Taylor & Francis
Taylor & Francis Group

VALIDATION OF THE ACTION RESEARCH ARM TEST USING ITEM RESPONSE THEORY IN PATIENTS AFTER STROKE

Chia-Lin Koh, BS[1], I-Ping Hsueh, MA[2], Wen-Chung Wang, PhD[3], Ching-Fan Sheu, PhD[4], Tzu-Ying Yu, BS[2], Chun-Hou Wang, BS[5] and Ching-Lin Hsieh, PhD[2]

From the [1]School of Health and Rehabilitation Sciences, the University of Queensland, [2]School of Occupational Therapy, College of Medicine, National Taiwan University and Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, [3]Department of Psychology, National Chung Cheng University, [4]Institute of Cognitive Science, National Cheng Kung University, Tainan, Taiwan and [5]School of Physical Therapy, Chung Shan Medical University and Department of Physical Therapy, Chung Shan Medical University Rehabilitation Hospital

Objective: **To validate the unidimensionality of the Action Research Arm Test (ARAT) using Mokken analysis and to examine whether scores of the ARAT can be transformed into interval scores using Rasch analysis.**

Subjects and methods: **A total of 351 patients with stroke were recruited from 5 rehabilitation departments located in 4 regions of Taiwan. The 19-item ARAT was administered to all the subjects by a physical therapist. The data were analysed using item response theory by non-parametric Mokken analysis followed by Rasch analysis.**

Results: **The results supported a unidimensional scale of the 19-item ARAT by Mokken analysis, with the scalability coefficient H = 0.95. Except for the item "pinch ball bearing 3rd finger and thumb", the remaining 18 items have a consistently hierarchical order along the upper extremity function's continuum. In contrast, the Rasch analysis, with a stepwise deletion of misfit items, showed that only 4 items ("grasp ball", "grasp block 5 cm^3", "grasp block 2.5 cm^3", and "grip tube 1 cm^3") fit the Rasch rating scale model's expectations.**

Conclusion: **Our findings indicated that the 19-item ARAT constituted a unidimensional construct measuring upper-extremity function in stroke patients. However, the results did not support the premise that the raw sum scores of the ARAT can be transformed into interval Rasch scores. Thus, the raw sum scores of the ARAT can provide information only about order of patients on their upper extremity functional abilities, but not represent each patient's exact functioning.**

Key words: psychometrics, cerebrovascular accident, arm.

J Rehabil Med 2006; 38: 375–380

Correspondence address: Ching-Lin Hsieh, School of Occupational Therapy, College of Medicine, National Taiwan University, 4F, No 17, Shiujou Rd, Taipei 100, Taiwan. E-mail: mike26@ha.mc.ntu.edu.tw

Submitted August 31, 2005; accepted May 2, 2006

INTRODUCTION

Upper extremity dysfunction occurs in approximately 30–66% of stroke survivors (1). For patients who have had a stroke, upper limb impairment is a major obstacle to re-acquiring competency in performing activities of daily living (2). These disabilities often produce long-term needs for assistance from caregivers and society (3). Accurately measuring the upper extremity function of patients with stroke is essential for appropriate treatment planning, clinical decision-making and research (e.g. outcome studies) (4–6). Therefore, a valid upper extremity functional measure for patients after stroke is crucial for both clinicians and researchers.

The Action Research Arm Test (ARAT) (2) is a measure widely used in evaluating the upper extremity function of patients after stroke. The ARAT has been found to have satisfactory psychometric properties (including intra-/inter-rater reliability, concurrent/convergent validity and responsiveness) using classical test theory (2, 7–12). However, 2 shortcomings remain in using this measure. First, the unidimensional construct of the ARAT has rarely been examined. Unidimensionality of the ARAT is crucial to determine what the ARAT is uniquely measuring and to ascertain whether the item scores of the ARAT can be summed up to quantify upper extremity function (13). To our knowledge, the unidimensionality of the ARAT had been examined only by van der Lee and co-workers (14) using the Mokken analysis (15). The Mokken analysis is a non-parametric modern item response theory (IRT) model that examines accuracy of ordering between persons' raw sum scores on a measure to determine unidimensionality (14, 16). With a sample of 63 patients in their study, van der Lee et al. found that the ARAT comprised a unidimensional scale. However, Mokken analysis typically requires a sample size larger than 200 to be reliably to estimate the unidimensionality of a scale (15). In addition, their sample could not be considered as representative of the total stroke population because neither slightly impaired nor severely impaired patients were included in the sample. The results of their study, in our opinion, therefore did not provide conclusive evidence supporting the unidimensionality of the ARAT.

