm to evaluate the muscle tone. In the second phase (score 2), the patient is asked to participate in the movement and the therapist evaluates the active contribution through normal muscle contractions. In the third phase (scores 3-5), the patient performs the movement independently and the therapist evaluates the range of motion. For each task only one score is given; the highest score that the patient obtains during that test item. For the hand function (MESUPES-hand test), the patient performs the movement actively. Score 0 is given if the performance conveys inadequate adaptation of tone, abnormal muscle contractions, synergic (flexor/extensor) or mass movement patterns. A more comprehensive description of MESUPES has been presented previously [6].

For this study, we translated the MESUPES into Swedish by using specific guidelines for cross-cultural adaptation [21]. Four Swedish versions were first made independently and by English-speaking Swedish persons. The authors of this study and another physiotherapist then merged these versions into one Swedish version, which was then retranslated back to English by a native English-speaking person also fluent in Swedish. The back translation was compared to the original English text, after which a first draft of the Swedish version was outlined. The research group also had access to the

original scale in Italian and one of the authors (CH) possesses the language skill to make a comparison to the original. A pilot study was performed to test the translation. This demonstrated some vagueness in the manual in how to interpret the definition of scores of 0 and 1 for the arm test. We corrected the scale after clarifications from Van de Winckel before a final Swedish version of the MESUPES was completed.

Procedure

Each subject was tested on the MESUPES by two physiotherapists independently. The patients were tested under standardized conditions and in the same environment. The two assessments were conducted within 48 h. For practical reasons, the physiotherapist responsible for the patient's intervention completed one of the assessments and a second assessor was randomly selected from the pool of available raters. Hence, two raters for one patient were not necessarily the same two raters for a different patient. Scoring was performed immediately after the assessment, without any discussion between the raters.

Data analysis

The inter-rater reliability for each item was determined by calculating percentage of agreement, percentage of maximum 1-score difference, and a linear weighted kappa analysis. Linear weighted kappa was chosen because it varies much less with the number of categories [22] and is preferred when comparing reliability between items with different numbers of categories [23]. The strength of agreement was interpreted according to Altman: <0.2 = poor, $0.21-0.40$ = fair, $0.41-0.60$ = moderate, $0.61-0.80$ = good, and $0.81-1.00$ = very good [24].

To evaluate whether there were systematic changes or not between the measurements, we calculated the mean difference between the two test occasions (mean diff), the standard deviation of the differences (SD_{diff}), and the Standard error (SE) of the mean difference. The 95% confidence intervals of d (95% CI) were derived as mean diff $\pm 2.018 \times SE$, whereas 2.018 was obtained from the t-table with 42 (n-1) degrees of freedom. If zero is included in the 95% CI, there is no systematic bias in the data. To evaluate heteroscedasticity, that is, that the variation would differ between higher and lower values, the correlation between the mean of the two test occasions and the absolute difference between test occasion 2 and 1 were tested using Spearman's rank correlation test. An outlier value was considered when the difference between two test occasions was outside 2 standard deviations (SD). Three outliers (>7.58 for total score) were identified but were not excluded from the analysis as to our knowledge there had been nothing that had occurred that would explain the differences during or between the test occasions.

For subscale scores and total score of the MESUPES, the relative reliability was quantified by the $ICC_{1,1}$ as the subjects were not all assessed by the same raters. With $ICC_{1,1}$, the difference among subjects is assessed without separating other sources of variances, including rater or measurement error [14]. A one-way analysis of variance (ANOVA) was used, and the 95% confidence interval was obtained from the ANOVA tables. The strength of correlation was interpreted according to Munro: 0.00-0.25 = little if any, 0.26-0.49 = low, 0.50-0.69 = moderate, 0.70-0.89 = high and > 0.90 = very high correlation [25].

