J Rehabil Med 2010; 42: 694–696

**SHORT COMMUNICATION**

# A COMPARISON OF TWO VALIDATED TESTS FOR UPPER LIMB FUNCTION AFTER STROKE: THE WOLF MOTOR FUNCTION TEST AND THE ACTION RESEARCH ARM TEST

Rinske Nijland, MSc[1], Erwin van Wegen, PhD[1], Jeanine Verbunt, PhD[3,4], Renske van Wijk, MSc[3], Joost van Kordelaar, MSc[1] and Gert Kwakkel, PhD[1,2]

*From the [1]Department of Rehabilitation Medicine, Research Institute MOVE, VU University Medical Centre Amsterdam, [2]Department Rehabilitation Medicine, Rudolf Magnus Institute of Neuroscience, University Medical Centre, Utrecht, [3]Adelante Center of Expertise in Rehabilitation and Audiology, Hoensbroek and [4]Research School CAPHRI, Maastricht University, Maastricht, The Netherlands*

***Objective:*** **To investigate the concurrent validity between the Action Research Arm Test (ARAT) and the Wolf Motor Function Test (WMFT) and to compare their reproducibility, internal consistency and floor and ceiling effects in the same sample of stroke patients.**

***Methods:*** **Forty patients participated in this study. Concurrent validity was determined with Spearman's rank correlation coefficients. Reproducibility was assessed with intraclass correlation coefficients (ICCs) and Bland-Altman plots, internal consistency by means of Cronbach's alphas, and floor and ceiling effects were considered to be present if more than 20% of patients fell outside a preliminary set lower and upper boundary.**

***Results:*** **Spearman's rank correlation coefficients ranged from 0.70 to 0.86. ICCs for inter-rater and intra-rater reliability ranged from 0.92 to 0.97. Bland-Altman plots showed a less stable way of scoring for the WMFT, compared with the ARAT. Cronbach's alpha was > 0.98 for both scales. No floor and ceiling effects were found.**

***Conclusion:*** **The present study showed good clinimetric properties for both assessments. The high concurrent validity suggests that ARAT and WMFT have significant overlap with regard to the underlying construct that is being measured.**

*Key words:* stroke; rehabilitation; upper extremity; outcome measure.

J Rehabil Med 2010; 42: 694–696

*Correspondence address: Erwin van Wegen, Department of Rehabilitation Medicine, VU University Medical Centre, Boelelaan 1117, NL-1081 HV Amsterdam, The Netherlands. E-mail: e.vanwegen@vumc.nl*

Submitted October 5, 2009; accepted February 25, 2010

## INTRODUCTION

A large number of assessments for upper extremity (UE) function after stroke have been published in the literature (1). However, they vary considerably in their focus, and a golden standard is lacking. As a consequence, the selection of a proper instrument is a complex process.

Despite the multitude of assessments for UE function after stroke, only a few valid and reliable clinical measurement tools are available to quantify UE function during the performance of unilateral motor tasks. The Action Research Arm Test (ARAT) (2, 3) and the Wolf Motor Function Test (WMFT) (4) both aim to do this. The clinimetric properties of both measures have been well established (3, 5–7). However, some limitations have also been described. For example, it has been suggested that there is an ambiguity in the way in which performance could be scored on the ARAT, which might lead to an important source of uncontrolled variation between observers or between clinical centres (8). Standardized guidelines need to be applied to reduce this variation (9). In addition, both instruments are often criticized regarding item redundancy (10, 11) and the presence of floor and ceiling effects (5). Finally, although both instruments attempt to provide an objective measure of UE function, it is unclear whether they can distinguish between restitution of function and compensation (12).

In order to provide clinicians and researchers a basis for the selection of an assessment instrument, the purpose of this study was to investigate the concurrent validity of ARAT and WMFT and to compare the reproducibility, internal consistency and floor and ceiling effects of both instruments in the same group of stroke patients. The outcomes measured with the ARAT are expected to be consistent with the WMFT outcomes.

