The Intra- and Interrater Reliability of the Action Research Arm Test: A Practical Test of Upper Extremity Function in Patients With Stroke

Johanna H. Van der Lee, MD, Vincent De Groot, MD, Heleen Beckerman, PT, PhD, Robert C. Wagenaar, PhD, Gustaaf J. Lankhorst, MD, PhD, Lex M. Bouter, PhD

ABSTRACT. van der Lee JH, de Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the Action Research Arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001;82:14-9.

Objectives: To determine the intra- and interrater reliability of the Action Research Arm (ARA) test, to assess its ability to detect a minimal clinically important difference (MCID) of 5.7 points, and to identify less reliable test items.

Design: Intrarater reliability of the sum scores and of individual items was assessed by comparing (1) the ratings of the laboratory measurements of 20 patients with the ratings of the same measurements recorded on videotape by the original rater, and (2) the repeated ratings of videotaped measurements by the same rater. Interrater reliability was assessed by comparing the ratings of the videotaped measurements of 2 raters. The resulting limits of agreement were compared with the MCID.

Patients: Stratified sample, based on the intake ARA score, of 20 chronic stroke patients (median age, 62yr; median time since stroke onset, 3.6yr; mean intake ARA score, 29.2).

Main Outcome Measures: Spearman's rank-order correlation coefficient (Spearman's rho); intraclass correlation coefficient (ICC); mean difference and limits of agreement, based on ARA sum scores; and weighted kappa, based on individual items.

Results: All intra- and interrater Spearman's rho and ICC values were higher than .98. The mean difference between ratings was highest for the interrater pair (.75; 95% confidence interval, .02–1.48), suggesting a small systematic difference between raters. Intrarater limits of agreement were −1.66 to 2.26; interrater limits of agreement were −2.35 to 3.85. Median weighted kappas exceeded .92.

Conclusion: The high intra- and interrater reliability of the ARA test was confirmed, as was its ability to detect a clinically relevant difference of 5.7 points.

Key Words: Arm; Cerebrovascular accident; Rehabilitation; Reproducibility of results; Treatment outcomes.

MANY DIFFERENT OUTCOME measures have been developed and used by various investigators to evaluate treatment for stroke patients. At least 27 different tests have been described in the literature, all of which measure function and dexterity of the affected upper limb.[1] The Action Research Arm (ARA) test has been used in several studies to measure upper limb function because of its presumed high reliability, high validity, and practical applicability.[1-4] Derived from the Upper Extremity Function Test,[5] the ARA test was first described by Lyle.[6] It is a performance test that consists of 4 subtests comprising 19 movements to be performed by the patient. Its concurrent validity has been confirmed by comparison with the Brunnstrom-Fugl-Meyer test,[7] with the Sollerman test,[1] and with the motor assessment scale,[8] and the intra- and interrater reliability of the ARA test were found to be satisfactory in earlier studies.[1,6,8] In 2 of these studies, reliability was assessed on the basis of a correlation coefficient,[1,6] which has been criticized for its inability to discern systematic differences between raters.[9] The third reliability study calculated the sum score of the ARA test in a very unusual way, combining the scores of both arms.[8] This approach might yield an unduly positive estimation of reliability because it can be expected that in most stroke patients, the ARA score of the upper extremity ipsilateral to the lesioned hemisphere will be maximal, thereby decreasing the chance of interrater disagreement. Furthermore, none of the earlier studies has examined the reliability of the individual items of the ARA test. The present reliability study aims to reconfirm the intra- and interrater reliability of the ARA test, and to identify items that are likely to cause disagreement between raters to formulate recommendations for improvement. We used 4 different measures of reliability because of their perceived complementary value.[10,11] The ARA test was administered as a primary outcome measure in a clinical trial evaluating the effectiveness of forced-use therapy of the hemiplegic arm in chronic stroke patients.[12] The minimal clinically important difference (MCID) was set at 10% of the scale's total range of 57 points, based on clinical experience and estimates reported in the literature for similar outcome measures in different domains.[13,14] One of the prerequisites for usefulness of the ARA test is that its measurement error is smaller than the estimated MCID. The following research questions were formulated:

1. To what extent do the ratings of 1 rater, made during the actual measurement, agree with his/her ratings of the videotape recording of that measurement after a period long enough to make recall of the actual measurement highly improbable? (We termed this condition intrarater reliability based on different sources of information.)