The second shortcoming of the use of the ARAT is that, with the Mokken analysis, the raw sum scores of the ARAT attain only the status of ordinal scores instead of interval scores, even if the unidimensionality of the ARAT has been verified. Because of the unequal interval between 2 scores, the numeric scores on an ordinal scale cannot exactly represent a person's functioning condition (17). Interval scores, on the other hand, represent an

DOI: 10.1080/16501970600803252

underlying trait in which equal intervals between any 2 points on a scale are of equal value. The interval property allows one to quantify change in a way which will support arithmetic operations such as subtraction and make sum scores of 2 different measuring results comparable. Furthermore, an interval score can be analysed by parametric statistics, which are often more powerful than non-parametric methods (18). Therefore, an interval-scale measure would enable clinicians and researchers to quantify upper extremity functional changes within patients and differences between patients who have had a stroke and to obtain a more accurate reflection of disease impact, functional recovery, and treatment effects in patients than is possible with ordinal-scale measure (19).

Parametric IRT models can be applied to examine whether sum scores of a measure can be transformed into interval scores. One well-known analysis of this approach is the Rasch analysis, which is a technique used to establish the interval scale property of a measuring instrument (20). Items that fit the Rasch model's expectations can be used to generate logit scores and can be viewed as interval scores (21, 22). "Logit" is a contraction of "Log-Odds Unit". The odds are the probability that an outcome does occur divided by the probability that the outcome does not occur. The logit score is the logarithm of the odds associated with the probability. When data fit the Rasch model's expectations, raw scores obtained from ordinal data can be transformed to logit scores which form an equal interval linear scale (23, 24).

The major difference between the Mokken and the Rasch analyses is that the Rasch model further requires parametric functional forms for item response function (IRF) of items, thus enabling transformed interval scores of a measure to be obtained (20). However, with this parametric assumption, the Rasch model tends to exclude items whose scores cannot be transformed into an interval scale, but which do fit the Mokken model's expectations, i.e. fit the unidimensional construct of a whole item set. Based on a basic definition of unidimensionality – that is, an item set is unidimensional if its true scores can be shown to be a monotonic increasing function of a single underlying latent variable (25) – the Mokken model is believed to exemplify the simplest form of unidimensionality (26). Therefore, the purpose of this study was 2-fold: (*i*) to validate the unidimensionality of the ARAT using Mokken analysis with a large sample; and (*ii*) to examine whether scores of the ARAT can be transformed into interval scores using Rasch analysis.

METHODS

Subjects

To select patients after stroke with a broad range of upper extremity dysfunction, subjects were recruited from 5 rehabilitation departments located in northern, central, southern and eastern Taiwan between October 2003 and January 2004. All inpatients and outpatients of the rehabilitation departments were invited to participate in the study if they met the following criteria: (*i*) diagnosis (International Classification of Diseases, Ninth Revision Clinical Modification [ICD-9-CM] codes) of cerebral haemorrhage (431) or cerebral infarction (434); (*ii*) ability to follow instructions; and (*iii*) absence of other major diseases (e.g. tumours or arthritis) or impairments (e.g. amputations or fractures) that would reduce or limit patients' ability to perform upper extremity tasks. Only patients who were able to give informed consent personally or by proxy (for those who were illiterate or unable to sign the informed consent form) were included in this study. The project was approved by the local ethics review boards.

Procedure

All of the 19 items of the ARAT were administered by the same physical therapist to the patients at the 5 rehabilitation departments. Patients' demographic details and data on co-morbidity were collected from their medical records.