## METHODS

### *Subjects*

Forty patients diagnosed with stroke were recruited from 2 rehabilitation centres in the Netherlands. Inclusion criteria were: (*i*) hemiparesis of the UE, with at least some voluntary muscle contraction (Medical Research Council (MRC) score ≥ 1); (*ii*) no severe deficits in communication, memory and understanding (Mini Mental State Examination (MMSE) score > 22); (*iii*) absence of orthopaedic UE limitations. The protocol was approved by the local ethics committees, and all patients gave written informed consent.

### *Outcome measures*

The ARAT consists of 19 tasks, which are categorized into 4 domains (grasp, grip, pinch and gross movements) (2). Quality of movement is scored on a 4-point scale (0 = can perform no part of the test, 3 = per-

forms test normally). The standardized method for scoring, developed by Yozbatiran et al. was used (9).

The WMFT consists of 17 items (6 joint-segment movements, 9 integrative functional movements and 2 strength items). Performance time of every item is measured between a precisely defined start- and end-point for each task with a maximum of 120 s. The WMFT also contains a 6-point Functional Ability Scale (FAS) that rates the quality of movement and has values ranging from 0 (no attempt made to use the more affected UE) to 5 (movement appears to be normal) (4).

*Procedures*

The patients recruited from the first rehabilitation centre ($n = 18$) participated in testing the reproducibility of both ARAT and WMFT. Both observers applied the measurements within one week to minimize the effect of spontaneous recovery. The subjects were assessed by both observers in random order. For intra-rater reliability, the same sample of 18 was observed twice by one observer, approximately 10 days apart.

Data from all 40 subjects were used to investigate internal consistency, concurrent validity and floor and ceiling effects of the WMFT and the ARAT. All assessments were executed by a trained observer in random order. To prevent the influence of fatigue on the results, a minimum break of 30 min physical rest was taken between the two tests.

*Statistical analysis*

Reproducibility was assessed by means of reliability and agreement (13). The inter- and intra-rater reliability for the total scores of both measures was analysed with the intraclass correlation coefficient (ICC). For the inter-rater reliability a 2-way random effects model with absolute agreement definition was used (14). The intra-rater reliability was determined by applying a 2-way mixed effects model with absolute agreement definition (14). ICCs were interpreted according to the classification of Fleiss (15). Agreement was assessed by means of the limits of agreement using the Bland & Altman method (16). Cronbach's alpha with corresponding confidence intervals (CI) were calculated to determine the internal consistency between the items of each scale. A Cronbach's alpha between 0.70 and 0.95 was considered satisfactory (17). Floor and ceiling effects were defined by means of the percentage of the subjects who scored beyond the lower or upper bound, respectively, of the total possible score. Cut-offs for floor and ceiling effects were set at 5% of the total score. As a consequence, scores below 3 points and scores above 54 points on the ARAT were determined as a floor and/or ceiling effect, respectively. For the WMFT the cut-off points were below 4 points and above 71 points, respectively. Significant floor and ceiling effects were considered to be present if more than 20% of the patients fell outside the lower or upper bound, respectively (18).

To determine the concurrent validity, Spearman's rank correlation coefficients ($r_s$) were calculated between the ARAT total score and the WMFT score, which was split into 4 variables: FAS score, median time score, item 7 and 14 (strength). A correlation coefficient $\geq 0.7$ was considered to reflect high concurrent validity (19). All tests were applied 2-tailed with a level of significance of 0.05.

## RESULTS

Patient characteristics are summarized in Table I. ICCs for inter-rater reliability of the ARAT and the WMFT were 0.92 and 0.94, respectively. The intra-rater ICCs were 0.97 and 0.95 for the ARAT and the WMFT, respectively.

The Bland-Altman plots (Fig. 1) showed higher limits of agreement in the between-observer plot of both assessments, suggesting a lower agreement between observers than within an observer. The within-observer plots reflect a less stable way of scoring for the WMFT, compared with the ARAT. Cronbach's alpha for the ARAT and the WMFT FAS were 0.985 (CI: 0.977–0.991) and 0.982 (CI: 0.972–0.989), respectively. No significant floor and ceiling effects were found. On both tests, approximately 17% of the patients scored beyond the upper 5% limits. Below the lower 5% limits was scored by 12.5% and by 5% on the ARAT and WMFT respectively.