Department of Rehabilitation Medicine, University Hospital Vrije Universiteit, (Van der Lee, De Groot, Beckerman, Lankhorst); Institute for Research in Extramural Medicine, Vrije Universiteit, (Van der Lee, De Groot, Beckerman, Lankhorst, Bouter); and the Department of Physical Therapy, University Hospital Vrije Universiteit, Amsterdam, The Netherlands (Wagenaar). Wagenaar's current affiliation is: Department of Physical Therapy, Sargent College of Health and Rehabilitation Sciences, Boston, MA.

Accepted in revised form April 28, 2000.

Supported by the Netherlands Organization for Scientific Research (NWO) Council for Medical and Health Research (project no. 904-65-045).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the authors is/are associated.

Reprint requests to Johanna H. Van der Lee, MD, Dept of Rehabilitation Medicine, University Hospital Vrije Universiteit, PO Box 7057, 1007 MB Amsterdam, The Netherlands, e-mail: *jh.vanderlee@azvu.nl*.

0003-9993/01/8201-5913$35.00/0

doi:10.1053/apmr.2001.18668

2. To what extent do the ratings of 1 rater of a videotape recording of a measurement agree with his/her ratings of the same videotape after a period long enough to make recall of the first rating of the videotape highly improbable? (Condition: intrarater reliability based on the same source of information.)
3. To what extent do 2 raters agree when they independently score the same videotaped measurement of a patient? (Condition: interrater reliability based on the same source of information.)
4. Is the ARA test capable of detecting an MCID of 5.7 points?
5. Do any of the items typically cause disagreement between raters?

METHODS

Subjects

A subsample of 20 patients involved in a randomized clinical trial (RCT) on the effectiveness of forced-use treatment in chronic stroke patients served as the study population in the reliability study.[12] The randomized clinical trial included 66 subjects who met the following inclusion criteria: (1) a history of a single stroke, at least 1 year previously, resulting in hemiparesis on the dominant side; (2) a minimum of 20° of active extension in the wrist and 10° of finger extension; (3) ARA test score below 51 (max score, 57); (4) age between 18 and 80 years; (5) ability to walk indoors without a cane, indicating no major balance problems; (6) no severe aphasia (score $>$P50 on the Stichting Afasie Nederland test[15]); and (7) no severe cognitive impairments (Mini-Mental State score $\geq$22).[16] The protocol was approved by the hospital's medical ethics committee, and all patients gave written informed consent. After randomization into 2 groups, 2 baseline measurements were performed before the intervention commenced (M1, M2), 2 measurements were performed during the 2 weeks of the intervention (M3, M4), and 4 follow-up measurements took place at 3, 6, 26, and 52 weeks after the start of the intervention (M5–M8). A complete set of 8 measurements was obtained from 58 patients.[12]

ARA Test

The ARA test material consists of a wooden box, which is placed on a table in front of the patient, containing blocks and objects of different sizes. In 3 subtests (grasp, grip, pinch), the ability to grasp, move, and release objects differing in size, weight, and shape is tested. Objects must be picked up and moved vertically (subtests of grasp and pinch) or horizontally (subtest of grip) to a standardized location. Two items in the subtest of grip not only consist of horizontal movement, but also involve a certain degree of vertical movement and pronation (pouring water from 1 glass into another) or supination (turning a washer). In the 6 items in the subtest of pinch, the patient is asked to pick up marbles of 2 different sizes with 2 fingers only (thumb and index finger, thumb and middle finger, thumb and ring finger, respectively) and move them to a holder on top of the box. The fourth subtest consists of 3 gross movements (move hand to mouth, place hand on top of head, place hand behind head). The quality of the movements per item is rated on a 4-point scale: 0 = no movement possible; 1 = movement partially performed; 2 = movement performed, but abnormally; 3 = movement performed normally. To allow for easier distinction between scores 2 and 3, Wagenaar et al[1] set time limits for each item, based on the performance of a sample of 20 healthy subjects of similar age (see appendix). The limits were set at mean plus twice the standard deviation of the performance times of the healthy elderly subjects. To enable the movement to be timed, the patient was asked to start and finish each movement task with his/her hand flat on the table.