Instrument

The ARAT, developed by Lyle (2), is based on the upper extremity function test of Carroll (27). It is designed to assess the recovery of upper extremity function following a cortical injury. The ARAT contains a total of 19 items and is divided into 4 subscales – "grasp" (6 items), "grip" (4 items), "pinch" (6 items), and "gross motor" (3 items). In the former 3 subscales, the ability to grasp, move, and release objects differing in size, weight, and shape is tested. The fourth subtest consists of 3 gross movements (place hand behind head, place hand on top of head, and move hand to mouth). The items are graded on a 4-point scale: 0 – cannot perform any part of the test; 1 – can partially perform the test; 2 – can complete the test but took abnormally long or had great difficulty; 3 – can perform the test normally. The maximum total score of 57 indicates the absence of upper extremity dysfunction.

Data analysis

Two models of Mokken scale analysis were performed using the MSP 5.0 computer program (15). First, the monotone homogeneity (MH) model for polytomous items was used to examine the unidimensionality of the ARAT (15). The MH model has 3 assumptions: (*i*) items form a unidimensional scale (measuring the same construct, e.g. upper extremity function); (*ii*) item scores are locally independent (e.g. the scores on a given set of items are stochastically independent of each other within a group of persons with the same level of upper extremity function); and (*iii*) the IRF for each item is a monotonically non-decreasing function of the underlying construct, which means that patients at a higher level of upper extremity function have a higher probability of scoring higher for an item. The fit of the MH model is evaluated by calculating the scalability coefficient H for the scale and H_i for each item i (15). The scalability coefficient H is a global indicator of the degree to which patients can be accurately ordered on the upper extremity function by means of their sum scores. Higher values of H indicate fewer violations of the assumptions and thus a better scale. A unidimensional scale is considered to be supported if $H \geq 0.50$ (15). Secondly, the double monotonicity (DM) model (15) (assuming that the IRFs of the scale do not intersect, in addition to the 3 assumptions of the MH model) was used to test whether the items of the ARAT possessed an invariant hierarchical ordering, which means that the difficulty order of all 19 items of the ARAT is the same for all patients suffering from a stroke. Thus, if item A is harder than item B for one patient, then item A is harder than item B for all patients. Moreover, this holds true for any pair of items on the scale. The fit of the DM model was investigated by 2 criteria: "Pmatrix crit" and "Restscore crit". A scale is considered to adequately meet the DM model if the largest Crit value per item is smaller than 40. If the values of both criteria for an item are found to be larger than 80, the invariant hierarchical ordering is seriously violated for this item (15).

To examine the parametric function of the ARAT, the Rasch rating scale model (28) was employed using the WINSTEPS program (29). The Rasch rating scale model is useful for polytomous items when one assumes that psychological distances, or thresholds, between scoring categories are the same for all items (21, 30). In this study, because the ARAT is a 4-point Likert scale (i.e. all items are rated 0, 1, 2, 3), the rating scale model was used in this study. In addition to the 4 assumptions of the Mokken analysis, the Rasch rating scale model requires a one-parametric functional form for the IRFs; that is, all IRFs have the same slope and differ only in item difficulty (26, 31). The same slope means the same value of the slope which is the average discrimination of all the items (29). Two fit statistics were used to examine whether the data fit the Rasch model's expectations. The infit mean square standardized residual (MNSQ) is sensitive to unexpected behaviour affecting responses to items near the person's functional ability in upper extremity function; the outfit MNSQ is sensitive to unexpected behaviour by persons on items far from the level (21, 22). Consequently a MNSQ expected value is close to 1.0 (32). Values greater than 1.0 (underfit) indicate the presence of unmodelled variance (noise) along with the useful information in the data. These degrade measurement. Values less than 1.0 (overfit) indicate better than expected fit to the model. These responses agree with but add little additional information to other responses (29, 32). Based on the publications of Wright & Linacre (33), the range of acceptable infit/outfit MNSQ values in this study is 0.6–1.4. The MNSQ value can be transformed to standardized value (called ZSTD) to test if the data fit the model's expectations. The ZSTD values follow approximately a t distribution, or the standard normal distribution, when the items fit the model's expectation. Z-scores reported in WINSTEPS are unit-normal deviates, in which only about 2.5% of the scores are larger (smaller) than 1.96 (-1.96) (29). The misfit criteria in this study were predefined as follows (21, 33): (*i*) infit ZSTD >1.96 and MNSQ >1.4 or outfit ZSTD >1.96 and MNSQ >1.4; and (*ii*) infit ZSTD < -1.96 and MNSQ <0.6 or outfit ZSTD < -1.96 and MNSQ <0.6. Items considered to misfit the Rasch model were removed in a stepwise manner by inspecting a series of infit to outfit statistics. In addition, the appropriateness of the scoring categories of the ARAT was investigated using the Rasch rating scale model. Estimates of the threshold difficulty between the adjacent scoring levels can be used to examine the appropriateness of the scoring categories of the ARAT (34). If disorderings of the step difficulty (i.e. the difficulty of a higher step is lower than that of its adjacent lower step) between any 2 adjacent levels were found, then the scoring categories of the items might be reorganized to achieve suitable scoring categories.

