EBRSR
[Evidence-Based Review of Stroke Rehabilitation]

20

Outcome Measures in Stroke Rehabilitation

Katherine Salter PhD (cand.), Nerissa Campbell PhD, Marina Richardson MSc, Swati Mehta PhD (cand.), Jeffrey Jutai PhD, Laura Zettler MSc, Matthew Moses BA, Andrew McClure MSc, Rachel Mays BSc (cand.), Norine Foley MSc, Robert Teasell MD

Last Updated: September 2013

Abstract

To enhance the clinical meaningfulness of the SREBR, the present review provides the best available information on how outcome measures might be classified and selected for use, based upon their measurement qualities. For this purpose, we have selected for review some of the most commonly-used measures in stroke rehabilitation. The ICF conceptual framework is used to classify measures in stroke rehabilitation and aspects of measurement theory pertinent for evaluating measures are discussed. Each measure reviewed in this chapter was evaluated in terms of appropriateness, reliability, validity, responsiveness, precision, interpretability, applicability and feasibility. All measures were assessed for the thoroughness with which its reliability, validity and responsiveness have been reported. The present document contains summary reviews of 38 assessment tools used in the evaluation of Body Structure (14 tools), Activity (15 tools) and Participation (9 tools) outcomes.

Table of Contents

20.1 Introduction

Measuring the effectiveness of interventions is accepted as being central to good practice. Van der Putten et al. (1999) pointed out that measuring the outcome of health care is a "central component of determining therapeutic effectiveness and, therefore, the provision of evidence-based healthcare," (van der Putten et al. 1999).

The Stroke Rehabilitation Evidence-Based Review (SREBR) is a landmark achievement in consolidating the best-available scientific evidence for the effectiveness of stroke rehabilitation. But, there are limitations to successfully transferring the research results to clinical practice and service delivery. Some are imposed by the current state of outcome measurement in stroke rehabilitation. Limitations include the lack of consensus on the selection of measures to best address and balance the needs and values of stakeholders in stroke rehabilitation, including patients and their caregivers, practitioners, and health care decision makers. Ultimately, the comparison of size and direction of statistical results across areas of stroke rehabilitation covered within the SREBR will be most meaningfully interpreted when it is clear that comparable approaches to outcome measurement have been used (Jutai & Teasell 2003). To enhance the clinical meaningfulness of the SREBR, we present the best available information on how outcome measures might be classified and selected for use, based upon their measurement qualities. For this purpose, we have selected for review only some of the more commonly used measures in stroke rehabilitation. We do not intend this to be a comprehensive compendium of stroke outcome measures.

In this chapter, we attempt to describe how the ICF (WHO 2001, 2002) conceptual framework can be used for classifying outcome measures in stroke rehabilitation, and summarize aspects of measurement theory that are pertinent for evaluating measures. We also give a template presentation on the characteristics, application, reliability, validity, and other clinimetric qualities of commonly used measures in a format for easy reference. For a more extensive discussion of outcome measurement theory and properties in rehabilitation, we refer the reader to the book authored by Finch et al. (2002). This chapter will present only the information most relevant for stroke rehabilitation.

20.1.1 Domains of Stroke Rehabilitation

Outcomes research requires a systematic approach to describing outcomes and classifying them meaningfully. The study and assessment of stroke rehabilitation has sparked the development of numerous outcome measures applicable to one or more of its many dimensions. In attempting to discuss some of the commonly used measures available for use within the field of stroke rehabilitation, it is useful to have guidelines available for classifying these tools. The WHO International Classification of Functioning, Disability and Health (ICF: WHO, 2001, 2002) provides a multi-dimensional framework for health and disability suited to the classification of outcome instruments.

Originally published in 1980, the WHO framework has undergone several revisions. In the most recent version, the ICF framework (2001, 2002) identifies three primary levels of human functioning – the body or body part, the whole person and the whole person in relation to his/her social context. Outcomes may be measured at any of these levels -- Body functions/structure (impairment); Activities (refers to the whole person – formerly conceived as disability in the old ICIDH framework) and Participation (formerly referred to as handicap). Activity and participation are affected by environmental and personal factors (referred to as contextual factors within the ICF).

