J Rehab Med 2001; 33: 110–113

THE RESPONSIVENESS OF THE ACTION RESEARCH ARM TEST AND THE FUGL-MEYER ASSESSMENT SCALE IN CHRONIC STROKE PATIENTS

Johanna H. van der Lee,[1,2] Heleen Beckerman,[1,2] Gustaaf J. Lankhorst[1,2] and Lex M. Bouter[2]

From the [1]Department of Rehabilitation Medicine, University Hospital Vrije Universiteit, [2]Institute for Research in Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands

The responsiveness of the Action Research Arm (ARA) test and the upper extremity motor section of the Fugl-Meyer Assessment (FMA) scale were compared in a cohort of 22 chronic stroke patients undergoing intensive forced use treatment aimed at improvement of upper extremity function. The cohort consisted of 13 men and 9 women, median age 58.5 years, median time since stroke 3.6 years. Responsiveness was defined as the sensitivity of an instrument to real change. Two baseline measurements were performed with a 2-week interval before the intervention, and a follow-up measurement after 2 weeks of intensive forced use treatment. The limits of agreement, according to the Bland–Altman method, were computed as a measure of the test-retest reliability. Two different measures of responsiveness were compared: (i) the number of patients who improved more than the upper limit of agreement during the intervention; (ii) the responsiveness ratio. The limits of agreement, designating the interval comprising 95% of the differences between two measurements in a stable individual, were −5.7 to 6.2 and −5.0 to 6.6 for the ARA test and the FMA scale, respectively. The possible sum scores range from 0 to 57 (ARA) and from 0 to 66 (FMA). The number of patients who improved more than the upper limit were 12 (54.5%) and 2 (9.1%); and the responsiveness ratios were 2.03 and 0.41 for the ARA test and the FMA scale, respectively. These results strongly suggest that the ARA test is more responsive to improvement in upper extremity function than the FMA scale in chronic stroke patients undergoing forced use treatment.

Key words: upper extremity, cerebrovascular disorders, responsiveness, reproducibility, Action Research Arm test, Fugl-Meyer Assessment scale.

J Rehab Med 2001; 33: 110–113

Correspondence address: J. H. van der Lee, MD, Department of Rehabilitation Medicine, University Hospital Vrije Universiteit, P.O. Box 7057, NL-1007 MB Amsterdam, The Netherlands. E-mail: jh.vanderlee@azvu.nl

(Accepted September 13, 2000)

INTRODUCTION

Rehabilitation of arm function after a stroke is an important research topic. An increasing number of patients survive the acute phase (1). In the Copenhagen Stroke Study, a population-based study involving 88% of all stroke patients in a well-defined geographical area, it was found that 21% of the surviving patients had not attained full upper extremity function (defined as independence while using both arms) (2), and 36% had not attained independent walking function after comprehensive rehabilitation (3). The affected arm remained useless in 56% of the surviving patients in the sub-group with severe initial upper extremity paresis (4). The function of the upper extremity differs from the lower extremity with respect to the possibility of compensation. Impaired upper extremity function can be alleviated to some extent by compensation with the contralateral upper extremity (4), even though this "unaffected" arm may not function as well as the arm of a healthy subject (5).

One of the problems encountered in clinical trials to evaluate the effect of rehabilitative interventions for the hemiparetic upper extremity is the choice of valid, reliable and responsive outcome measures. Outcome measures that focus on independence in activities of daily living (ADL) are not specific for the motor function of the affected arm, because, at least theoretically, complete independence can also be achieved using only one arm (2). Therefore, instruments focussing on independence lack responsiveness to change in the function of the affected arm itself. The responsiveness of a measurement instrument is its sensitivity to true, clinically meaningful change (6–8). Different criteria have been used as indicators of clinically meaningful change, such as the change after an intervention of known efficacy (6) or the change in patients receiving the most effective therapy in a randomized clinical trial (8).

The Action Research Arm (ARA) test (9) and the upper extremity motor section of the Fugl-Meyer Assessment (FMA) scale (10) have frequently been used in clinical trials to measure improvement in the motor function of the affected arm itself (see for instance references 11–14). The ARA test and the upper extremity motor section of the FMA scale have been found to be equally responsive in the first 8 weeks post-stroke, when most of the recovery in arm function takes place (15). In a chronic population the changes in arm function as a result of treatment are expected to be much less impressive than in the first weeks after a stroke (2). The purpose of this study is to compare the responsiveness of the ARA test and the FMA scale in a chronic stroke population receiving intensive forced use treatment, which is aimed at improving dexterity and functional recovery of the hemiparetic arm (14).