Reliability Study

Half the measurements (M2, M4, M6, M8) were videotaped. The camera was placed on a tripod at the unaffected side of the patient, at an angle of approximately 30° behind the frontal plane, at a distance of approximately 2 to 3 meters. The video camera, which also recorded sound, focused on the patient's upper body, and care was taken that the upper part of the ARA test box, containing the holders, and the back of the patient's chair were visible. The patient was seated with his/her back against the chair and instructed to keep contact with the back of the chair as much as possible throughout the test. The distance from the test box was standardized in such a way that the fingertips of the patient's passively extended paretic arm, supported by the investigator, could just touch the rear end of the top of the box. The starting position for the patient was with his/her paretic hand on the table. The patient was instructed to start performing the movement task after the investigator counted to 3, to try to complete the task at his/her usual movement speed, and to put the hand back on the table immediately after completion of the task. The patients were asked to try to perform all 19 movement tasks, and all their attempts were videotaped.

For this reliability study, the videotapes were stratified based on the patients' intake ARA score (ie, the score measured before enrollment for the trial). After arranging the patients by their intake ARA score, every third patient was selected for the sample, thus obtaining a sample size of 20. The aim of stratification was to obtain a sample with a variability similar to the original study population. The sample of videotaped measurements was chosen in a standardized way, so that the different measurements (M2, M4, M6, M8) were equally represented. The videotaped measurements were rated by the original examiner 4 to 27 months after the original measurement (intrarater reliability based on different sources of information), and again 4 to 6 weeks later (intrarater reliability based on the same source of information). Both intervals were considered to be long enough to exclude recall of the earlier ratings. The same sample of videotaped measurements was also rated once by a second rater (interrater reliability based on the same

Table 1: Intake Characteristics of Subjects (*n* = 20)

Median age in yr (IQR)	62 (52.5–71.8)
Median years since stroke (IQR)	3.6 (2.5–4.9)
Women	11 (55%)
Diagnosis of hemorrhage	3 (15%)
Left-sided hemiparesis	6 (30%)
Sensory disorders present	10 (50%)
Hemineglect present	2 (10%)
Intake ARA score*	29.2 ± 12.5
Intake FMA score*	49.2 ± 9.9

Abbreviations: IQR, interquartile range; FMA, Fugl-Meyer Assessment Scale.
*Intake (ie, pre-enrollment) ARA and FMA scores are expressed as mean ± SD assessed by first rater; higher ARA (maximum score, 57) and FMA (maximum score, 66) scores indicate better arm function, less impairment, respectively.[19]

Table 2: Results of 3 Different Statistical Methods to Estimate Intra- and Interrater Reliability of the Sum Score of the ARA Test, Based on Original Laboratory Measurements and Videotaped Measurements of 20 Chronic Stroke Patients

	Intrarater Reliability		
	Different Information* Sources	Same Information† Source	Interrater Reliability
Spearman's rho	.993	.995	.995
ICC‡	.997	.996	.989
Mean difference	0.15	0.30	0.75
(95% CI)	(−.40 to .70)	(−.16 to .76)	(.02–1.48)
Limits of agreement	−2.21 to 2.51	−1.66 to 2.26	−2.35 to 3.85

* Comparison of original laboratory measurement and videotaped measurement rated by the same rater.

† Comparison of repeated ratings of videotaped measurement by the same rater.