RESULTS

A total of 351 patients were recruited in the study. The characteristics of the subjects are presented in Table I. The participants had a wide range of upper extremity function deficits, and their sum scores of the ARAT were scattered throughout the full range of possible scores (0–57).

Table II shows that the range of scalability coefficient H_i of each item of the ARAT fells between 0.92–0.97. The scalability coefficient H of the 19-item ARAT is 0.95, which is well above the criterion of 0.5. The Pmatrix and Restscore Crit values of each item of the ARAT were all below the benchmark of 80, except for the "pinch ball bearing 3rd finger and thumb" (Pmatrix Crit $=93$), indicating little violation of the assumption of invariant item ordering. After removing the item "Pinch ball bearing 3rd finger and thumb", the 18-item ARAT fitted the Mokken DM model's expectations well ($H = 0.95$; Pmatrix <57 and Restscore <15). Thus, we concluded that the 18-item ARAT fitted the DM model's expectations.

Because parameter estimates and fit statistics of the Rasch analysis depended on other items in the test and test length, misfit items were generally removed in a stepwise manner. We found that only 4 out of 19 items of the ARAT ("grasp ball", "grasp block 5 cm^{3}", "grasp block 2.5 cm^{3}", and "grip tube 1 cm^{3}") fitted the Rasch model's expectations. The values in Table III were tentative to give a general impression of the 19-item ARAT Rasch model-data fit, showing 12 of the ARAT items did not fit the Rasch model's expectations in the initial analysis (infit or outfit ZSTD >1.96 and MNSQ >1.4; or infit or outfit ZSTD < -1.96 and MNSQ <0.6). The Rasch partial credit model had also been used to examine the 19-item ARAT. Similarly, only 6 items fitted the expectations of the Rasch partial credit model. The threshold difficulty estimates of the ARAT were far apart ($>$ 0.90 logits). In addition, the ordering of the threshold difficulty estimates was not reversed. These results indicate that the scoring categories of the ARAT are acceptable.

Table I. *Characteristics of the patients after stroke ($n = 351$)*

Characteristics	
Gender (male/female)	222/129
Age, median (interquartile range)	63 (53–71)
Month after onset, median (interquartile range)	12.5 (4–30)
Diagnosis, *n* (%)	
Cerebral haemorrhage	113 (32)
Cerebral infarction	238 (68)
Side of paresis, *n* (%)	
Right	175 (50)
Left	176 (50)
ARAT sum score, median (interquartile range)	5.0 (0–40)
Severity of UE function, *n* (%)	
Severe (ARAT <5)	175 (50)
Moderate	117 (33)
Mild (ARAT >51)	59 (17)

ARAT = Action Research Arm Test; UE = upper extremity.