METHODS

Patients and measurements

In a randomized clinical trial on the effectiveness of intensive forced use therapy to improve the arm function in chronic stroke patients, the ARA test and the FMA scale were used as outcome measures (14). Patients were included if they met the following criteria: (i) a history of a single stroke, at least 1 year previously, resulting in a hemiparesis on the dominant side; (ii) a minimum of 20° of active extension in the wrist and 10° of finger extension; (iii) ARA test score at intake below 51 (maximum score: 57); (iv) age 18–80 years; (v) able to walk indoors without a stick, indicating no major balance problems; (vi) no severe aphasia [score above P50 on the SAN (Stichting Afasie Nederland) test (16)]; (vii) no severe cognitive impairments (Mini Mental State Examination score of 22 or higher) (17). Sensory disorders were rated on a dichotomous scale. Any sensory deviations reported by the patient during the interview, or in a test involving alternating and simultaneous touching of both hands (with eyes closed), were rated as positive. Hemineglect was operationalized as a difference of at least two letters between the unaffected and the affected side in the letter cancellation test, or a significant ($p < 0.05$) deviation from the centre in a line bisection test comprising of 10 lines of 10 cm, assessed by means of a Wilcoxon signed rank sum test.

The protocol was approved by the University Hospital Medical Ethics Committee, and all patients gave written informed consent. Two baseline measurements were performed with a 2-week interval before the intervention commenced. A follow-up measurement took place within 3 weeks after the start of the 2-week intervention period. The experimental intervention, consisting of immobilization of the unaffected arm by means of a splint and sling, combined with intensive training of the affected arm for a period of 2 weeks, 5 days a week, 6 hours a day, was compared with an equally intensive reference intervention of bimanual training (14).

A sub-sample of the patient population was included in the present responsiveness study, based on the following additional criteria: (i) FMA score at intake below 60 (maximum score: 66); (ii) allocated to the experimental treatment; (iii) no missing values for any of the three measurements; (iv) both baseline measurements performed by the same rater. This sub-sample was compiled in such a way that improvement was equally possible on the FMA (score <60; maximum: 66) and on the ARA (score <51; maximum: 57), and the intervention currently perceived to be most powerful was applied to these patients (18).

Measurement instruments

The FMA scale is a performance test, in which the patient is asked to make movements that are considered to reflect the sequential stages of hyperreflexia, flexion and extension synergies, and the ability to perform selective movements (10). The upper extremity motor function section consists of 32 items which represent movement components, rated on a three-point ordinal scale (0–2). The score of one item, reflex activity, is doubled before calculating the sum score. The maximum sum score is 66, inferring optimal recovery. The FMA scale has been shown to be valid (10) and reliable (19).

For the domain in which the function of one upper extremity in purposeful activities is measured, Wade uses the term "focal disability" (20). The ARA test is an example of a measurement instrument that measures arm function in this domain (9). It is a performance test, in which the ability to perform gross movements and the ability to grasp, move and release objects differing in size, weight and shape, is tested. It consists of 19 items rated on a four-point ordinal scale (0–3). Summation of the 19 scores yields a sum score which ranges from 0 (none of the movements can be performed) to 57 (all movements are performed without difficulty). The ARA test has been shown to be valid and reliable (9, 15, 21).

Responsiveness measures

Many different approaches have been described to quantify responsiveness (8). The responsiveness of an instrument cannot be evaluated separately from its reliability (7). If an instrument shows a considerable change in the mean score of patients after an intervention, this can only be considered as an indication of the instrument's responsiveness if it has been shown to be reliable in a stable population (test-retest reproducibility) (8). We used the following responsiveness measures shown below.

Number of patients who improved more than the upper limit of agreement during intensive therapy. The Bland–Altman method was first used to evaluate the test-retest reproducibility (22). The limits of agreement are defined as the mean difference between the two measurements per individual ± twice the standard deviation. The upper and lower limits of agreement represent the "error thresholds" for improvement and deterioration, respectively. Assuming a normal distribution of the differences, just over 95% of the differences between the two measurements per individual in a stable population will be between these limits (22). The study population was considered to be stable between both baseline measurements because the patients were all in the chronic phase post-stroke (2). To compare the sensitivity to "real" (i.e. greater than the "error threshold") improvement for the ARA test and the FMA scale, the numbers of patients who improved more than the upper limit of agreement during the experimental intervention were compared.