‡ Interrater ICC: $\text{ICC} = \dfrac{\text{BMS} - \text{EMS}}{\text{BMS} + (\text{k} - 1)\text{EMS} + \text{k}(\text{RMS} - \text{EMS})/\text{n}}$,

Intrarater ICC: $\text{ICC} = \dfrac{\text{BMS} - \text{EMS}}{\text{BMS} + (\text{k} - 1)\text{EMS}}$, in which n = the number of subjects; k = the number of raters/ratings; BMS; between-subjects mean square; EMS; error mean square; RMS; between-raters mean square.[10]

source of information). The raters, both clinical investigators, were experienced in the use of the ARA test.

Statistics

The reliability of the sum scores of all 19 items was assessed in 3 ways: (1) Spearman's rank-order correlation coefficient (Spearman's rho)[17]; (2) intraclass correlation coefficient (ICC)[11]; and (3) mean difference and limits of agreement (Bland and Altman plot).[9] The ICC was calculated from the mean squares of the sources of variance obtained by analysis of variance, with the computer program SPSS for Windows, version 8.0.[a] For the computation of the ICC for the interrater pair, the formula treating rater as a random effect was applied, considering the 2 raters as a sample of all possible raters, to allow for generalization of the findings.[11] For the intrarater ICC, the formula treating rater as a fixed effect was applied because there was only 1 rater involved in this particular trial.[12]

The Bland and Altman–defined limits of agreement ($D - 2s$, $D + 2s$) have been put into standard mathematical expression as $\Delta - 2SD$ and $\Delta + 2SD$, in which Δ is the mean of the differences between 2 ratings of the same subject, and SD is the standard deviation of the differences. Because the measurement errors will very likely follow a Gaussian distribution, 95% of the differences will lie between these limits of agreement (or more precisely, between $\Delta - 1.96SD$ and $\Delta + 1.96SD$).[9] A test is considered to be capable of detecting a difference of at least the magnitude of the limits of agreement.

To detect items that are more liable to cause rater disagreement, weighted kappas were calculated by individual item.[17] Weighted kappa takes the magnitude of the difference between ratings into account; differences between raters have more (diminishing) influence on weighted kappa when they are greater.

RESULTS

The sample included 20 patients (9 men, 11 women; median age, 62yr), with a median time since stroke of 3.6 years. The baseline characteristics are presented in table 1. Spearman's rho, ICCs, and the parameters according to the Bland and Altman plot (mean difference, limits of agreement) are presented in table 2. The values for Spearman's rho and ICC, which are all higher than .98, indicate good intra- and interrater reliability. The mean difference does not differ significantly from zero in the intrarater pairs, but it is slightly greater than zero in the interrater pair (.75; 95% confidence interval [CI]; .02–1.48), indicating a small systematic difference between the 2 raters.

Scatter plots of the difference between ratings against the mean of ratings for each of the 3 pairs of sum-score ratings are presented in figures 1, 2, and 3. The horizontal lines in these graphs show the mean of the differences and the limits of agreement. Visual inspection of the scatter plot of the interrater difference against the mean (fig 3) showed an outlier that had a 6-point difference between the sum-score ratings. Examina-

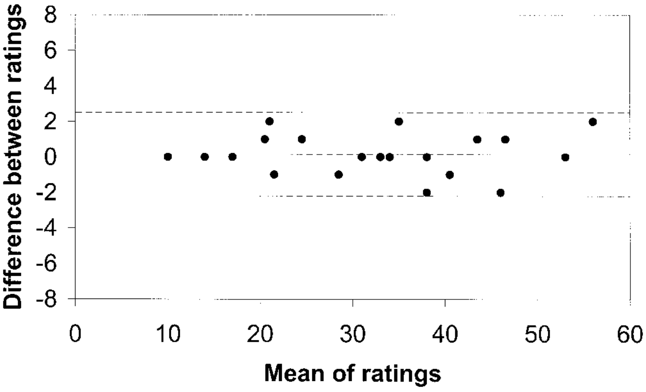

Fig 1. Difference between ratings against the mean of ratings (sum scores): intrarater pair based on different sources of information.

