# 行政院國家科學委員會專題研究計畫　成果報告

## 中風病人手功能量表之發展(3/3)

計畫類別：個別型計畫
計畫編號：NSC93-2314-B-002-028-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立臺灣大學醫學院職能治療學系

計畫主持人：謝清麟

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 27 日

說明：目前已完成 2 篇稿件投稿中，其中一篇如下，已接近被接受刊登。另外完整的「中風病人手功能量表」稿件，亦接近完稿。

# Validation of the Action Research Arm Test using item response theory in stroke patients


## Abstract

**Objective:** To validate the unidimensionality of the Action Research Arm Test (ARAT) by Mokken scale analysis and to examine whether the ordinal raw sum scores of ARAT can be transformed to interval scores via the Rasch rating scale model.

**Subjects and Methods:** A total of 351 patients were recruited from five rehabilitation departments located in four regions of Taiwan. The 19-item ARAT was administered to all the subjects by a physical therapist. The data were analyzed by Mokken analysis, followed by Rasch analysis.

**Results:** The results strongly 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 ARAT items form a constant hierarchical scale for all stroke patients. On the other hand, the Rasch analysis, with a stepwise deletion of misfit items, showed that only 4 items (“grasp ball”, “grasp block 5 cm3”, “grasp block 2.5 cm3”, and “grip tube 1 cm”) fit the rating scale model's expectation.

**Conclusions:** Our results provided strong evidence that the ARAT forms a unidimensional scale for patients with stroke. However, the results did not support the premise that the ordinal total raw sum scores of the ARAT can be transformed into interval Rasch scores. Thus, patients can be ranked on their upper extremity functional abilities by the raw sum scores, but equal differences in these scores do not imply equal differences in functional status. The score changes within patients and the score differences between patients can only give us the information about order of the patients, but not represent exact function of them.



**Key Words:** Psychometrics, Cerebrovascular Accident, Arm.

**Journal of Rehabilitation Medicine**



**Correspondence address:** Ching-Lin Hsieh, School of Occupational Therapy, College of Medicine, National Taiwan University, 7, Chung-Shan S. Rd, Taipei 100, Taiwan.

**E-mail address:** mike26@ha.mc.ntu.edu.tw

## Introduction

Upper extremity (UE) dysfunction occurs in approximately 30 to 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 for patients with stroke (3). Accurately measuring the UE 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 UE functional measure for stroke patients is crucial for both clinicians and researchers.

The Action Research Arm Test (ARAT) (2) is a measure widely used in evaluating the UE function of stroke patients. Many researchers have examined its psychometric properties (including intra-/inter-rater reliability, concurrent/convergent validity, and responsiveness) using classical test theory and have found satisfactory results (2, 7-13). However, at least two shortcomings remain in using this measure in clinical and research settings. First, the unidimensional construct of the ARAT has rarely been examined, where the unidimensional construct of a scale indicates whether all items of the scale measure the same construct and is required if one is to justify the summation of scores to quantify characteristics of interest (14). To our knowledge, only van der Lee and her coworkers had examined this property of the ARAT using Mokken scale analysis (15) and found that the measure comprised a unidimensional scale (16). Mokken scale analysis is a nonparametric modern item response theory (IRT) model that examines accuracy of ordering between persons' raw sum scores on a measure to determine unidimensionality (15, 17). However, since Mokken analysis typically requires a sample size larger than 200 to reliably estimate the unidimensionality of a scale (15), the sample size of 63 subjects in the study by van der Lee et al. appears to have been too small. Furthermore, their sample could not be considered as representative of the total stroke population because neither slightly impaired nor severely impaired patients were included in their sample. Therefore, the results of their study do not provide conclusive evidence supporting the unidimensionality of the ARAT.