DISCUSSION

This study was the first to use both a non-parametric Mokken analysis and a parametric Rasch analysis with a large enough sample to reliably validate the measurement properties of the ARAT in patients who have suffered a stroke. We found that the ARAT was consistent with the MH and DM models' expectations, except for one item in the DM model. This result demonstrated that the 19 items of ARAT belong to the same construct, which can be named the upper extremity function based on Lyle's original design of the ARAT (2). Also, except for the item "pinch ball bearing 3rd finger and thumb", the remaining 18 items have a consistently hierarchical order along the upper extremity function's continuum. However, the measure was not consistent with the Rasch rating scale model, indicating that raw scores from this measure cannot be transformed into interval scores.

We found that 18 items of the ARAT (except "pinch ball bearing 3rd finger and thumb") fit the DM model of the Mokken scale analysis, meaning that the difficulty of ordering of these items was the same for all individuals. The misfit to the DM model of "pinch ball bearing 3rd finger and thumb" was also found in the study by van der Lee et al. (14), indicating that the difficulty ordering of this item varied from the other items and should be removed. However, the other 3 items these authors found as misfitting the DM model – "pinch marble 3rd finger and thumb", "pinch ball bearing 2nd finger and thumb", and "pinch ball bearing 1st finger and thumb" – were not

Table II. *Mokken scale analysis of the Action Research Arm Test (ARAT) arranged in ascending order of mean, indicating item difficulty from high to low*

Item	Mean	ItemH (H_i)	Pmatrix[†]	Restscore[†]
Pinch ball bearing 3rd finger and thumb[‡]	0.60	0.92	93	
Pinch marble 3rd finger and thumb	0.71	0.93	60	
Pinch ball bearing 2nd finger and thumb	0.76	0.95		
Pour water glass to glass	0.79	0.94	1	
Grasp block (10 cm^3)	0.81	0.93		
Pinch ball bearing 1st finger and thumb	0.84	0.94	4	
Pinch marble 2nd finger and thumb	0.85	0.95	18	
Pinch marble 1st finger and thumb	0.94	0.94		
Grasp block (7.5 cm^3)	0.97	0.96	36	
Grip washer over bolt	0.97	0.95	2	
Grasp ball	1.00	0.96	51	
Grip tube (2.25 cm^3)	1.03	0.97	37	2
Grasp stone	1.05	0.97	44	5
Grip tube (1 cm^3)	1.06	0.96	46	
Grasp block (5 cm^3)	1.07	0.96	38	
Place hand behind head	1.12	0.92	44	7
Grasp block (2.5 cm^3)	1.14	0.96	15	14
Place hand on top of head	1.29	0.94	51	15
Hand to mouth	1.45	0.96		5

[†] Values of items with violations smaller than the minimum criteria of the MSP 5.0 computer program are not shown.
[‡] Item that showed violation ordering (Pmatrix >80).
Note: Because the ARAT generally did not fit the Rasch model's expectations, it is not appropriate to calculate mean score for the ARAT. We show the mean score of the ARAT here, only because it can be used for inter-study comparison.

found to deviate from the DM model's expectations in the current study. The differences between their sample characteristics and ours might account for these discrepancies: Our sample covered the full range of possible scores of the ARAT (0–57), whereas their sample did not include patients with severe upper extremity dysfunction (i.e. ARAT <5) and patients with mild upper extremity dysfunction (i.e. ARAT >51). In particular, "pinch marble 3rd finger and thumb" and "pinch ball bearing 2nd finger and thumb" were the 2 most difficult items that showed a good fit to the Mokken model in our study, but a poor fit in the study by van der Lee et al. (14). Thus, the presence of subjects with mild upper extremity dysfunction (i.e. ARAT >51) in the sample of this study may have caused the differences in the results of the studies.

Because collapsing of categories might improve model-data fit of the Rasch analysis (35, 36), we had tried to collapse the middle categories (i.e. recoding 0123 to 0112) due to the low gap between the first 2 thresholds to determine whether the fit of the ARAT would be improved. We found that seven items of the revised ARAT fitted the Rasch model's expectations. However, because up to 63% of the 19 items remained misfit, this result was still not satisfactory. Future studies may endeavour to redefine the scoring categories of the ARAT to improve model-data fit.