The standard error of measurement (SEM) was calculated by obtaining the square root of the error variance from the ANOVA table [13]. The SEM was divided by the mean from test occasion 1 and 2 and multiplied by 100 to get a percentage value: SEM%. The lower the SEM, the more reliable is the measurement [26]. In order to provide a direct comparison of the effect of calculating with different statistical rigour, three levels of MDC (MDC₉₅, MDC₉₀ and MDC₈₀) were calculated to determine whether the change score of an individual patient is real at the 95%, 90% and 80% confidence level. The MDC₉₅ was calculated as $SEM \times 1.96 \times \sqrt{2}$ and then divided by the mean from test occasion 1 and 2 and multiplied by 100 to get a percentage value: the MDC%. The MDC₉₀ ($SEM \times 1.65 \times \sqrt{2}$), and the MDC₈₀ ($SEM \times 1.28 \times \sqrt{2}$) were calculated in similar manner. Lower MDC values reflect greater responsiveness [27].

Concurrent validity was investigated by computing the two-tailed Spearman's rho between the MESUPES score (mean of two test occasions) and the score for the upper extremity part of the M MAS (taken at the first test occasion), which is considered as a reliable and valid scale for stroke patients [17]. In this study we used three subsets of M MAS, namely upper-arm function, hand movements and advanced hand activities. The total score of the three subsets of the M MAS and the total score of MESUPES were compared in the analysis. We also calculated the correlation between the sub scores of the three subsets of the M MAS and the sub scores of the MESUPES-arm test and MESUPES-hand test. The strength of correlation was interpreted according to Munroe [25]. The SPSS (Statistical Package for Social Sciences, Version 15.0, Inc., Chicago, IL, USA) was used for data analysis.

Results

Interrater reliability

There was good to very good agreement between raters, as shown by linear weighted kappa values ranging from 0.63 to 0.96 for all items of the MESUPES (Table II). Ten of seventeen items reached a kappa value of 0.86 or more. Regarding exact agreement between raters they scored the same patient identically for 32 of 42 (76%) patients in 14 of 17 items. No item had a score difference of more than 2. There were few ratings at score level 2 and even more so at score level 1 for all items across the arm subscale, while about two thirds of the sample scored between 3 or above. The ratings at score levels for hand items were distributed across the entire hand subscale.

The absolute difference of inter-ratings of total score of the MESUPES was 4 or lower for almost 90% of the patients (Table III). The mean absolute difference of the total score was 2.0 (SD=3.3). There was no heteroscedasticity in the data (Spearman 2-tailed $p=0.220$).

- *Insert tables II-III about here* -

Table IV shows that the mean differences between tests of sub scores and total score of the MESUPES was low. The 95% CI for the mean difference included zero, which indicated no systematic biases. The ICCs for the sub scores and the total score were all 0.98, indicating very high correlation.

- *Insert table IV about here* -

As reported in Table V, the SEM for the total score was 2.68, representing sufficient absolute reliability (see discussion). The table also presents the range of MDC if calculating with confidence intervals with different statistical rigor. The MDC for the total score increases from 5 to 8, if choosing a calculation based on MDC₉₅ rather than the MDC₈₀. Note that a value of 3 corresponds to 5% of the total MESUPES score.

- *Insert Table V here* -

Concurrent validity

The correlation was high ($r_s=0.87$) between the total scores of MESUPES and the upper extremity part of the M MAS, indicating high concurrent validity. The relationship between the arm items of the MESUPES and the M MAS, as well as the relationship between the hand items of the MESUPES and the M MAS also showed high correlation values ($r_s=0.84$ and $r_s=0.80$, respectively).

Discussion

The present study shows that the MESUPES has high reliability for assessment of upper extremity in stroke patients, and thus expands the results of the earlier study of the MESUPES [6]. The MDC of the MESUPES, obtained by different raters, varies with the strength of the statistical rigor but provides useful information to rehabilitation staff of the change over time or after intervention. Data suggest that a score change of 8, 7 and 5 is necessary to be 95%, 90% and 80% confident, respectively, of a genuine change of function. High correlation values between the MESUPES and the arm scores of the M MAS were also found, indicating high concurrent validity.