The responsiveness ratio. According to Guyatt et al. (6) the responsiveness ratio (RR) was computed as the ratio of the mean change after the experimental intervention and the standard deviation of the mean change during the baseline period.

$$\mathrm{RR} = \frac{\text{mean improvement during intervention}}{\text{standard deviation of mean difference baseline}}$$

A higher responsiveness ratio indicates greater responsiveness.

In both these approaches, undergoing the experimental intervention was defined as a criterion for improvement and the baseline period represented stability.

RESULTS

The baseline characteristics of the 22 patients included in the study are presented in Table I. The Bland–Altman scatter-plots of the difference between baseline measurements against the mean baseline score per individual are shown in Figs 1 and 2. The mean difference and the limits of agreement, designating the interval comprising 95% of the differences between two measurements in a stable individual, are shown by the horizontal lines. The means and standard deviations of the ARA and FMA scores at the three subsequent measurement points and means and standard deviations of the differences between these measurements, as well as the limits of agreement and the responsiveness ratios, are presented in Table II.

During the baseline period (14–20 days, median 15 days) the mean change on either of the two instruments was small, 0.3 and 0.8 points for the ARA and the FMA, respectively, as is shown in Figs. 1 and 2 and in Table II. The limits of agreement were −5.7 to 6.2 (Fig. 1) and −5.0 to 6.6 (Fig. 2) for the ARA test and the

Table I. *Characteristics of the 22 chronic stroke patients included in the present study*

Median age (inter-quartile range)	58.5	(53.2–63.2)
Median years since stroke (inter-quartile range)	3.6	(2.1–6.3)
Females (%)	9	(40.9)
Diagnosis of hemorrhage (%)	6	(27.3)
Left-sided hemiparesis (%)	4	(18.2)
Sensory disorders present (%)	12	(54.5)
Hemineglect present (%)	2	(9.1)
Median baseline* ARA score (interquartile range)	38.0	(20.5–40.5)
Median baseline* FMA score (interquartile range)	49.2	(41.5–54.9)

* Mean of baseline 1 and 2 per patient.

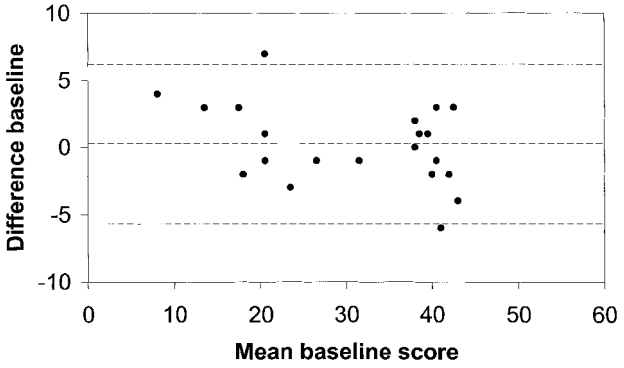

Fig. 1. Scatter-plot of the difference between the two baseline Action Research Arm (ARA) scores against the mean baseline ARA score per individual. The horizontal lines show the mean of the differences (middle) and the limits of agreement.

FMA scale, respectively. This means that the difference between two measurements has to be at least approximately 10% of the total range of each scale (10.9% for the ARA, and 10.0% for the FMA) to make measurement error unlikely, and to allow for the conclusion that real change has occurred. The mean difference after the intervention (20–27 days, median 21 days after the second baseline measurement) was much larger for the ARA (6.1 points) than for the FMA (1.2 points). The number of patients who improved more than the upper limit of agreement on the ARA test during the intervention period was 12 (54.5%), compared to 2 (9.1%) on the FMA scale, indicating a substantially larger responsiveness of the ARA test. This is also reflected in the responsiveness ratios, 2.03 and 0.41 for the ARA and the FMA, respectively.