Table 20.1.1 ICF Definitions

Old Terminology	New Terminology	Definition
Impairment	Body function/structure	-Physiological functions of body systems including psychological. Structures are anatomical parts or regions of their bodies and their components. Impairments are problems in body function or structure.
Disability	Activity	-The execution of a task by an individual. Limitations in activity are defined as difficulties an individual might experience in completing a given activity.
Handicap	Participation	-Involvement of an individual in a life situation. Restrictions to participation describe difficulties experienced by the individual in a life situation or role.

Outcome measures can also be conceived of as falling along a continuum of measurement moving from measurements at the level of body function or structure to those focused on participation and life satisfaction. The number of other, non-treatment, variables external to healthcare present that could account for change increases as one moves away from body structure toward life satisfaction, making outcomes much more difficult to define and assess (Brenner et al. 1995; Roberts & Counsell 1998).

If a classification is to be useful for scientific research, the basic categories and concepts within it need to be measurable, and their boundaries clear and distinct. It is not yet clear from the research evidence that the three ICF categories completely fulfill these criteria. Nonetheless, when applied to outcome assessment in stroke rehabilitation the ICF conceptual framework can be used to place outcome measures into one of the three categories depending upon what it is they purport to measure. However, outcome measures rarely fit neatly into a single category. More often, they assess elements belonging to more than one domain. For the purposes of this discussion, measures have been classified according to the level of assessment they include furthest along a continuum from body function, through activity, to participation. The instruments appearing in the Participation domain, for instance, assess elements from all domains including those reflective of participation in life situations such as social functioning or roles. While these measures have been used to assess health-related quality of life, it is not our intent to define such a construct or its assessment here.

Table 20.1.2 Classification of Outcome Measures*

Body structure (*impairments*)	**Activities (*limitations to activity–disability*)**	**Participation (*barriers to participation--handicap*)**
Beck Depression Inventory Behavioral Inattention Test Canadian Neurological Scale Clock Drawing Test Frenchay Aphasia Screening Test Fugl-Meyer Assessment General Health Questionnaire -28 Geriatric Depression Scale Hospital Anxiety and Depression Scale Line Bisection Test Mini Mental State Examination Modified Ashworth Scale Montreal Cognitive Assessment Motor-free Visual Perception Test National Institutes of Health Stroke Scale	Action Research Arm Test Barthel Index Berg Balance Scale Box and Block Test Chedoke McMaster Stroke Assessment Scale Chedoke Arm and Hand Activity Inventory Clinical Outcome Variables Scale Functional Ambulation Categories Functional Independence Measure Frenchay Activities Index Motor Assessment Scale Nine-hole Peg Test Rankin Handicap Scale Rivermead Mobility Scale Rivermead Motor Assessment	Canadian Occupational Performance Measure EuroQol Quality of Life Scale LIFE-H London Handicap Scale Medical Outcomes Study Short- Form 36 Nottingham Health Profile Reintegration to Normal Living Index Stroke Adapted Sickness Impact Profile Stroke Impact Scale Stroke Specific Quality of Life

Orpington Prognostic Scale Stroke Rehabiliation Assessment of Movement	Six Minute Walk Test Timed Up and Go Wolf Motor Function Test	

**Based on tables presented in Roberts & Counsell (1998) and Duncan et al. (2000).*

20.1.2 Evaluation Criteria for Outcome Measures

While it is useful to have this framework within which to classify levels of outcomes measures, it is necessary to have a set of criteria to guide the selection of outcomes measures. Reliability, validity and responsiveness have widespread usage and are discussed as being essential to the evaluation of outcome measures (Duncan et al. 2002; Law & MacDermid 2002; Roberts & Counsell 1998; van der Putten et al. 1999). Finch et al. provide a good tutorial on issues for outcome measure selection (Finch et al. 2002).

The Health Technology Assessment (HTA) programme (Fitzpatrick et al. 1998) examined 413 articles focusing on methodological aspects of the use and development of patient-based outcome measures. In their report, they recommend the use of 8 evaluation criteria. Table 20.1.2.1 lists the criteria and gives a definition for each one. It also identifies a recommended standard for quantifying (rating) each criterion, where applicable, and how the ratings should be interpreted. The table, including some additional considerations described below, was applied to each of the outcome measures reviewed in this chapter.