Second, with the Mokken scaling analysis, the raw sum score of the ARAT attains only the status of an ordinal score, even if the unidimensionality of the ARAT has been verified. This means that a given difference in raw sum scores at one end of on the scale does not necessarily represent the same amount of functional change for an identical difference at the other end of the scale (18). For example, suppose that a patient, in two subsequent evaluation periods, gained 5 points of progress (e.g. 5 to 10) and then 10 points of progress (e.g. 10 to 20) on the ARAT. It would be tempting to interpret these score changes to mean that at the second evaluation the patient's UE function improved by twice as much as that at the first evaluation. But, it is not necessarily so because these scores are ordinal. Interval scores, on the contrary, represent an underlying trait in which equal intervals between any two points on a scale are of equal value. The interval property maintains the numerical meaning of score gains from a scale and allows the scores to serve beyond being just categories on an ordinal scale. Therefore, clinicians and researchers can know exactly how much functional ability patients have gained or how a certain two patients with different scores differ from others in their functional status. In the above example, 10 points of progress on an interval scale would indicate a doubling of the gain of 5 points in the UE function of the patient. Furthermore, an interval measure can be analyzed by parametric statistics, which are often more powerful than non-parametric methods (19). Therefore, an interval-scale measure would enable clinicians and researchers to numerically quantify UE 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 measures (20).

To determine whether the ARAT sum score show an interval scale, the Rasch analysis was conducted. The Rasch analysis is a technique to establish the interval scale property of a measuring instrument (21). Items that fit the Rasch model's expectations can be used to generate logit scores and can be viewed as interval scores (22, 23). The purposes of this study were to validate the unidimensionality of the ARAT by Mokken analysis with a large sample and to examine whether the ARAT fits the Rasch model's expectation, thus producing interval scores.

## Methods

### Subjects

To select stroke patients with a broad range of UE 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: (1) diagnosis (International Classification of Diseases, Ninth Revision Clinical Modification [ICD-9-CM] codes) of cerebral hemorrhage (431) or cerebral infarction (434), (2) ability to follow instructions, and (3) absence of other major diseases (e.g., tumors or arthritis) or impairments (e.g., amputations or fractures) that would reduce or limit patients' ability to perform UE 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 ethical review boards.

### Procedure

The ARAT was administered by the same physical therapist to the patients at the 5 rehabilitation departments. Patients' demographic details and data on comorbidity were collected from their medical records.

### Instrument

The ARAT, developed by Lyle(2), is based on the UE function test of Carroll.(24) It is designed to assess the recovery of UE 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- can not 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 UE 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 three assumptions: (1) items form a unidimensional scale (measuring the same construct, e.g., UE function); (2) 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 UE function); and (3) the item characteristic curve (ICC) for each item is a monotonically nondecreasing function of the underlying construct, which means that patients at a higher level of UE 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 Hi 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 UE 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 strongly supported if $H \geq 0.50$ (15). Second, the double monotonicity (DM) model (15) (in additional to the three assumptions of the MH model, the DM model assumes also that the ICCs of the scale do not intersect) was used to test whether the items of the ARAT possessed an invariant hierarchical ordering, which means that the difficulty ordering 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 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 two criteria values: “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 (25) was employed using the WINSTEPS program (26). In addition to the four assumptions of the Mokken analysis, the Rasch model requires a one-parametric functional form for the ICCs; that is, all ICCs have the same slope and differ only in item difficulty (27, 28). The same slope means the same value of the slope which is the average discrimination of all the items (26). 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 behavior affecting responses to items near the person’s functional ability in UE function; the outfit MNSQ is sensitive to unexpected behavior by persons on items far from the level (22, 23). The MNSQ value can be transformed to a t statistic, termed the standardized Z value, which follows approximately the t, or standard normal distribution, when the items fit the model’s expectation. The misfit criteria in this study were predefined as follows: (22, 29) (1) infit ZSTD > 1.96 and MNSQ > 1.4 or outfit ZSTD > 1.96 and MNSQ > 1.4; and (2) infit ZSTD < -1.96 and MNSQ < 0.6 or outfit ZSTD < -1.96 and MNSQ < 0.6. A MNSQ value more than 1.4 indicates 40% greater variation in the observed data than the Rasch model predicted, suggesting either that the item does not belong with the other items on the same continuum or that there are problems in item definition. A MNSQ value of less than 0.6 indicates 40% less variation in the observed response pattern than was modeled; that is, the item fails to discriminate individuals with different abilities or the item is redundant with other items that measure a similar amount of challenge (30). Items considered to misfit to the Rasch model were removed in a stepwise manner by inspecting a series of infit to outfit statistics.

## Results

A total of 351 patients were recruited in the study. The characteristics of the subjects are presented in Table 1. The participants had a wide range of UE function deficits, and their sum scores of the ARAT were scattered throughout the full range of possible scores (0-57).

Table 2 shows that the range of scalability coefficient Hi of each item of the ARAT fell 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.