The present study addressed both the absolute and relative reliability of the MESUPES, which has not previously been reported. The SEM is useful for clinicians to estimate how “true” a score is (i.e. beyond measurement errors). The SEM% value of the total score was below 7, which is considered small [28]. To our knowledge, there is no clear definition of acceptable MDC “cut-off” values and in addition, several different terms represent similar calculations [29-31]. For example, the MDC% is defined exactly the same way as the smallest real difference in percentage (SRD %) [30]. Other studies have considered a SRD% less than 30 or 23, respectively, as an acceptable cut-off for some commonly used measurements in stroke patients for timing aspects [11, 28] and a SRD% less than 33 as a suitable cut-off for strength aspects [32]. Hence, we consider the MDC% of 18 for MESUPES as sufficiently low. In analogy with MESUPES, Bergs Balance Scale (BBS) is based on ordinal items summarized to a total score (maximum score 56)[33]. Interestingly, our MDC values are comparable to previous studies of BBS in stroke patients where a score difference of 7 is necessary to be 95% confident of a genuine change [34, 35]. Although the 95% confidence level is the most common level in research, some authors believe that lower levels of confidence might be enough in clinical settings [29, 34, 36]. Hence, we also present the 90% and 80% levels of confidence to provide the clinician with more options when evaluating an individual patient’s function and to pinpoint the difference of increasing statistical rigor. It is worth pointing out that the MDC obtained by the same rater is expected to be lower than the MDC obtained by different raters [37]. Therefore, although intrarater reliability was not investigated in the present study, a score difference lower than 8 would be sufficient to assume a MDC of enough certainty when the same rater repeatedly assesses hand and arm function using the MESUPES. Outliers in the data will nevertheless in either case strongly influence the calculations. In the present study, three outliers were located in

the middle score range of the scale. If excluding the outliers, the MDC_{95} and the $MDC\%$ of the MESUPES should have been considerably lower (4.3 points and 10, respectively). We chose, however, to present and base our calculations from data for the whole study population, also because it is likely that such a situation could reflect the clinical context.

The weighted kappa for the MESUPES arm test varied between 0.80-0.92, which is higher than the findings of Van de Winckel et al (0.62-0.79) [6]. The first two arm items ('Hand to stomach', and 'Hand from stomach to side') had lower kappa values than the rest of the arm items. To further clarify the instructions of these items, we suggest a more specific location of the placement of the hand on the stomach, i.e. the navel. In the study by Van de Winckel et al [6], weighted kappa statistics for the hand items could not be derived as a majority of the patients scored zero. In our study, the scores of the MESUPES hand test were distributed across the entire subscale, and the weighted kappa values indicated good or very good agreement. Some hand items ('Wrist extension', 'Spread fingers 2 and 3', 'Rotate dice fingers 1 and 2', 'Index top + rotate dice') showed lower kappa values than the rest of the hand items.

In the study by Van de Winckel et al [6], the calculations of relative reliability were made with the $ICC_{2,1}$, a model that counts that the subjects are assessed by the same raters. For our study design the $ICC_{1,1}$ seemed most appropriate as the study consisted of a pool of second raters. It is rigorous, in the sense that it usually gives a lower value than $ICC_{2,1}$ [38]. However, for our data both ICC models gave equal results. A wider range of total scores among individual patients was likely one reason for the

higher ICCs in our study (0.98, CI 0.97-0.99) compared to the study by Van de Winckel et al (0.95, CI 0.91-0.97) [6].

Van de Winckel and co-workers also addressed construct validity, which was proven to be sufficient. In our study, despite a lack of perfect gold standard measures, we investigated the concurrent validity in relation to M MAS, which seemed a good choice since the upper extremity part of MAS has previously been used as a standard measure to examine the association between upper extremity instruments [39]. Indeed, the results confirmed high concurrent validity. However, our findings need to be interpreted with some caution as the M MAS is not a perfect gold standard to evaluate quality of movement. In the process of selecting a criterion measure, we excluded time-based instruments like the Wolf Motor Function Test (WMFT)[40] as the MESUPES evaluates quality of movement in a self-selected speed. Further, the FMA and Action Research Arm Test (ARAT) [41] did not seem appropriate as they have other theoretical backgrounds, and thus evaluate movement performance differently than MESUPES [6]. However, since there are several other instruments of the upper extremity in stroke that are categorized in the ICF categories body functions/structures and activities [42], we would welcome more studies to further evaluate the concurrent validity of MESUPES.