DISCUSSION

Both methods applied in this study showed that the ARA test is substantially more responsive to improvement in upper extremity function than the FMA scale. This result is not in accordance with the findings of De Weerdt & Harrison, who concluded that the mean improvement on the ARA test and the FMA scale was

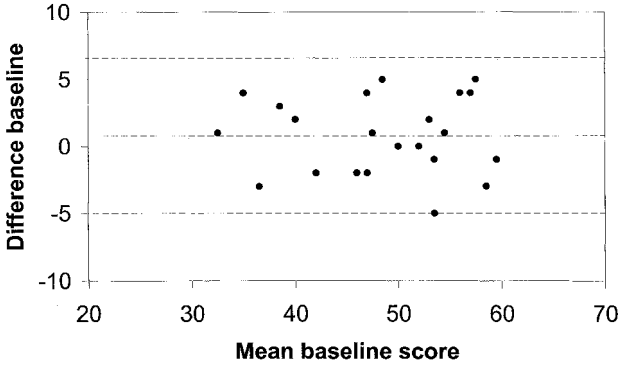

Fig. 2. Scatter-plot of the difference between the two baseline Fugl-Meyer Assessment (FMA) scores against the mean baseline FMA score per individual. The horizontal lines show the mean of the differences (middle) and the limits of agreement.

Table II. *Means and standard deviations (SDs) of Action Research Arm (ARA) and Fugl-Meyer Assessment (FMA) scores at three subsequent measurement points and of differences between these measurements; limits of agreement and responsiveness ratios (n = 22)*

	ARA test		FMA scale	
	Mean	SD	Mean	SD
Baseline 1	30.9	11.9	48.0	8.2
Difference baseline	*0.3*	*3.0*	*0.8*	*2.9*
Baseline 2	31.2	10.8	48.8	8.1
Improvement	*6.1*	*5.2*	*1.2*	*3.2*
Follow-up	37.3	13.4	50.0	7.8
Limits of agreement[1]	−5.7 to 6.2		−5.0 to 6.6	
Responsiveness ratio[2]	2.03		0.41	

[1] Mean difference baseline ±2 × SD difference baseline.

[2] Responsiveness ratio = $\frac{\text{mean improvement during intervention}}{\text{standard deviation of mean difference baseline}}$.

very similar, based on a scatter-plot of improvement on either test between 2 and 8 weeks post-stroke in a cohort of 53 patients (15). This difference in findings may be the result of the difference in chronicity of the patients, or could be due to methodological differences. It is conceivable that the improvement during the first weeks to months of recovery takes place at a level that differs from the domain in which the forced use treatment has its effect. During the first weeks to months after a stroke, the effect of rehabilitation combined with spontaneous recovery results in improvement at both the impairment level and at the disability level. De Weerdt & Harrison distinguished "motor recovery", represented by the subsequent stages of movement in and out of synergy patterns, as measured by the FMA scale, from "functional recovery", which is more at the disability level, as measured by the ARA test (15). There is no simple relationship between improvement at the impairment and disability levels (23). Therefore, and also because the forced use treatment was specifically aimed at the disability level, the greater responsiveness of the ARA test found in the present study is not surprising. The differences between these results and those of De Weerdt & Harrison may also be due to a difference in the initial arm function level of the included patients. Only one patient in the present study population had a first baseline score of less than 11 points on the ARA, and none of the patients scored less than 11 points on the FMA, whereas more than half of the patients in the De Weerdt & Harrison study scored less than 11 points on each test at the first measurement (15).

Although two different methods were used in this study to assess responsiveness, the underlying assumption was the same: the arm function of chronic stroke patients is stable between two baseline measurements 2 weeks apart, and it improves during a 2-week intervention of intensive forced use therapy. Since an established gold standard for improvement in arm function is lacking, undergoing the intensive forced use treatment was used as an external criterion for improvement. Although this

intervention has not yet been proven beyond doubt to be effective (14), the finding of a similar test-retest reproducibility of ARA and FMA, as expressed in the limits of agreement, supports the validity of the difference in responsiveness. In the clinical trial (14), the experimental intervention of forced use treatment was compared with a reference treatment of bimanual training. The patients undergoing the reference treatment were not included in this responsiveness study because bimanual training in the chronic phase is not generally considered to be effective. Alternatively, if the reference treatment could have been considered to be merely placebo-treatment, the results in the reference group could have been used to define stability. However, taking the mean improvement (1.7 points on the ARA test) during the reference intervention into account (14), the use of the baseline period (with a mean improvement of 0.3 points on the ARA test) to define stability was considered to be more valid.