The poor Rasch model-data fit suggests that the current items of the ARAT cannot meet the parametric assumption of the Rasch model, indicating that raw scores of the ARAT cannot be transformed into interval scores. Given the aforementioned advantages for further usage of interval scores, which are especially important for the calculation of change scores in medical outcome studies, more effort should be invested into revising the ARAT to make it fit the Rasch model's expectations. Researchers who are interested in constructing an interval level measure of upper extremity function may base their work on the 4 remaining items ("grasp ball", "grasp block 5 cm^3", "grasp block 2.5 cm^3", and "grip tube 1cm") to revise the items of the ARAT. They may also use the generalized one-parameter logistic model (37, 38), a less stringent model than the Rasch model, to analyse the ARAT before revision.

Further revisions should also address the issue of local dependence. Because of the original design of the ARAT, items in each subscale are similar and correlated, which may cause local item dependence (LID). The item fit statistics can detect LID to some extent. If 2 items are overly related (a sign of LID), the item fit MNSQ statistics would be much smaller than 1.0 and the ZSTD statistics would be extremely negative (39). In the initial Rasch analysis, seven items were found to have MNSQ smaller than the critical value of 0.6, indicating they were overly related: for example, "grip tube 2.25 cm^3" and "grip tube 1 cm^3", "grasp ball" and "grasp block 5 cm^3". There are other sophisticated procedures to detect LID, (for example ref. 40–42). Because these procedures are beyond the scope of this study and because the item fit statistics did a good job in detecting LID, we did not apply these sophisticated procedures here.

Because the ARAT fit the Mokken scale analysis but not the Rasch analysis, the sum scores of this measure have only ordinal scale properties, rather than interval ones. Some concerns for further applications of the sum scores of the ARAT in clinical and research settings are as follows. First, it cannot be assumed that the same amount of change in scores means the same amount of functional improvement independent of the positions

Table III. *Initial Rasch analysis of the 19 items of the Action Research Arm Test (ARAT)*

Item	Difficulty Logit*	SE Logit	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Pinch ball bearing 3rd finger and thumb	3.59	0.19	$\underline{2.66}$	$\underline{7.2}$	1.32	0.6
Pinch marble 3rd finger and thumb	2.31	0.17	$\underline{1.90}$	$\underline{5.0}$	1.19	0.5
Pinch ball bearing 2nd finger and thumb	1.77	0.17	1.16	1.1	0.6	−0.8
Pour water glass to glass	1.46	0.17	$\underline{1.43}$	$\underline{2.8}$	0.88	−0.2
Grasp block (10 cm^3)	1.32	0.17	$\underline{1.75}$	$\underline{4.5}$	1.16	0.5
Pinch ball bearing 1st finger and thumb	0.99	0.16	$\underline{1.41}$	$\underline{2.7}$	0.9	−0.2
Pinch marble 2nd finger and thumb	0.88	0.16	1.10	0.8	0.68	−1.0
Pinch marble 1st finger and thumb	0.06	0.16	1.02	0.2	0.72	−1.2
Grasp block (7.5 cm^3)	−0.13	0.15	0.63	$\underline{-3.2}$	$\underline{0.43}$	$\underline{-3.0}$
Grip washer over bolt	−0.13	0.15	0.92	−0.6	0.64	−1.6
Grasp ball	−0.38	0.15	$\underline{0.57}$	$\underline{-3.9}$	$\underline{0.36}$	$\underline{-3.6}$
Grip tube (2.25 cm^3)	−0.65	0.15	$\underline{0.42}$	$\underline{-6.0}$	$\underline{0.32}$	$\underline{-3.9}$
Grasp stone	−0.78	0.15	$\underline{0.41}$	$\underline{-6.1}$	$\underline{0.30}$	$\underline{-3.9}$
Grip tube (1 cm^3)	−0.85	0.15	$\underline{0.52}$	$\underline{-4.6}$	$\underline{0.37}$	$\underline{-3.3}$
Grasp block (5 cm^3)	−0.95	0.14	$\underline{0.53}$	$\underline{-4.7}$	$\underline{0.36}$	$\underline{-3.3}$
Place hand behind head	−1.32	0.14	$\underline{1.68}$	$\underline{4.8}$	$\underline{3.48}$	$\underline{5.6}$
Grasp block (2.5 cm^3)	−1.46	0.14	0.6	$\underline{-4.0}$	$\underline{0.48}$	−2.1
Place hand on top of head	−2.41	0.13	1.1	1.0	$\underline{1.71}$	1.4
Hand to mouth	−3.31	0.13	0.82	−1.9	1.03	0.3
Threshold 1[†]	−1.92	0.06	–	–	–	–
Threshold 2	−0.98	0.07	–	–	–	–
Threshold 3	2.90	0.07	–	–	–	–