Our study design, as well as the one by Van de Winckel et al [6], reflects common clinical scenarios in the sense that patients both in the sub acute and chronic phase, respectively, were included and the assessments were performed by different pairs of raters. To minimize unwanted variability of clinical ratings, it is preferable that the same rater assesses the patient twice [12]. However, in the present study we wanted to compare the results with the Van de Winckel's study [6]. The re-test was made within 48 h to minimize

changes in the individual patient's condition that would introduce patient variability. No systematic changes were found, indicating that there were no learning effects or effects of fatigue. When using the MESUPES, the therapist can repeat the task three times if the patient performs the task inadequately, and the therapist is also allowed to make the patient aware of the abnormal components of the movement. This strategy gives the patient a chance to truly understand and correctly perform the various tasks on test occasions. In addition, the tasks are not physically challenging or cognitively exhausting. Even though the average age of patients in the present study was younger than what would be usual for many patients after stroke, this did not appear to affect the test outcome, as the previous Rasch analysis by Van de Winckel et al has showed that age does not contribute to the response of the items [6]. There are three important differences in our sample compared to that in the study by Van de Winckel et al [6]. Firstly, the sample size was somewhat smaller in our study but yet considered sufficient for establishing clinically important changes for MESUPES [30]. Secondly, there were more patients who showed better function in the upper extremity as, in contrast to the study by Van de Winckel [6], we had no patients in the acute phase and many of our patients had already undergone rehabilitation. Consequently, the more even distribution across the hand items enabled us to calculate weighted kappa of the hand test. Thirdly, the sample in our study had a wider range of stroke severity which contributed to higher ICC on the total score. While two thirds of the sample had moderate to good function in the upper extremity, some patients scored 0 in the arm and hand items.

In the MESUPES arm test, the tone adaptation is evaluated during slow passive movements (score 0-1) before proceeding to assisted active movements. This hierarchical assessing order assumes that a normal tone adaptation is a prerequisite to activate muscles normally, which in turn is a requirement to perform movements

normally. The way of assessing tone in functional tasks according to the MESUPES is an interesting and new approach of assessing tone after stroke compared to other well-established instruments for instance the Ashworth scale [43]. In the MESUPES the performance of the affected arm is compared to the non affected arm or, in presence of an interfering pathology to that side, to what is accepted as “normal movement”. The concept of “normal movement” is defined in the manual of MESUPES as painless, without tremor, executed with a normal range of motion using adequate muscle contraction and a normal orientation of the various body segments [6]. The MESUPES provides valuable information of quality of movement in the affected arm, which is helpful for clinicians when designing individual exercise programs. The scale is possible to use regardless of the degree of hemiparesis since the scale encompasses passive, assisted and active movements. Therefore, we believe that MESUPES is useful for measuring recovery after stroke in adult patients at all stages ranging from the acute phase through to a more stable phase. Further research is nevertheless needed to investigate other aspects of the MESUPES, for instance its sensitivity. It would likewise seem valuable in future publications to present Rasch converted scores in addition to raw scores, since the previous represents an interval scale and the converted score of MDC might be more meaningful to the clinicians.

Conclusion

The MESUPES shows high interrater reliability, and our study provides useful estimates of MDC for different levels of certainty. The MDC of the MESUPES, obtained by different raters, varies with the strength of the statistical rigor but provides useful information to rehabilitation staff of the change over time or after intervention. Data suggest that a score change of 8, 7 and 5 is necessary to be 95%, 90% and 80%

confident, respectively, of a genuine change of function. Although there was a high correlation between MESUPES and M MAS as indicative of concurrent validity, further investigation of validity aspects is encouraged. Additional research to examine other psychometric properties of the MESUPES such as sensitivity is also required to further establish its utility.

Acknowledgements

We wish to thank all participants in the study, and physiotherapists Folke Forzelius, Marianne Sandström and Elisabeth Svanberg for collecting data and Elisabeth Anens and Hilda Mulligan for valuable points of view. Gunilla Janzén, manager of Neurorehab Sävar, is gratefully acknowledged for her support.