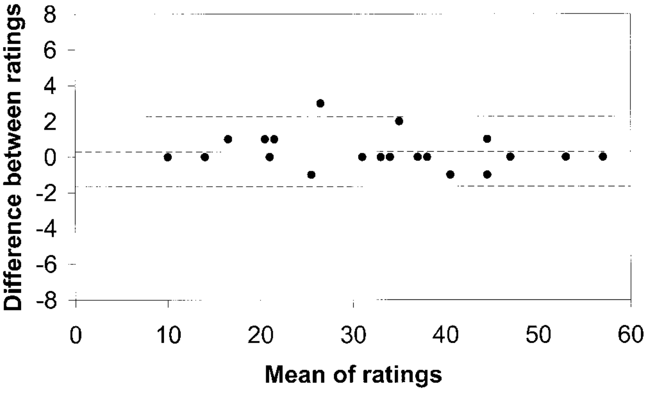

Fig 2. Difference between ratings against the mean of ratings (sum scores): intrarater pair based on the same source of information.

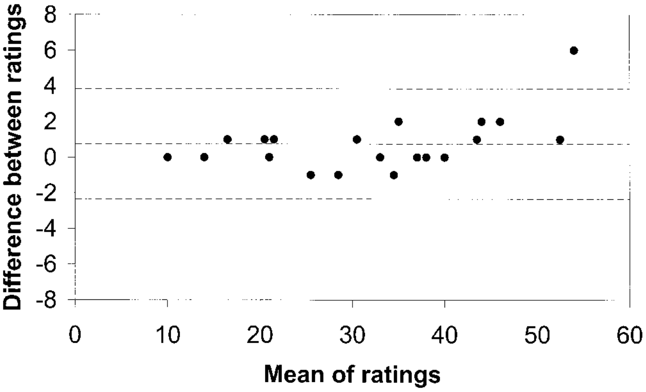

Fig 3. Difference between ratings against the mean of ratings (sum scores): interrater pair (both ratings of videotapes).

tion of the scores per individual item for this patient showed that despite his rapid performance, which was consistently within the time limit, his performance on several items was not normal, according to the second rater, because of inappropriate trunk and shoulder movements. Excluding this patient from the calculations, as was suggested as a possibility by Bland and Altman,[9] produced a mean difference of .47 points (95% CI = .00–.93) and limits of agreement of −1.46 to 2.40. Comparing the magnitude of the limits of agreement with the MCID, which we arbitrarily set at 10% of the total range of the scale (ie, 5.7 points), the values for the intra- and interrater pairs made it likely that a difference of 5.7 points could be distinguished from measurement error.

The range and median values of the weighted kappas per ARA subtest are presented in table 3. Agreement was good to very good on all individual items.[17] The lowest values were found for the gross movements subtest for all 3 pairs of ratings. From the cross-tabulations (data not shown), it became clear that the difference between ratings was never greater than 1 point on the 4-point scale.

DISCUSSION

Although the items on the ARA test are scored on an ordinal 4-point scale, performance on this test is usually expressed as a sum score, which is generally treated as an interval scale ranging from 0 to 57.[1-4,6-8] Statistical methods to estimate reliability for ordinal scales are different from those for interval or continuous scales, although the underlying statistical principles are similar.[10] Because the sum score of the ARA test is used in the analysis of clinical trials, this is the most appropriate level to present. Although Pearson's or Spearman's correlation coefficients are commonly considered insufficient for establishing reliability, because of their incapacity to detect systematic differences,[9] they were nevertheless presented in early literature reports on the reliability of the ARA test.[1,6] The Spearman's rho found in the present study are similar to the Pearson's product moment correlation coefficient of .99 found for the interrater reliability reported by Lyle,[6] and to the Spearman's rho of .99 for the intrarater reliability reported by Wagenaar et al.[1] The interrater ICC value of .99, found in this study, was also similar to the interrater ICC of .98 reported by Hsieh et al.[8] It should be noted that the sum scores were calculated differently in the above-mentioned studies. In all 3 studies, the sum score was based on a Guttman scale, and the sum score calculated by Hsieh et al[8] differed from the other 2 studies and from the present study in that the scores for the affected and the unaffected arm were added together, resulting in a maximum possible score of 114.