As stated earlier, responsiveness has to do with real, clinically meaningful change. The definition used to determine improvement during the forced use treatment as clinically meaningful remains arbitrary. Although attempts are made to assess the minimal clinically important difference (MCID) empirically, this is difficult and, typically, very subjective (24). The limits of agreement designate the smallest signal that can be detected surmounting the test-retest "noise". As such, these values are merely statistical and bear no relationship to the MCID. However, if the limits of agreement were found to encompass the MCID, detection of a difference equal to the MCID would be impossible. This would imply inability of the instrument to detect changes that are considered to be clinically relevant (24). At the start of the clinical trial (14), the MCID was arbitrarily set at 10% of the total range of the scale, based on clinical experience and estimates reported for similar outcome measures in different domains (25). The resulting MCIDs of 5.7 and 6.6 points for the ARA test and the FMA scale, respectively, are very similar to the limits of agreement found in this study.

It is therefore concluded that the ARA test and the upper extremity motor section of the FMA scale are both reliable enough to detect clinically relevant changes, but the ARA test is substantially more responsive to improvement in upper extremity function in chronic stroke patients. Therefore, the ARA test is recommended to evaluate changes in arm motor function in chronic stroke patients.

ACKNOWLEDGEMENT

The present study has been financially supported by the NWO Council for Medical and Health Research, project number 904-65-045.

REFERENCES

1. Stegmayr B, Asplund K. Exploring the declining case fatality in acute stroke—population-based observations in the Northern Sweden MONICA project. J Int Med 1996; 240: 143–149.
2. Nakayama H, Jørgensen 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.
3. Jørgensen HS, Nakayama H, Raaschou HO, Olsen TS. Recovery of walking function in stroke patients: the Copenhagen Stroke Study. Arch Phys Med Rehabil 1995; 76: 27–32.
4. Nakayama H, Jørgensen 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.
5. Desrosiers J, Bourbonnais D, Bravo G, Roy PM, Guay M. Performance of the "unaffected" upper extremity of elderly stroke patients. Stroke 1996; 27: 1564–1570.
6. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171–178.
7. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992; 45: 1341–1345.
8. Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change score. Phys Ther 1996; 76: 1109–1123.
9. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Internat J Rehabil Res 1981; 4: 483–492.
10. Fugl-Meyer AR, Jääskö 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.
11. Feys HM, De Weerdt WJ, Selz BE, Cox Steck GA, Spichiger R, Vereeck LE, Putman KD, Van Hoydonck GA. Effect of a therapeutic intervention for the hemiplegic upper limb in the acute phase after stroke. A single blind, randomized, controlled multicenter trial. Stroke 1998; 29: 785–792.
12. Sonde L, Gip C, Fernaeus SE, Nilsson CG, Viitanen M. Stimulation with low frequency (1.7 Hz) transcutaneous electric nerve stimulation (low-TENS) increases motor function of the post-stroke paretic arm. Scand J Rehabil Med 1998; 30: 95–99.
13. Kwakkel G, Wagenaar RC, Twisk J. W. R, Lankhorst GJ, Koetsier JC. Intensity of leg and arm training after primary middle-cerebral artery stroke: a randomised trial. Lancet 1999; 354: 191–196.
14. Van der Lee JH, Wagenaar RC, Lankhorst GJ, Vogelaar TW, Devillé WL, Bouter LM. Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. Stroke 1999; 30: 2369–2375.
15. De Weerdt W, Harrison MA. Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnstrom–Fugl-Meyer test and the Action Research Arm test. Physiother Canada 1985; 37: 65–70.
16. Deelman BG, Koning-Haanstra M, Liebrand WBG, Van de Burg W. Handleiding van de SAN test. Lisse: Swets en Zeitlinger, 1987.
17. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 12: 189–198.
18. Duncan PW. Synthesis of intervention trials to improve motor recovery following stroke. Top Stroke Rehabil 1997; 3: 1–20.
19. Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Phys Ther 1983; 63: 1606–1610.
20. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press, 1995.
21. 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: 107–114.
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1 (8476): 307–310.
23. Roth EJ, Heinemann AW, Lovell LL, Harvey RL, McGuire JR, Diaz S. Impairment and disability: their relation during stroke rehabilitation. Arch Phys Med Rehabil 1998; 79: 329–335.
24. Hébert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. Arch Phys Med Rehabil 1997; 78: 1305–1308.
25. Brønfort G, Bouter LM. Responsiveness of general health status in low back pain: a comparison of the COOP charts and the SF-36. Pain 1999; 83: 201–209.