Declaration of interest: The authors report no declarations of interest. The authors alone are responsible for the content and writing of the paper. The study was funded by the Swedish Stroke Association, the Memory Foundation of the Swedish Association of Registered Physiotherapists, and the County Council of Västerbotten (FoU).

References

1. Kwakkel G, Kollen B, Wagenaar R. Therapy impact on functional recovery in stroke rehabilitation: A critical review of the literature. *Physiotherapy*. 1999;85:377-91.
2. Shumway-Cook A, Woollacott MH. Motor control : translating research into clinical practice. 3 ed. Baltimore, Md: Lippincott Williams & Wilkins; 2007.

3. Thielman G, Dean C, Gentile A. Rehabilitation of reaching after stroke: task-related training versus progressive resistive exercise. *Arch Phys Med Rehabil.* 2004;85:1613-18.
4. Levin MF, Desrosiers J, Beauchemin D, Bergeron N, Rochette A. Development and validation of a scale for rating motor compensations used for reaching in patients with hemiparesis: the reaching performance scale. *Phys Ther.* 2004;84:8-22.
5. Pomeroy VM, Pramanik A, Sykes L, Richards J, Hill E. Agreement between physiotherapists on quality of movement rated via videotape. *Clin Rehabil.* 2003;17:264-72.
6. Van de Winckel A, Feys H, van der Knaap S, Messerli R, Baronti F, Lehmann R, Van Hemelrijk B, Pante F, Perfetti C, De Weerd W. Can quality of movement be measured? Rasch analysis and inter-rater reliability of the Motor Evaluation Scale for Upper Extremity in Stroke Patients (MESUPES). *Clin Rehabil.* 2006;20:871-84.
7. 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;40:787-95.
8. Dal Pezzo F FB, Rigodanzo D, Rodighiero L SM. Cartella valutativa per il recupero dei movimenti della mano nell'emiplegico. In 'Dall'Osservazione all'esercizio: la pianificazione del trattamento riabilitativo' Collana di Riabilitazione diretta da Carlo Perfetti, Aldo Pieroni . Guido Gnocchi Editore srl, Casa editrice Idelson dal 1908;1995:75-77.
9. Perfetti C. *Der hemiplegische Patient*: Pflaum; 1997.
10. Perfetti C. *L'exercice thérapeutique cognitif pour la rééducation du patient hémiplégique*: Masson; 2001.

11. Chen HM, Chen CC, Hsueh IP, Huang SL, Hsieh CL. Test-retest reproducibility and smallest real difference of 5 hand function tests in patients with stroke. *Neurorehabil Neural Repair*. 2009;23:435-40.
12. Domholdt E. *Rehabilitation research: principles and applications*: Elsevier Saunders; 2005.
13. Stratford PW, Finch E, Solomon P, Binkley J, Gill C, Moreland J. Using the Roland-Morris Questionnaire to make decisions about individual patients. *Physiother Can*. 1996;48:107-18.
14. Portney L, Watkins M. *Foundations of clinical research: applications to practice*: Prentice Hall Upper Saddle River, NJ; 2000.
15. Fugl-Meyer AR JL, Leyman I, Ollson S, Steglind S. The post-stroke hemiplegic patient. A method for evaluation of physical performance. *Scand J Rehabil Med*. 1975;7:13-31.
16. Carr JH, Shepherd RB, Nordholm L, Lynne D. Investigation of a new motor assessment scale for stroke patients. *Phys Ther*. 1985;65:175-80.
17. Barkelius K, Johansson A, Korm K, Lindmark B. Reliabilitets-och validitetsprövning av modifierad motor assessment scale enligt Uppsala Akademiska Sjukhus-95 (Swedish). *Nordisk Fysioterapi*. 1997;1:121-26.
18. Wade DT. *Measurement in neurological rehabilitation*. Oxford; 1992.
19. www.euroqol.org.
20. Geyh S, Cieza A, Schouten J, Dickson H, Frommelt P, Omar Z, Konstanjek N, Ring H, Stucki G. ICF core sets for stroke. *J Rehabil Med*. 2004;36:135-41.
21. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol*. 1993;46:1417-32.

22. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85:257-68.
23. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology.* 1996;7:199-202.
24. Altman DG. *Practical Statistics for Medical Research*: Chapman & Hall;1991.
25. Munro B. *Statistical methods for health care research*: Lippincott Williams & Wilkins; 2004.
26. Atkinson G, M NA. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.* 1998;26:217-38.
27. Wagner JM, Rhodes JA, Patten C. Reproducibility and minimal detectable change of three-dimensional kinematic analysis of reaching tasks in people with hemiparesis after stroke. *Phys Ther.* 2008;88:652-63.
28. Flansbjer UB, Holmback AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med.* 2005;37:75-82.
29. Conradsson M, Lundin-Olsson L, Lindelof N, Littbrand H, Malmqvist L, Gustafson Y, Rosendahl E. Berg Balance Scale: Intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. *Phys Ther.* 2007;87:1155-62.
30. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil.* 2005;84:719-23.
31. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther.* 2006;86:735-43.

32. Flansbjer UB, Holmback AM, Downham D, Lexell J. What change in isokinetic knee muscle strength can be detected in men and women with hemiparesis after stroke? Clin Rehabil. 2005;19:514-22.
33. Berg K, Wood-Dauphinee S, Williams JJ, Gayton D. Measuring balance in the elderly: preliminary development of an instrument. Physiother Can. 1989;41:304-11
34. Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. Aust J Physiother. 2001;47:29-38.
35. Liaw LJ, Hsieh CL, Lo SK, Chen HM, Lee S, Lin JH. The relative and absolute reliability of two balance performance measures in chronic stroke patients. Disabil Rehabil. 2008;30:656-61.
36. Donoghue D, Stokes EK. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. J Rehabil Med. 2009;41:343-46.
37. Lu WS, Wang CH, Lin JH, Sheu CF, Hsieh CL. The minimal detectable change of the simplified stroke rehabilitation assessment of movement measure. Journal of Rehabilitation Medicine. 2008;40:615-9.
38. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil. 1998;12:187-99.
39. 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;27:107-13.
40. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. Stroke. 2001;32:1635-39.
41. Carroll D. A quantitative test of upper extremity function. J Chronic Dis. 1965;18:479-91.

42. Chen SY, Winstein CJ. A systematic review of voluntary arm recovery in hemiparetic stroke: critical predictors for meaningful outcomes using the international classification of functioning, disability, and health. *J Neurol Phys Ther.* 2009;33:2-13.
43. Pandyan AD, Johnson GR, Price CI, Curless RH, Barnes MP, Rodgers H. A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. *Clin Rehabil.* 1999;13:373-83.

Implications for rehabilitation

- This study shows that the Motor Evaluation Scale of Upper Extremity in Stroke patients (MESUPES) has high absolute and relative reliability.
- The MESUPES is suggested to be a useful tool to evaluate quality of movement in the upper extremity of stroke patients.

Table I. Characteristics of the subjects (N = 42)

Age in years, mean $\pm$ SD (range)	56 $\pm$ 12 (31-83)
Gender, male/female, n	27/15
Side of paresis, right/left, n	21/21
Stroke subtype, ischemic/hemorrhagic, n	30/12
Months after stroke, median (IQR)	7 (3.0-15.3)
Handedness, right/left, n	41/1
UE M MAS (maximum score = 15), mean $\pm$ SD (range)	
Affected arm	10.6 $\pm$ 5.5 (0-15)
Non affected arm	14.9 $\pm$ 0.3 (14-15)
EQ-5D	
Mobility: No problems in walking about, n	21
Some problems in walking about, n	21
Self-care: No problems with self-care, n	34
Some problems washing or dressing, n	8
Usual activities:	
No problems with performing, n	20
Some problems with performing, n	19
Unable to perform usual activities, n	3

SD = standard deviation of mean; IQR = interquartile range (1st quartile to 3rd quartile); UE M MAS = upper extremity part of the Modified Motor Assessment Scale; EQ-5D, Euroqol 5 Dimensions self report questionnaire.

