Outcome Measurement in Stroke
A Scale Selection Strategy

Karen Baker, MSc, MCSP; Stefan J. Cano, PhD, CPsychol; E. Diane Playford, MD, FRCP

Background and Purpose—Evaluating the impact of new treatments requires the use of reliable, valid, and responsive outcome measures. However, given the wide range of instruments currently available, it is not always straightforward for healthcare professionals to select the most appropriate tool. In this review, we propose a potential approach to scale selection.

Methods—In designing a new study of the impact of a robotic device in stroke rehabilitation, we developed a three-stage scale selection strategy. First, two guidance documents (Medical Outcome Trust and Food and Drug Administration PRO Guidance) were reviewed to identify key scale assessment criteria. Second, consideration was given at a theoretical level of the concepts and domains relevant to the goals our study. Third, a comprehensive literature search strategy and review were developed in conjunction with healthcare professionals and psychometricians. Identified scales were appraised regarding their psychometric properties and clinical content.

Results—Forty-five measures were initially identified and appraised. From a clinical content perspective, none of the measures were considered to be sufficient on their own to capture all the important outcome domains in this study. However, 3 measures were identified that best met our review criteria: Stroke Rehabilitation Assessment of Movement, Chedoke Arm and Hand Inventory, and ABILHAND. After the final stage of scale appraisal, two further upper limb scales (Fugl-Meyer and Action Research Arm Test) were included based on clinical content and study design issues.

Conclusions—Our three-stage review process appears to be a potentially useful approach for evidence-based scale selection in stroke rehabilitation studies. (Stroke. 2011;42:1787-1794.)

Key Words: cerebrovascular disorders ● outcomes ● rehabilitation ● scales ● upper extremity

Stroke is currently the single largest cause of adult disability in the United Kingdom with one third of people who have had a stroke left with long-term disability. New therapies such as repetitive practice through the use of robotics have the potential to enhance the recovery of neurological function post-stroke. The evaluation of such treatments increasingly depends on the use of valid, reliable, and responsive outcome measures. Therefore, selecting the right measure for the right study is essential and rests on a clear understanding of the scientific quality of the rating scales.

Systemic reviews can help to select outcome measures. For example, Ashford et al reviewed measures for the hemiparetic upper limb, identified 6 scales that met their selection criteria, but concluded that currently there was no single reliable and valid measure available to capture the full range of functional tasks in the hemiparetic upper limb. This type of systematic review provides invaluable information but also raises important questions. For example, how well targeted are different measures to the goals of specific interventions? How do we select the best available measures for future clinical studies? What criteria should be used and why?

The aim of this article is to present a scale selection strategy for evidence-based scale selection in stroke research. We describe three steps. First, we reviewed two recent psychometric guidance documents that define key scale assessment criteria. Second, we considered, at a theoretical level, clinical issues, concepts, and domains important to include in stroke outcome research. Third, we performed a comprehensive literature search strategy and review with input from healthcare professionals, psychometricians, and librarians. Having these three building blocks in place, we selected the most scientifically sound and clinically relevant outcome measures for a study examining a robotic device designed to improve upper limb function after stroke.

Step I: Psychometric Guidance Documents
We selected the two most widely used guideline documents for psychometric standards for rating scale research...
to provide appropriate criteria against which to examine existing scales: (1) the Scientific Advisory Committee of the Medical Outcome Trust (MOT) guidance for the development and validation of health-related quality of life instruments; and (2) the recently finalized US Food and Drug Administration (FDA) guidelines for PRO measures in clinical trials.