*Items are arranged in descending order of difficulty logit; Underlining indicates misfit items.
[†]Threshold means difficulty between the adjacent scoring levels. The items of the ARAT have 4 levels of scaling and thus have 3 thresholds.
MNSQ = mean square standardized residual; ZSTD = standardized value.

where score changes are calculated (17). For example, if an individual gains a greater score than on a previous assessment, this can be considered only as showing improvement on his/her functional ability; how much he/she exactly improved would still be unknown. Secondly, score differences between individuals and groups of patients are not necessarily comparable unless they are based on the same evaluation scores initially. For instance, a patient with lower upper extremity function may experience larger numerical gains than a patient with relatively good upper extremity function, but it cannot be concluded that the former patient has improved more than the latter or that the treatment is more effective for those patients with lower upper extremity function. Furthermore, the sum scores of the ARAT should be subjected to non-parametric statistical analysis.

One limitation of this study is that we did not assess the sensory functions of our participants. Sensory deficits may influence their performance, especially on the items involving picking up the small ball bearing with 2 fingers only. Lacking information about sensory function of sample characteristics may compromise the interpretation of the results.

In summary, our findings indicate that the 19-item ARAT constitutes a unidimensional construct measuring upper extremity function in patients after stroke. Except for one item, 18-item ARAT fit the DM model representing a consistently hierarchical order along the upper extremity function's continuum. Since the 19-item ARAT forms a unidimensional structure, this indicates the raw scores of the test can be summed. Thus, clinicians and researchers are recommended to use the 19 items of the ARAT as a whole instead of using them as 4 subscales. However, they should be aware that the raw sum scores of the test are an ordinal scale rather than an interval scale, implying that differences in scores on the ARAT should be interpreted with great care. Further efforts may be needed to revise the ARAT so that the resulting sum scores can be considered as having interval scale properties.

ACKNOWLEDGEMENT

The study was supported by 2 research grants form National Science Council (NSC 93-2314-B-002-028, NSC 93-2314-B-002-284).