Table II. Agreement between raters for each item of the MESUPES for the two test occasions (N = 42)

Item	Exact Agreement n (%)	≤ 1 score difference difference in agreement n (%)	Linear-weighted Kappa (95% CI)
MESUPES arm test			
1. Hand to stomach	31 (74)	39 (93)	0.82 (0.70-0.94)
2. Hand from stomach to side	29 (69)	40 (95)	0.80 (0.69-0.92)
3. Arm abduction	35 (83)	41 (98)	0.89 (0.81-0.98)
4. Arm adduction to side	36 (86)	41 (98)	0.90 (0.82-0.99)
5. Hand from knee to table	32 (76)	41 (98)	0.86 (0.77-0.95)
6. Hand to mouth	38 (90)	40 (95)	0.92 (0.84-1.00)
7. Reach for bottle	33 (79)	41 (98)	0.87 (0.78-0.96)
8. Hand on top of head	35 (83)	42 (100)	0.91 (0.84-0.98)
MESUPES hand test			
1. Pinch grip fingers 1 and 2	41 (98)	41 (100)	0.96 (0.88-1.00)
2. Wrist extension	37 (88)	41 (98)	0.78 (0.58-0.98)
3. Opposition fingers 1 and 5	40 (95)	42 (100)	0.93 (0.83-1.00)
4. Extension 3rd finger	39 (93)	42 (100)	0.91 (0.81-1.00)
5. Spread fingers 2 and 3	31 (74)	41 (98)	0.63 (0.41-0.85)
6. Extension 5th finger*	35 (85)	41 (100)	0.83 (0.71-1.96)
7. Lift little bottle 2 cm	37 (88)	42 (100)	0.86 (0.74-0.98)
8. Rotate dice fingers 1 and 2	32 (76)	42 (100)	0.73 (0.59-0.88)
9. Index top + rotate dice	34 (81)	42 (100)	0.76 (0.60-0.91)

* N = 41 (one patient had an amputation of the 5th finger since childhood)

MESUPES, Motor Evaluation Scale for Upper Extremity in Stroke patients; CI, confidence interval.

Table III. Absolute difference in total scores of the MESUPES between two test occasions

Differences in MESUPES points	Participants, n (N = 42)	Percentage	Cumulative Percentage
0	22	52.4	52.4
1	5	11.9	64.3
2	3	7.1	71.4
3	4	9.5	81.0
4	3	7.1	88.1
6	1	2.4	90.5
7	1	2.4	92.9
11	2*	4.8	97.6
13	1*	2.4	100

* Outliers according to definition in the Method section.

MESUPES, Motor Evaluation Scale for Upper Extremity in Stroke patients.

Table IV. Reliability analysis of the sub scores and total score of the MESUPES (N = 42)

	Mean (SD) Test1	Mean (SD) test 2	Mean diff (95% CI)	ICC (95% CI)
Sub score arm (maximum 40)	29.1 (14.8)	29.3 (14.6)	0.24 (-0.85-1.08)	0.98 (0.96-0.99)
Sub score hand (maximum 18)	12.0 (6.2)	12.3 (6.2)	0.21 (-0.38-0.47)	0.98 (0.96-0.99)
Total score (maximum 58)	41.1 (20.7)	41.6 (20.5)	0.45 (-0.91-1.43)	0.98 (0.97-0.99)

MESUPES, Motor Evaluation Scale for Upper Extremity in Stroke patients; SD, standard deviation; ICC, intraclass correlation coefficients; CI, confidence interval.

Table V. The Minimal Detectable Change of the sub scores and total score of the MESUPES (N = 42)

	SEM	SEM%	MDC ₈₀	MDC ₉₀	MDC ₉₅	MDC%
Sub score arm (maximum 40)	2.20	7.8	3.73	5.12	6.10	21.6
Sub score hand (maximum 18)	0.94	7.7	1.60	2.19	2.61	21.4
Total score (maximum 58)	2.68	6.5	4.55	6.23	7.43	18.0

MESUPES, Motor Evaluation Scale for Upper Extremity in Stroke patients; SEM, standard error of measurement; MDC, minimal detectable change