Because the parameters and fit statistics of the Rasch analysis depend on the nature of the items as well the number of items on a scale are included, the misfit items are generally removed in a stepwise manner. Those parameters and fit statistics in Table 3 were tentative to give a general impression of the fit of the ARAT to the Rasch model. Twelve of the ARAT

items did not fit the Rasch model's expectations (infit or outfit ZSTD > 1.96 and MNSQ >1.4; or infit or outfit ZSTD < -1.96 and MNSQ < 0.6). After deleting misfit items stepwise according to the preset criteria, only four items ("grasp ball", "grasp block 5 cm3", "grasp block 2.5 cm3", and "grip tube 1 cm") were found to fit the model's expectation.

## Discussion

This study was the first to use both a non-parametric Mokken analysis and a parametric Rasch analysis to examine 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, excepting for one item in the DM model. This result indicated that the 18-item ARAT can be considered a unidimensional hierarchical ordering measure. 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.

The property of unidimensionality is fundamental to a measure and forms the key component of content validity; (27) that is, we can know what we are measuring. The current study, using a sample size capable of yielding a reliable estimate of the unidimensional construct of the ARAT, demonstrated that the ARAT indeed measures the UE function as such a construct. The unidimensional scale of the ARAT justifies the use of the sum scores of this measure. Our results also confirm van der Lee et al's suggestion that it is inappropriate to divide the 19-item ARAT into the four subscales proposed by Lyle, (2) as it was found to be a unidimensional scale.

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 van der Lee et al's study, indicating that the difficulty ordering of this item varied from the other items and should be removed. However, the other three misfitting items, "pinch marble 3rd finger and thumb", "pinch ball bearing 2nd finger and thumb", and "pinch ball bearing 1st finger and thumb", to the DM model found in van der Lee et al's study were not found to deviate from the DM model's expectation in the current study (16). 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 were not included patients with severe UE dysfunction (i.e. ARAT < 5) and patients with mild UE dysfunction (i.e. ARAT > 51). In particular, "pinch marble 3rd finger and thumb" and "pinch ball bearing 2nd finger and thumb" were the two most difficult items which fit in our study but misfitted in van der Lee et al's study. Thus, the presence of subjects with mild UE dysfunction (i.e. ARAT > 51) in the sample of this study may have caused the differences in the results of the studies.

The poor data-Rasch model fit suggests that the current items of the ARAT cannot meet the parametric form assumption of the Rasch model. These results were not unique to our study. Cook et al (31) found 4 shoulder function scales that misfit the Rasch model. In another study, 3 functional assessments (the Fibromyalgia Impact Scale, the Health Assessment Questionnaire, and the Medical Outcome Survey Short Form (SF-36)) were all found to fail to meet the assumptions of the Rasch model (32). These other well known and used measures have also failed to stand up to the Rasch analysis. Thus, the results support previous arguments that the Rasch model is a stringent model for tests with parametric properties (14, 27). For the aforementioned advantages of interval scores in this study, it is valuable to establish interval level assessments for research and clinical settings. Researchers who are interested in constructing an interval level measure of UE function may base their work on the 4 remaining items ("grasp ball", "grasp block 5cm3", "grasp block 2.5cm3", and "grip tube 1cm") to revise the items of the ARAT. However, it is expected that they will have

to devote great deal of resources and time to construct items that can fit a parametric IRT model and generate interval scores.

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 where score changes are calculated, nor can the differences between scores be directly compared within individuals or between a group of patients (18). For example, clinicians and researchers may find that treatment A results in twice of the ARAT total score gain of treatment B in an individual or in a group of patients. Although the numerical value of the former is twice that of the latter, all that can be concluded is that treatment A has “greater” effectiveness than treatment B, not that it is twice as effective. Second, 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 UE function may experience larger numerical gains than a patient with relatively good UE 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 UE function. Furthermore, the sum scores of the ARAT should be subjected to non-parametric statistical analysis.

In summary, our results provide strong evidence that the ARAT is a unidimensional hierarchical scale for patients with stroke, excepting one item. Since the 19-item ARAT forms a unidimensional structure, this indicates the raw scores of the test can be summed. Thus, it is the scale that is being recommended to the clinicians or researchers are recommended to use the 19-item ARAT. However, they should be aware 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 the resulting sum scores can be considered as having interval scale properties.

