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.

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 below 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 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) 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 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)

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

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.
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