Table 20.1.2.1 Evaluation Criteria and Standards

Criterion	Definition	Standard
Appropriateness	The match of the instrument to the purpose/question under study. One must determine what information is required and what use will be made of the information gathered (Wade 1992)	Depends upon the specific purpose for which the measurement is intended.
Reliability	- Refers to the reproducibility and internal consistency of the instrument. - *Reproducibility* addresses the degree to which the score is free from random error. Test re-test & inter-observer reliability both focus on this aspect of reliability and are commonly evaluated using correlation statistics including ICC, Pearson's or Spearman's coefficients and kappa coefficients (weighted or unweighted). - *Internal consistency* assesses the homogeneity of the scale items. It is generally examined using split-half reliability or Cronbach's alpha statistics. Item-to-item and item-to scale correlations are also accepted methods.	*Test-retest or interobserver reliability (ICC; kappa statistics):* [1] Excellent: ≥ 0.75; Adequate: 0.4 – 0.74; Poor: ≤ 0.40 Note: Fitzpatrick et al. (1998) recommend a minimum test-retest reliability of 0.90 if the measure is to be used to evaluate the ongoing progress of an individual in a treatment situation. *Internal consistency (split-half or Cronbach's α statistics):* Excellent: ≥ 0.80; Adequate: 0.70 – 0.79; Poor < 0.70[2] **Note:** Fitzpatrick et al. (1998) caution α values in excess of 0.90 may indicate redundancy. *Inter-item & item-to-scale correlation coefficients*: -Adequate levels -- inter-item: between 0.3 and 0.9; item-to-scale: between 0.2 and 0.9[3]
Validity	Does the instrument measure what it purports to measure? Forms of validity include face, content,	*Construct/convergent and concurrent correlations*: Excellent: ≥ 0.60, Adequate: 0.31 - 0.59, Poor: $\leq$

	construct, and criterion. Concurrent, convergent or discriminative, and predictive validity are all considered to be forms of criterion validity. However, concurrent, convergent and discriminative validity all depend on the existence of a "gold standard" to provide a basis for comparison. If no gold standard exists, they represent a form of construct validity in which the relationship to another measure is hypothesized (Finch et al. 2002).	0.30[4] ROC analysis – AUC: Excellent: ≥0.90, Adequate: 0.70 – 0.89, Poor: <0.70 [5] There are no agreed on standards by which to judge sensitivity and specificity as a validity index (Riddle & Stratford, 1999)
Responsiveness	Sensitivity to changes within patients over time (which might be indicative of therapeutic effects). Responsiveness is most commonly evaluated through correlation with other change scores, effect sizes, standardized response means, relative efficiency, sensitivity & specificity of change scores and ROC analysis. Assessment of possible floor and ceiling effects is included as they indicate limits to the range of detectable change beyond which no further improvement or deterioration can be noted.	*Sensitivity to change*: Excellent: Evidence of change in expected direction using methods such as standardized effect sizes: <0.5 = small; 0.5 – 0.8 = moderate ≥0.8 = large) Also, by way of standardized response means, ROC analysis of change scores (area under the curve – see above) or relative efficiency. Adequate: Evidence of moderate/less change than expected; conflicting evidence. Poor: Weak evidence based solely on p-values (statistical significance) [6] *Floor/Ceiling Effects*: Excellent: No floor or ceiling effects Adequate: floor and ceiling effects ≤20% of patients who attain either the minimum (floor) or maximum (ceiling) score. Poor: >20%. [7]
Precision	Number of gradations or distinctions within the measurement. E.g. Yes/no response vs. a 7-point Likert response set	Depends on the precision required for the purpose of the measurement (e.g., classification, evaluation, prediction).
Interpretability	How meaningful are the scores? Are there consistent definitions and classifications for results? Are there norms available for comparison?	Jutai & Teasell (2003) point out these practical issues should not be separated from consideration of the values that underscore the selection of outcome measures. A brief assessment of practicality will accompany each summary evaluation.
Acceptability	How acceptable the scale is in terms of completion by the patient – does it represent a burden? Can the assessment be completed by proxy, if necessary?	
Feasibility	Extent of effort, burden, expense & disruption to staff/clinical care arising from the administration of the instrument.	

Unless otherwise noted within the table, criteria and definitions: Fitzpatrick et al. (1998); McDowell & Newell (1996). Sources for evaluation standards: [1]Andresen (2000); Hseuh et al. (2001); Wolfe et al. (1991); [2]Andresen (2000);[3]Hobart et al. (2001); Fitzpatrick et al. (1998); [4,6]Andresen (2000); McDowell & Newell (1996); Fitzpatrick et al. (1998); Cohen et al. 2000; [5]McDowell & Newell (1996); [7]Hobart et al. (2001).