PRO measures are patient-derived questionnaires that aim to quantify any aspect of a patient’s health status ranging from symptoms to other complex concepts such as quality of life. Both the MOT and FDA documents emphasize the importance of patients’ views in clinical research. Both documents identify key properties for psychometrically robust measures. Tables 1 and 2 summarize the key psychometric issues identified by these documents. Taken together these documents provide an essential basis in the process of selecting PRO measures, because they provide rigorous standards that scales should meet. It is important to note that these properties of a PRO measure also have relevance to clinician-rated scales, which should be evaluated in the same rigorous manner.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual and measurement model</td>
<td>Rationale for and depiction of the concept(s) the measure is intended to assess (eg, scale and subscale structure)</td>
</tr>
<tr>
<td>Reliability</td>
<td>Degree to which the measure is free from random error, including internal consistency and test–retest reproducibility</td>
</tr>
<tr>
<td>Validity</td>
<td>Degree to which instrument measures what it purports to measure, including content, construct, and criterion validity</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Instrument’s ability to detect change in outcomes that matter to persons with a health condition, their significant others, or their providers</td>
</tr>
<tr>
<td>Interpretability</td>
<td>Degree to which one can assign clinical or commonly understood meaning to quantitative scores</td>
</tr>
<tr>
<td>Respondent and administrative burden</td>
<td>Time, energy, and other difficulty placed on respondents and on those who administer the instrument</td>
</tr>
<tr>
<td>Alternative forms</td>
<td>Equivalent information on versions of measures other than original source</td>
</tr>
<tr>
<td>Cultural and language adaptations (translations).</td>
<td>Conceptual and linguistic equivalence and psychometric properties of cross-cultural adaptation</td>
</tr>
</tbody>
</table>

MOT indicates Medical Outcome Trust.

Step II: Stroke-Specific Issues in the Selection of Appropriate Outcome Measures

There is no consensus on the battery of outcome measure to use when assessing physical recovery post-stroke. Selecting the appropriate scales to assess recovery in stroke is a difficult task given the heterogeneity of stroke etiology, symptoms, severity, and even recovery itself. However, despite these complexities, several clinically anchored strategies can assist in selecting the right measure for this population in clinical practice.

<table>
<thead>
<tr>
<th>Psychometric Property</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data quality</td>
<td>Assessed by completeness of data and score distributions</td>
</tr>
<tr>
<td>Scaling assumptions</td>
<td>Items in a scale should measure a common underlying construct</td>
</tr>
<tr>
<td></td>
<td>Items in a scale should contain a similar proportion of information concerning the construct being measured</td>
</tr>
<tr>
<td></td>
<td>Items should be correctly grouped into scales</td>
</tr>
<tr>
<td>Targeting</td>
<td>Indicates whether a scale is acceptable as a measure for the sample</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
</tr>
<tr>
<td>Internal consistency</td>
<td>The extent to which items in a scale measure the same construct</td>
</tr>
<tr>
<td>Test–retest reproducibility</td>
<td>The stability of a scale between repeat administrations of the scale on two occasions</td>
</tr>
<tr>
<td>Validity</td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td>The extent to which the content of a scale is representative of the conceptual domain it is intended to cover</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Evidence that the scale measures a single construct, that items can be combined to form a summary score, and that subscales measure distinct but related constructs</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The ability of a scale to detect clinically significant change</td>
</tr>
</tbody>
</table>

FDA indicates Food and Drug Administration.
trial research and practice. Thus, in addition to the psychometric guidelines described above, we included the International Classification of Functioning, Disability and Health (ICF) framework\(^\text{15}\) to help identify scales with relevant domains for our study.

The ICF framework\(^\text{15}\) provides a conceptual framework for the selection and classification of outcome measures.\(^\text{16–18}\) The domains contained in the ICF include Body Functions and Structures (impairments) and Activities and Participation (disabilities). In the context of stroke research, it may be important for researchers to know what impact an intervention has had at an impairment level, but it is equally important to identify what impact these changes have for individuals at an activity or participation level and how this affects more complex multidimensional concepts such as quality of life.\(^\text{19}\) Novel treatments need refining until their impact at the level of body structure and function translates into something meaningful for the patient.

The use of robotic aids as an adjunct to therapy to increase intensity of repetitive arm movement is a promising novel intervention.\(^\text{3}\) We are currently conducting a feasibility study of a robotic device, which targets reach, pronation/supination, and grasp to assist upper limb recovery in the acute phase after stroke of all etiologies. Patients ranging from those with only those with only a few proximal flickers to those with near normal strength can use this device. To appropriately assess this new potential therapy, we require scales that capture clinically important improvements in arm movements at an impairment level and activity.