In the present study, we assumed that agreement would be highest when information from the same source (ie, videotapes) was rated by the same person, and lowest in the interrater pair. This assumption was not confirmed because all 3 rating pairs had high reliability scores, and the differences between them were negligible. The videotaped measurements do not appear to be much more difficult to rate than the measurements observed in the real laboratory situation. However, the existence of an outlier (see fig 3) shows that, despite the use of the time limits set by Wagenaar,[1] some disagreements still occurred between raters with regard to scores 2 and 3. This issue could probably be resolved by applying a more explicit criterion to assess whether the patient's back is actually in contact with the back of the chair throughout the entire performance of the tasks. The examiner could perhaps hold his hand behind the patient's back but, of course, this method could not be used for videorecordings. Another alternative might be to install an electric sensor in the back of the chair.

It is important to know the variability between patients in a reliability study to assess the comparative value of the reported ICC.[10] A large variability between patients automatically enhances the ICC. The way in which the sample was selected in the present study made it representative for the entire population involved in the trial. The mean intake ARA score ± SD of all 66 subjects was 29 ± 12.7, indicating that the sample of 20 patients (mean intake ARA score, 29.2 ± 12.5) was representative in this respect. As can be seen in the figures, the sample of patients in the present study was composed in such a way

Table 3: Range (Median) of Weighted Kappa* as a Measure of Intra- and Interrater Reliability of Individual Items, Grouped per Subtest, of the ARA Test, Based on Original Laboratory Measurements and Videotaped Measurements of 20 Chronic Stroke Patients

		Intrarater Reliability		
Subtests	No. of Items	Different Information† Sources	Same Information‡ Source	Interrater Reliability
Grasp	6	.87–1 (.92)	.93–1 (1)	.83–1 (.90)
Grip	4	.92–1 (.95)	.92–1 (.98)	.83–1 (.95)
Pinch	6	.90–1 (.98)	.92–1 (1)	.90–.95 (.95)
Gross movements	3	.78–.87 (.83)	.78–.92 (.88)	.83–.91 (.87)
All items	19	.78–1 (.94)	.78–1 (1)	.83–1 (.93)

* Weighted kappa takes the magnitude of the difference between ratings into account: $\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$, in which $p_{o(w)}$ is the weighted observed proportional agreement, and $p_{e(w)}$ is the weighted expected proportional agreement.[17]

† Comparison of original laboratory measurement and videotaped measurement rated by the same rater.

‡ Comparison of repeated ratings of videotaped measurement by the same rater.

that almost the entire range of the scale was represented, with the exception of the lower part.

The ICC and the Bland and Altman plot yield complementary information.[11] The ICC makes comparison between different measurement instruments possible, when applied in similar study populations, whereas the scatter plots of differences between ratings against the mean of ratings according to Bland and Altman provide insight into the distribution of differences between 2 measurements. One characteristic of the Bland and Altman plot, which can be either an advantage or a disadvantage,[9,10] is that it is expressed in the same units as the scale itself. The limits of agreement represent an estimate of the range of rating-pair differences within which 95% of the differences between 2 ratings will lie. The small systematic difference between the 2 different raters in the present study could not have been found by merely estimating the ICC, which confirms the additional value of the Bland and Altman plot.

Another advantage of the limits of agreement is that they can be compared with the somewhat arbitrary value of the MCID (5.7) and with the difference between treatment groups found in the RCT (3; 95% CI = 1.3–4.8).[12] Had the limits of agreement been greater than the MCID of 5.7, this would suggest that the ARA test would not be reliable enough to detect a difference that is considered to be clinically relevant.[18] The difference between groups that was found in the RCT exceeds the intrarater limits of agreement, which supports the validity of this finding as a "signal" surmounting the "noise" because of the variability of ratings. The somewhat greater interrater limits of agreement do not affect this conclusion because (with a few exceptions) there was only 1 rater involved in the trial.