Each measure reviewed in this chapter was also assessed for the thoroughness with which its reliability, validity and responsiveness have been reported in the literature. Standards for evaluation of rigor were adapted from McDowell & Newell (1996) and Anderson (2000).

Table 20.1.2.2 Evaluation Standards – Rigor

Thoroughness or Rigor of testing	Excellent – most major forms of testing reported. Adequate – several studies and/or several types of testing reported Poor – minimal information is reported and/or few studies (other than author's) N/a – no information available

Assessments of rigor using the above standards are given along with evaluation ratings for reliability, validity and responsiveness for each measure (see Table 20.1.2.3, below).

Table 20.1.2.3 Evaluation Summary

Reliability		**Validity**		**Responsiveness**		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

Ratings of +++(excellent), ++ (adequate) and + (poor) are assigned based on the criteria and evidence presented in the standards column of Table 20.3. For example, If a rating of "+++" or excellent is given for validity, it means that evidence has been presented demonstrating excellent construct validity based on the standards provided and in various forms including convergent and discriminant validity.

In addition to the criteria outlined above, 3 additional issues were considered:
Has the measure been used in a stroke population?
Has the measure been tested for use with proxy assessment?
What is the recommended time frame for measurement?

20.1.3 Has the Measure Been Used in a Stroke Population?

Reliability and validity are not fixed qualities of measures. They should be regarded as relative indicators of how well the instrument might function within a given sample or for a given purpose (Fitzpatrick et al. 1998; Lorentz et al. 2002). Responsiveness, too, may be condition or purpose specific. Van der Putten et al., (1999) for example, found the Barthel Index and the FIM exhibited greater effect sizes among stroke patients than among MS patients concluding that responsiveness of instruments seems disease- or condition- dependent. Therefore, it is important for a measure to have been tested for use in the population within which it will be used.

Measures developed for generic use cannot focus on the problems associated with any one condition and, therefore, may not be sensitive to problems inherent in the stroke population (Buck et al. 2000). In a discussion of health-related quality of life measurement, Williams et al. (1999) point out that generic measures may not include particular assessments of importance in stroke (such as arm and hand or language assessments).

20.1.4 Has the Measure Been Tested for Use with Proxy Assessment?

When assessment is conducted in such a way as to require a form of self-report (e.g. interview or questionnaire – in person, by telephone or by mail), stroke survivors who have experienced significant cognitive or speech and language deficits may not be able to complete such measures and therefore, may be excluded from assessment. In such cases, the use of a proxy respondent becomes an important alternative source of information. However, the use of proxy respondents should be approached with a degree of caution.

In studies of proxy assessments, a tendency has been reported for family members or significant others to assess the patient as more disabled than they appear on other measures of functional disability, including self-reported methods. This discrepancy becomes more pronounced among patients with more impaired levels of functioning (Hachisuka et al. 1997; Segal et al. 1996; Sneeuw et al. 1997). Hachisuka et al. (1997) suggested that this discrepancy could be explained by a difference in interpretation. Proxy respondents may be rating actual, observable performance, while patients may rate their perceived capability – what they think they are capable of doing rather than what they actually do.

Unfortunately, use of a healthcare professional as a substitute for the family member or significant other as proxy does not solve the problem of reliability. A similar discrepancy has been noted in ratings when using healthcare professionals as proxy respondents though in the opposite direction. They may tend to rate patients higher than the patients themselves would (McGinnis et al. 1986; Sneeuw et al. 1997). It has been suggested that, in this case, the discrepancy is due to a difference in frame of reference. A healthcare professional may use a different, more disabled group, as a reference norm whereas the patient would only compare him/herself to pre-stroke conditions (McGinnis et al. 1986).

20.1.5 What is the Recommended Timeframe for Measurement?

The natural history of stroke presents problems in assessment in that the rate and extent of change in outcomes varies across the different levels of ICF classification (Duncan et al. 2000). The further one moves along the outcome continuum from body structure toward participation, the more time it may take to reach a measurement end point, that is, social context may take longer to stabilize than the impaired body structure (Duncan et al. 2000).