**Step III: Literature Review**

A comprehensive literature search strategy and review were developed in conjunction with healthcare professionals, psychometricians, and librarians. An electronic bibliographic search was conducted in the following databases: Medline, Embase (Excerpta Medica), CINAHL, and PsycINFO. The databases were searched from 1966 to the present. Limits were placed on each search to exclude non-English citations and nonhuman subjects using a variety of key terms, including upper limb, upper extremity, arm function, outcome measure, stroke, cerebral vascular accident, assessment, scale, score, quality of life, and questionnaire.

All instruments included in the review were identified as an upper limb outcome measure. A follow-up review of references was performed to find relevant articles not detected in the electronic searches. From this search, scales that measure global motor function but also included a specific upper limb subsection were included (eg, the Rivermead Motor Assessment). However, scales in which upper limb function was not separate from other functions were excluded (eg, the Stroke Impact Scale). Scales that had not been psychometrically evaluated in patients with stroke were also excluded. The Figure illustrates this search.

**Results**

The findings from the literature review were cross-referenced against the guideline documents criteria and clinical considerations underpinned by the ICF framework. The review itself identified 25 outcome measures used to evaluate upper limb recovery post-stroke. These measures were separated into stroke-specific clinician-rated and PRO measures. The properties and initial evaluation of these 25 measures against the specific criteria identified on the MOT/FDA guidance are shown in Table 3 and 4.

Three measures were identified that best met the criteria of the MOT and FDA guidelines: Chedoke Arm and Hand Inventory, Stroke Rehabilitation Assessment of Movement upper limb section, and ABILHAND. Details regarding the development and psychometric properties of these measures are described below. A fourth measure, the Upper Limb-Motor Assessment Scale, also fit many of the MOT criteria, but a closer inspection of psychometric properties of the measure found that the upper limb items should be used with caution.\(^\text{20}\) This scale was therefore not felt to be suitable for use in our study. A summary of the three identified measures is presented below.

**Literature Review-Identified Scales**

**Chedoke Arm and Hand Inventory**

This scale was developed to measure functional tasks in people poststroke.\(^\text{21}\) Initially 109 patients with stroke and their caregivers were interviewed. From this literature review and expert opinion 751 items were generated. Item reduction was then carried out by statistical analysis and expert opinion.

The test consists of 13 functional tasks that reflect domains deemed important by survivors of stroke. This included bilateral activities; non-gender specific tasks; and the full range of movements, pinches, and grasps covering all stages of motor recovery post-stroke.

Psychometric data included face, content, and factorial validity. Correlations with the Action Reach Arm Test (ARAT; \(r=0.93\)) and the Chedoke Master Stoke Assessment \(r=0.87\)) were high. Internal consistency \((r=0.98)\) and single-item factor loading exceeded recommended criteria (range, 0.76 to 0.96) as did interrater reliability (intraclass correlation coefficient, 0.98).

**Stroke Rehabilitation Assessment of Movement Upper Limb Subscale**

The purpose of the Stroke Rehabilitation Assessment of Movement scale was to provide a comprehensive, objective, and quantitative evaluation of motor functioning of individuals with stroke.\(^\text{22}\)

It consists of 30 items that are distributed among three subscales: upper limb movement; lower limb movements; and basic mobility items. Limb movements are scored on a 3-point scale, whereas mobility items are scored on 4-point scale. There is a maximum score of 70 with each limb subscale scored out of 20. Only the upper limb subscale was to be used in the proposed study.

Inter- and intrarater reliability correlation coefficients were 0.99 for total score and 0.96 to 0.99 for subscale scores. Cronbach \(\alpha\) was \(>0.98\) (\(n=26\)). Wang et al\(^\text{23}\) on found interrater reliability-weighted \(\kappa\) of individual items
ranged from 0.55 to 0.94 an intraclass correlation coefficient total score of 0.96 (n=54). These meet recommended criteria. Construct validity has been evaluated through comparisons against the Barthel Index, Fugl-Meyer Assessment (FMA), and Box and Block test, indicating moderate to good levels of concordance. Predictive validity was comparable to Barthel Index and gait speed. Evaluations of responsiveness supported the ability of the Stroke Rehabilitation Assessment of Movement to reflect change over time.