## Acknowledgement

The study was supported by two 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. Dromerick AW, Edwards DF, Hahn M. Does the application of constraint-induced movement therapy during acute rehabilitation reduce arm impairment after ischemic stroke? Stroke 2000, 31:2984-2988.
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. van der Lee JH, Snels IAK, Beckerman H, Lankhorst GJ, Wagenaar RC, Bouter LM. Exercise therapy for arm function in stroke patients: a systematic review of randomized controlled trials. Clin Rehabil 2001, 15:20-31.
12. Hsueh IP, Hsieh CL. Responsiveness of two upper extremity function instruments for stroke inpatients receiving rehabilitation. Clin Rehabil 2002, 16:617-624.
13. 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.
14. 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.
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. 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.
17. Sijtsma K. Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. Appl psychol meas 1998, 22:3-31.
18. Tennant A, Geddes JML, Chamberlain MA. The Barthel Index. An ordinal score or interval level measure? Clin Rehabil 1996, 10:301-308.
19. 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.
20. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989, 70:857-860.
21. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.
22. Wright BD, Masters GN. Rating Scale Analysis. Chicago, IL: MESA Press, 1982.
23. Wright BD, Mok M. Rasch models overview. J Appl Meas 2000, 1:83-106.
24. Carroll D. A Quantitative Test of Upper Extremity Function. J Chronic Dis 1965, 18:479-491.
25. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978, 43:561-573.
26. WINSTEPS [program]. Version 3.51. Chicago, IL: http://www.winsteps.com, 2004.
27. 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.
28. 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.
29. Wright BD, Linacre JM. Reasonable item mean-square fit values. Rasch Meas Trans 1994, 8:370.
30. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Erlbaum, 2001.

31. Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. Arch Phys Med Rehabil 2001, 82:1558-1565.
32. Wolfe F, Hawley DJ, Goldenberg DL, Russell IJ, Buskila D, Neumann L. The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. J Rheumatol 2000, 27:1989-1999.

Table 1. Characteristics of the stroke patients (n=351)

| Characteristic | |
|---|---|
| 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 hemorrhage | 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%) |

Table 2. The Mokken scale analysis of the ARAT

| 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 |

*Items are arranged in ascending order of mean, indicating item difficulty from high to low.
†Values of items with violations smaller than the minimum criteria of MSP 5.0 were not shown.
‡Item that showed violation ordering (Pmatrix > 80).

Table 3. The initial Rasch analysis of the 19 items of the 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 | 2.66 | 7.2 | 1.32 | 0.6 |
| Pinch marble 3rd finger and thumb | 2.31 | 0.17 | 1.90 | 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 | 1.43 | 2.8 | 0.88 | -0.2 |
| Grasp block (10 $cm^3$) | 1.32 | 0.17 | 1.75 | 4.5 | 1.16 | 0.5 |
| Pinch ball bearing 1st finger and thumb | 0.99 | 0.16 | 1.41 | 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 | -3.2 | 0.43 | -3.0 |
| Grip washer over bolt | -0.13 | 0.15 | 0.92 | -0.6 | 0.64 | -1.6 |
| Grasp ball | -0.38 | 0.15 | 0.57 | -3.9 | 0.36 | -3.6 |
| Grip tube (2.25 $cm^3$) | -0.65 | 0.15 | 0.42 | -6.0 | 0.32 | -3.9 |
| Grasp stone | -0.78 | 0.15 | 0.41 | -6.1 | 0.30 | -3.9 |
| Grip tube (1 $cm^3$) | -0.85 | 0.15 | 0.52 | -4.6 | 0.37 | -3.3 |
| Grasp block (5 $cm^3$) | -0.95 | 0.14 | 0.53 | -4.7 | 0.36 | -3.3 |
| Place hand behind head | -1.32 | 0.14 | 1.68 | 4.8 | 3.48 | 5.6 |
| Grasp block (2.5 $cm^3$) | -1.46 | 0.14 | 0.6 | -4.0 | 0.48 | -2.1 |
| Place hand on top of head | -2.41 | 0.13 | 1.1 | 1.0 | 1.71 | 1.4 |
| Hand to mouth | -3.31 | 0.13 | 0.82 | -1.9 | 1.03 | 0.3 |

* Items are arranged in descending order of difficulty logit
The figures underlined indicated misfit items.