REFERENCES

1. Kwakkel G, Kollen BJ, Wagenaar RC. Therapy impact on functional recovery in stroke rehabilitation. Physiotherapy 1999; 85: 377–391.
2. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981; 4: 483–492.
3. Mayo NE, Wood-Dauphinee S, Ahmed S, Gordon C, Higgins J, McEwen S, et al. Disablement following stroke. Disabil Rehabil 1999; 21: 258–268.
4. Wade DT. Measuring arm impairment and disability after stroke. Int Disabil Stud 1989; 11: 89–92.
5. Duncan PW, Lai SM, van Culin V, Huang L, Clausen D, Wallace D. Development of a comprehensive assessment toolbox for stroke. Clin Geriatr Med 1999; 15: 885–915.
6. Croarkin E, Danoff J, Barnes C. Evidence-based rating of upper-extremity motor function tests used for people following a stroke. Phys Ther 2004; 84: 62–74.
7. de Weerdt WJG, Harrison MA. Measuring recovery of arm-hand function in stroke patients: a comparison of the Bruunstrom-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: 107–113.
9. Dekker CL, van Staalduinen AM, Beckerman H, van der Lee JH, Koppe PA, Zondervan RCJ. Concurrent validity of instruments to measure upper extremity performance: the Action Research Arm Test, the Nine-Hole-Peg Test and the Motricity Index [Dutch]. Ned Tijdschr Fysioter 2001; 111: 110–115.
10. van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. J Rehabil Med 2001; 33: 110–113.
11. Hsueh IP, Hsieh CL. Responsiveness of two upper extremity function instruments for stroke inpatients receiving rehabilitation. Clin Rehabil 2002; 16: 617–624.
12. Hsueh IP, Lee MM, Hsieh CL. The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? Clin Rehabil 2002; 16: 382–388.
13. Sodring KM, Bautz-Holter E, Ljunggren AE, Wyller TB. Description and validation of a test of motor function and activities in stroke patients. The Sodring Motor Evaluation of Stroke Patients. Scand J Rehabil Med 1995; 27: 211–217.
14. van der Lee JH, Roorda LD, Beckerman H, Lankhorst GJ, Bouter LM. Improving the Action Research Arm test: a unidimensional hierarchical scale. Clin Rehabil 2002; 16: 646–653.
15. Molenaar IW, Sijtsma K. User's Manual MSP5 for Windows: a program for Mokken Scale analysis for polyromous items. Groningen: lec ProGamma; 2000.
16. Sijtsma K. Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. Appl Psychol Meas 1998; 22: 3–31.
17. Tennant A, Geddes JML, Chamberlain MA. The Barthel Index. An ordinal score or interval level measure? Clin Rehabil 1996; 10: 301–308.
18. Avery LM, Russell DJ, Raina PS, Walter SD, Rosenbaum PL. Rasch analysis of the Gross Motor Function Measure: validating the assumptions of the Rasch model to create an interval-level measure. Arch Phys Med Rehabil 2003; 84: 697–705.
19. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989; 70: 857–860.
20. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.
21. Wright BD, Masters GN. Rating Scale Analysis. Chicago, IL: MESA Press; 1982.
22. Wright BD, Mok M. Rasch models overview. J Appl Meas 2000; 1: 83–106.
23. Wright BD. Logits? Rasch Meas Trans 1993; 7: 288.
24. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Erlbaum; 2001.
25. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley; 1968.
26. van der Heijden PG, van Buuren S, Fekkes M, Radder J, Verrips E. Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF-36. Qual Life Res 2003; 12: 189–198.
27. Carroll D. A quantitative test of upper extremity function. J Chronic Dis 1965; 18: 479–491.
28. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978; 43: 561–573.
29. WINSTEPS [program]. Version 3.51. Chicago, IL: http://www.winsteps.com; 2004.
30. Fox CM, Jones JA. Uses of Rasch modeling in counseling psychology research. J Counsel Psychol 1998; 45: 30–45.
31. van Alphen A, Halfens R, Hasman A, Imbos T. Likert or Rasch? Nothing is more applicable than good theory. J Adv Nurs 1994; 20: 196–201.
32. Smith RM. Polytomous Mean-Square Fit Statistics. Rasch Meas Trans 1996; 10: 516–517.
33. Wright BD, Linacre JM. Reasonable item mean-square fit values. Rasch Meas Trans 1994; 8: 370.
34. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002; 3: 85–106.
35. Nilsson AL, Sunnerhagen KS, Grimby G. Scoring alternatives for FIM in neurological disorders applying Rasch analysis. Acta Neurol Scand 2005; 111: 264–273.
36. Zhu W, Updyke WF, Lewandowski C. Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. J Outcome Meas 1997; 1: 286–304.
37. Verhelst ND, Glas CAW. The one-parameter logistic model. In: Fischer GH, Molenaar IW, eds. Rasch models: foundations, recent developments, and applications. New York: Springer-Verlag; 1995, p. 215–237.
38. One-Parameter Logistic Model OPLM [program]. Arnhem: CITO: National Institute for Educational Measurement, The Netherlands; 1995.
39. Linacre JM, Wright BD. Chi-Square Fit Statistics. Rasch Meas Trans 1994; 8: 350.
40. Yen WM. Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas 1984; 8: 125–145.
41. Tuerlinckx F, De Boeck P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychol Methods 2001; 6: 181–195.
42. Wang WC, Wilson MR. The Rasch testlet model. Appl Psychol Meas 2005; 29: 126–149.