The spearman correlation coefficient ($r_s$) between the ARAT total score and the total WMFT FAS was 0.86 ($p < 0.01$) and between the ARAT score and the WMFT median time score was –0.89 ($p < 0.01$). Finally, both strength tasks of the WMFT (i.e. items 7 and 14) showed a correlation coefficient $r_s$ of 0.70 ($p < 0.01$) with the ARAT.

## DISCUSSION

The main findings of the present study are that both assessments show excellent inter- and intra-observer reliability and are highly correlated with each other. However, the Bland-Altman plots showed that the between-observer agreement of both instruments was lower than the within-observer agreement, confirming that there still might be an ambiguity in the way in which performance could be scored and that proper training of observers is important for uniform application of the tests. Additionally, the within-observer plots showed a less stable way of scoring for the WMFT, suggesting a relatively higher measurement error for the WMFT.

Our results showed Cronbach's alpha's of 0.98 or higher for both scales, which is consistent with previous findings of high internal consistency (6, 10). This suggests that both assessments measure a single, unidimensional construct. However, an alpha-score of 0.98 could also suggest item redundancy. A challenge for future studies is to include more stroke patients in order to determine whether the number of items of each test can be reduced following an item-response theory model and to further investigate the dimensionality of both tests.

In contrast to claims from the literature (5, 10), floor and ceiling effects were not found in either instrument. This may have been caused by the fact we included mainly patients with mild to moderate hemiparesis. However, for this group, the evaluation and quantification of upper limb function is most relevant.

Table I. *Patient characteristics* (n = 40)

| Characteristic | |
|---|---|
| Sex, M/F, *n* | 23/17 |
| Age, years, mean (SD) [range] | 60.0 (13.6) [31–82] |
| Side of hemiplegia, L/R, *n* | 23/17 |
| MRC score (0–5), median (IQR) | 4.0 (4–5) |
| Time since stroke onset in years, median (IQR) | 0.41 (0.25–0.77) |
| ARAT total score, median (IQR) | 38 (22–46) |
| WMFT median time in s, median (IQR) | 3.29 (2.31–5.91) |
| WMFT FAS, median (IQR) | 53 (32.75–67.75) |

M/F: male/female; SD: standard deviation; L/R: left/right; MRC: Medical Research Council (muscle power); MMSE, Mini Mental State Examination; ARAT: Action Research Arm Test; WMFT: Wolf Motor Function Test; FAS: Functional Ability Score; IQR: interquartile ranges.

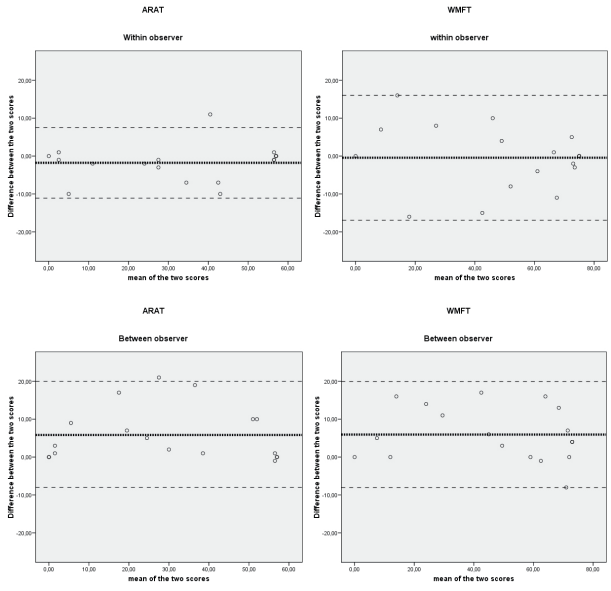


*Fig. 1.* Graphic representation according to the Bland-Altman technique. The dashed bold lines represent the mean difference score. The dashed lines represent the limits of agreement (mean ± 1.96 × the standard deviation of the difference score). ARAT: Action Research Arm Test total score; WMFT: Wolf Motor Function Test total functional ability score.