**ABILHAND**

ABILHAND was initially devised to measure “manual (dis)ability” in patients with rheumatoid arthritis who had undergone arthrodesis. The test was then administered to patients with chronic stroke.

This scale measures both unilateral and bimanual activities done without other assistance. For each question, the patient provided his or her feeling of difficulty irrespective of the limb actually used to do the activity. An inventory of 56 manual activities that patients were originally asked to judge on a 4-level scale: 0 (impossible), 1 (very difficult), 2 (difficult), and 3 (easy).

The ABILHAND was evaluated for chronic stroke using the Winsteps Rasch analysis computer program. The measure was found to have a reliability of 0.90 and item difficulty hierarchy was stable. Further information on Rasch analysis is described elsewhere.5

**Additional Scales**

There were immediate drawbacks to using the three identified scales in isolation. All are relatively new scales that have not been widely used in the upper limb intervention trial literature. Using these scales alone would have made it difficult to compare the results of our study with others in the stroke literature.

Therefore, it was felt to be appropriate to add further scales that would address this key point but would still be psychometrically valid and cover impairment and activity. From the initial electronic bibliographic search, two further measures of impairment were selected. The FMA upper limb section and ARAT were chosen because they have been widely used in stroke literature and have been used as “gold standard” for comparison with other measures. A summary of these measures follows.
The FMA is an impairment-based measure developed to assess motor recovery after stroke and was based on early works of Twitchell and Brunnstrom. It is widely used in stroke research and has been used as a gold standard to compare the reliability and validity of other outcome measures.

Scoring ranges from 0 to a maximum of 66 for upper limb movement. The upper limb section has 33 items, which include reflex testing, movement observation, grasp testing, and assessment of coordination.

The validity, reliability, and responsiveness of the FMA have been extensively reported and meet evaluation criteria.

### Action Research Arm Test

The ARAT was devised in 1965 as the upper extremity function test with the objective of developing a testing procedure that was representative of the major activities of the upper limb in everyday activities of daily living. The scale was reorganized by Lyle to measure at upper limb dysfunction post-cerebral cortical injury using a Guttman scale and renamed “ARAT.” The test has been widely used in rehabilitation and treatment trials.

The ARAT is a performance test that consists of 4 domains: grasping (lifting up different size objects), gripping (holding and moving objects), pinching (picking up small objects), and gross movement (e.g., hand to mouth) involving 19 movements.

Reliability, validity, and responsiveness have been investigated and reported to meet recommended criteria. Like the FMA, it has been extensively examined against other measures and is used as a “gold standard” for comparison of other upper limb measures.

### Discussion

In this study, we attempted to bring together best practice psychometric guidelines, a theoretical framework, and clinically important criteria to identify the most appropriate outcome measures for a new study. The scales were then categorized into the ICF framework to ensure clinical relevance. “Gold standard” measures were also included to ensure the results can be interpreted by the wider research community and are reproducible. Using this method, the following scales were chosen to be used in this research trial: Stroke Rehabilitation Assessment of Movement, Chedoke

---

**Table 3. Stroke Clinician-Rated Scales: Development and Validation Criteria**

<table>
<thead>
<tr>
<th>Item generation</th>
<th>ARAT</th>
<th>CAHAI</th>
<th>10-s Test</th>
<th>AMAT</th>
<th>WMFT</th>
<th>FMA</th>
<th>MSS</th>
<th>UL-MAS</th>
<th>DeSouza</th>
<th>RMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient interviews</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert opinion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop conceptual model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item reduction</th>
<th>ARAT</th>
<th>CAHAI</th>
<th>10-s Test</th>
<th>AMAT</th>
<th>WMFT</th>
<th>FMA</th>
<th>MSS</th>
<th>UL-MAS</th>
<th>DeSouza</th>
<th>RMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert opinion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item redundancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endorsement frequencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychometric analyses</th>
<th>ARAT</th>
<th>CAHAI</th>
<th>10-s Test</th>
<th>AMAT</th>
<th>WMFT</th>
<th>FMA</th>
<th>MSS</th>
<th>UL-MAS</th>
<th>DeSouza</th>
<th>RMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal consistency reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item total correlations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interrater reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test–retest reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity: within scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity: comparison with other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity: hypothesis testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ARAT indicates Action Research Arm Test; CAHAI, Chedoke Arm and Hand Inventory; 10-s Test, 10-s tests; AMAT, The Arm Motor Ability Test; WMFT, Wolf Motor Function Test; FMA, Fugl-Meyer Assessment; MSS, Motor Status Score; UL-MAS, upper limb of Motor Assessment Scale; DeSouza, Deouza Arm Test; RMA, Rivermead Motor Assessment; STREAM, Stroke Rehabilitation Assessment of Movement; MESUPE, Motor Evaluation Scale for the Upper Extremity in Stroke Patients; NHPT, 9-Hole Peg Test; FAT, Frenchay Arm Test; Sodring, Sodring Motor Evaluation Test; Sollerman, Sollerman Hand Function Test; MCA, Motor Club Assessment; MMAC, Modified Motor Assessment Chart.