The weighted kappas provide insight into certain subtests or items that are relatively easier or more difficult to rate than others. When comparing the weighted kappas for the different ARA subtests (table 3), agreement on the gross movements subtest seemed to be lower than on the other 3 subtests. From examination of the cross-tabulations of the individual items, it became clear that most of the inconsistencies were found for the item "hand to mouth." For this item, the obvious difficulty was in distinguishing between scores 2 and 3. This is probably connected with 2 aspects of this item: (1) the small range of scores in the sample of patients, who all scored higher than 1 on this item, and (2) the short time limit set for this item. The time limit is 2.4 seconds, above which the score is 2 instead of 3, implying that there is a greater risk for timing errors for this item than for other items, which have a longer time limit. Considering the overall good reliability of the ARA test, it is not recommended that this item be changed or omitted because it is perceived to be an important item if there is a lower level of upper extremity function—a condition that was not represented in the study sample.

CONCLUSION

The present study confirms the high intra- and interrater reliability of the ARA test in a population of chronic stroke patients with a moderate residual loss of arm function. It is capable of detecting a difference of 10% of its maximum possible sum score of 57 points, which is considered to be clinically relevant. To make a clear distinction between scores 2 and 3, it is recommended that an explicit criterion be applied to assess patients' contact with the back of the chair, in combination with the time limits.

References

1. Wagenaar RC, Meijer OG, van Wieringen PCW, Kuik DJ, Hazenberg GJ, Lindeboom J, et al. The functional recovery of stroke: a comparison between neuro-developmental treatment and the Brunnstrom method. Scand J Rehabil Med 1990;22:1-8.
2. Feys HM, De Weerdt WJ, Selz BE, Cox Steck GA, Spichiger R, Vereeck LE, et al. 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: 785-92.
3. Kwakkel G, Wagenaar RC, Twisk JWR, Lankhorst GJ, Koetsier JC. Intensity of leg and arm training after primary middle-cerebral artery stroke: a randomised trial. Lancet 1999;354:191-6.
4. Powell J, Pandyan AD, Granat M, Cameron M, Stott DJ. Electrical stimulation of wrist extensors in poststroke hemiplegia. Stroke 1999;30:1384-9.
5. Carroll D. A quantitative test of upper extremity function. J Chronic Dis 1965;18:479-91.
6. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981;4:483-92.
7. 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 Can 1985;37:65-70.
8. 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;27:107-14.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i:307-10.
10. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford Univ Pr; 1995. p 104-27.
11. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998;12:187-99.
12. 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;30:2369-75.
13. Brønfort G. Efficacy of manual therapies of the spine [dissertation]. Amsterdam: Thesis Publishers Amsterdam; 1997.
14. 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;83:201-9.
15. Deelman BG, Koning-Haanstra M, Liebrand WBG, van de Burg W. Handleiding van de SAN test. Lisse: Swets en Zeitlinger; 1987.
16. Dick JPR, Guilloff RJ, Stewart A, Blackstock A. Mini-mental state examination in neurological patients. J Neurol Neurosurg Psychiatry 1984;47:496-9.
17. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991. p 406-7.
18. Hébert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. Arch Phys Med Rehabil 1997;78:1305-8.
19. 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.

Supplier

a. SPSS, Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.

APPENDIX

Time limits (mean + 2 SD of the performance times of 20 healthy elderly subjects) for each of the 19 items of the ARA test.[1] If performance is slower than the time limit or if the patient loses contact with the back of the chair during performance, the score is 2 instead of 3.

Subtest	Items	Time Limit (s)
Grasp	Block 2.5cm	3.6
	Block 5cm	3.5
	Block 7.5cm	3.9
	Ball 7.5cm	3.8
	Stone	3.6
	Block 10cm	4.2
Grip	Tube 2.25cm	4.2
	Tube 1cm	4.3
	Place washer over bolt	4
	Pour water from glass to glass	7.9
Pinch	Large marble first finger and thumb	3.8
	Large marble second finger and thumb	3.8
	Large marble third finger and thumb	4.1
	Small marble first finger and thumb	4
	Small marble second finger and thumb	4.1
	Small marble third finger and thumb	4.4
Gross Movements	Move hand to mouth	2.4
	Place hand on top of head	2.7
	Place hand behind head	2.7