The high concurrent validity between both tests suggests that ARAT and WMFT have significant overlap with regard to the underlying construct that is being measured. Unfortunately, based on our results no direct insight into the nature of the underlying construct that both assessments are assumed to quantify can be given. However, because of the large number of instruments, knowledge concerning their underlying construct is needed in order to compare them and to classify them meaningfully. The International Classification of Functioning, Disability and Health (ICF) (20) can facilitate classification of a measurement instrument in what it does and does not intend to assess, by making a distinction between the domains of Body Functions and Structure, Activity and Participation. The ARAT and the WMFT can be distinguished from most other tests since they both intend to assess unilateral performance on functional tasks as well as gross movements of the upper paretic limb. However, to determine exactly what both assessments measure, we need to improve our understanding of the required motor performance and coordination to execute items on the WMFT and ARAT. Future studies should implement electromyography and kinematic analysis in order to distinguish between restitution of function and the use of compensation strategies (12, 21). Monitoring of parallel changes in test scores and actual (kinematic) performance in a longitudinal manner will shed light on what actually changes during functional recovery (21).

Some limitations of the present study should be noted. First, this study was based on a modest sample size. Secondly, only patients with mild to moderate disease severity were included in this study. This obviously limits the generalization of the present findings to other patients with different characteristics.

## ACKNOWLEDGEMENTS

The authors want to thank Han Franck for his cooperation in the preparation phase of the study. This study is co-financed by the EXPLICIT stroke programme of ZonMw (grant number 89000001).

## REFERENCES

1. Ashford S, Slade M, Malaprade F, Turner-Stokes L. Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review. J Rehabil Med 2008; 40: 787–795.
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. van der Lee JH, De G 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: 14–19.
4. Wolf SL, Thompson PA, Morris DM, Rose DK, Winstein CJ, Taub E, et al. The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. Neurorehabil Neural Repair 2005; 19: 194–205.
5. Lin JH, Hsu MJ, Sheu CF, Wu TS, Lin RT, Chen CH, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys Ther 2009; 89: 840–850.
6. Morris DM, Uswatte G, Crago JE, Cook EW, III, Taub E. The reliability of the wolf motor function test for assessing upper extremity function after stroke. Arch Phys Med Rehabil 2001; 82: 750–755.
7. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. Stroke 2001; 32: 1635–1639.
8. Donaldson C, Tallis R, Pomeroy V. Outcome measures in neurophysiotherapy for the arm and hand: have we lost our grip? Clin Rehabil 2006; 20: 459–460.
9. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. Neurorehabil Neural Repair 2008; 22: 78–90.
10. 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.
11. Bogard K, Wolf S, Zhang Q, Thompson P, Morris D, Nichols-Larsen D. Can the Wolf Motor Function Test be streamlined? Neurorehabil Neural Repair 2009; 23: 422–428.
12. Levin MF, Kleim JA, Wolf SL. What do motor "recovery" and "compensation" mean in patients following stroke? Neurorehabil Neural Repair 2009; 23: 313–319.
13. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; 59: 1033–1039.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–428.
15. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–310.
17. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34–42.
18. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J. Issues for selection of outcome measures in stroke rehabilitation: ICF Body Functions. Disabil Rehabil 2005; 27: 191–207.
19. Munro BH. Statistical methods for health care research. 5th edn. Philadelphia: Lippincott Williams & Wilkins; 2005.
20. World Health Organization (WHO). International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
21. Kwakkel G, Meskers CG, van Wegen EE, Lankhorst GJ, Geurts AC, van Kuijk AA, et al. Impact of early applied upper limb stimulation: the EXPLICIT-stroke programme design. BMC Neurol 2008; 8: 49.