*Psychometric assessment indicated was performed on the scale in each instance.
Arm and Hand Inventory, ABILHAND, FMA, ARAT. These scales represent not only all domains of the ICF framework, but also incorporate a mixture of clinical-rated and patient-reported outcome measures.

This approach to scale selection was felt to add extra value to traditional systematic reviews by using a three-stage approach to aid in the selection of appropriate and valid outcome scales. Using the FDA and MOT guidelines ensured the selected measures that are scientifically rigorous. The ICF ensured the scales were clinically meaningful. We believe this three-stage strategy may be a useful tool in selection of scales for future clinical studies.

Our review has also highlighted three key limitations of stroke-specific rating scales. First, the question still remains of how different measures capture change from different interventions. Currently relatively little responsiveness data have been reported on the scales we describe, and therefore it is difficult to make an accurate assessment of a clinically meaningful differences based on their scores. Second, our review calls attention to the fact that current scales have their limitations in terms of content coverage. For example, there is currently no single valid and reliable scale available that comprehensively captures the complete range of function hemiparetic upper limb. Furthermore, scales that focus on the participation component of the ICF are extremely limited.

Finally, the use of PRO measures in stroke research is a relatively new development and is complicated in this population by cognitive and communication difficulties that can occur in stroke survivors. With care, most of these problems can be overcome. A more specific difficulty that occurs in stroke is the presence of anosognosia (denial of deficit) or neglect. For these patients and those with severe cognitive and communication difficulties, the inclusion of proxy data can assist, although the use of proxy respondents should be approached with caution. Ideally scales need to be validated for proxy use separately.

Modern technologies offer opportunities to address some of these difficulties. The development of item banks, with computerized adaptive testing, may mean stroke survivors need only answer a limited number of questions to obtain a meaningful value rather than the many that may be asked in a battery of measures addressing impairment, activity, and participation. In addition, touch-sensitive screens in which the questions are associated with images, and spoken as well
as written questions, may improve access for communication and cognitively impaired patients.

Conclusions
The need to choose appropriate and valid outcome measures in clinical research has become of increasing importance. If clinically meaningful interpretations are to be made from studies, it is of vital importance that the scales used are rigorous measures of the effects they purport to quantify.3 There is no general consensus on the battery of scales that should be used in clinical stroke trials and clinical practice. Our current review proposes some strategies that can be used to improve the selection of scales.

The following four recommendations are offered for consideration. First, the use of established rigorous guidelines such as the MOT and FDA should be applied to scales to establish important factors regarding the scale, such as its psychometric properties and development. Second, the relevance of stroke outcome scales can be increased by incorporating the ICF framework. Third, currently studies need to use scales that cover a range of activities, including measuring outcome in a robust manner. This has to be the primary aim because this alone will allow us to compare studies accurately. However, in the short term, we need to recognize that studies will be compared and amalgamated using meta-analysis and thus some core measure need to be used to facilitate this. Finally, with increasing understanding and development of psychometrically robust scales, the stroke community needs to migrate to using rating scales that fulfill these important criteria.

Sources of Funding
The research referred to in this article is funded by the Stroke Association.

Disclosures
None.

References