Jorgensen et al. (1995) demonstrated that recovery in Activities of Daily Living (ADL) occurs, in most patients, within the first 13 weeks following a stroke even though the time course of both neurological and functional recovery is strongly related to initial stroke severity. They suggest that a valid prognosis of functional recovery might be made within the first 6-months. According to Mayo et al., by 6 months post-stroke, physical recovery is complete, for the most part, with additional gains being a function of learning, practice and confidence (Mayo 1999). Duncan et al. (2000) support this suggested time frame for assessment of neurological impairment and disability outcomes but suggest that participation outcomes wait at least 6 months to provide the opportunity for the patient's social situation to stabilize. They also suggest that assessments at the time of discharge not be used as endpoint measurements. The variability in treatment interventions and length of stay practices decreases the comparative usefulness of this information.

20.2 Body Structure/Impairment Outcome Measures

This section corresponds to the first level or category of the ICF classification system. While keeping in mind that the fit of a given instrument within a single category is rarely perfect, measures appearing in this section focus primarily on the identification or assessment of impairments in body function, structure or system (including psychological).

20.2.1 Beck Depression Inventory (BDI)

The Beck Depression Inventory was developed to provide a quantitative expression of the intensity of depression (Beck et al. 1961). Items appearing on the inventory were derived through clinical observation and were not intended to reflect any particular theoretical approach to depression or its diagnosis. Since its introduction, it has become a widely used instrument for detection and assessment of intensity of depression.

The inventory consists of 21 items, which represent symptoms or attitudes associated with depression. Each item is presented as a multiple choice response set comprised of 4 self-evaluative statements graded from 0-3 in severity. The respondent is to choose the statement that fits him/her best relative to the past week up to and including today (Beck et al. 1961; McDowell & Newell 1996). Ratings are summed to provide a total score ranging from 0 – 63. The generally accepted threshold for presence of depression is 10 (Aben et al. 2002). Additionally, classifications of 10-18 (mild), 19-29 (moderate) and 30 – 63 (severe) are commonly used (Beck et al. 1988). Originally administered by a trained interviewer, it has become most common for the BDI to be administered as a self-completion questionnaire. In this form, it takes approximately 5 – 10 minutes to complete (Beck et al. 1988; McDowell & Newell 1996). A 13-item short form was developed by Beck and Beck (1972). Copies of the scale and permission to use it can be obtained from The Psychological Corporation, Texas, USA.

Advantages

The BDI is short and simple to administer (McDowell & Newell 1996). It does not require training to administer. Aben et al. (2002) found no substantial differences between the BDI and 3 other depression-screening tools when used with stroke populations. Its brevity and simplicity, together with the fact that it does not rely heavily on the somatic components of depression, may recommend it as the most suitable depression scale for administration among stroke patients (Aben et al. 2002; Turner-Stokes & Hassan 2002).

Beck et al. (2000) developed a shortened version for use as a screening tool to identify the possible presence of depression in medical patients. This 7-item version does not include items representative of somatic symptoms of depression. A single study was identified that examined the use of this scale in individuals with stroke (Healey et al. 2008). Although the authors reported evidence of acceptable reliability and validity, sensitivity and specificity for the identification of major and minor depression were somewhat low (0.62 and 0.78, respectively). Use of the BDI-FS missed 2 patients diagnosed with major depression and produced 11 false positives (Healey et al. 2008). However, these results were based on a cut-off score derived from a sample of geriatric outpatients rather than individuals with stroke, Further research with a larger sample of stroke patients is necessary in order to determine optimum cut-offs for the BDI-FS within this population.

To reflect the updated DSM-IV criteria for depression, Beck et al. (1996) published the BDI-II in 1996. Although the BDI-II may be used relatively frequently in the assessment of depression in adults, there is little evidence within the research literature to suggest that it is used routinely in the assessment of either elderly individuals or of individuals who have experienced stroke. Apart from a single study examining the factor structure of the BDI-II (Siegert et al. 2009), we could identify no published evaluations of the reliability or validity of this version of the BDI when used to assess depression within our population of interest.

Limitations

Although the standardized cutoff for the presence of depression seems to be optimal for use in a stroke population, the inventory still yields a high rate (approx. 31%) of misdiagnosis among the stroke population especially among women (Aben et al. 2002). Aben et al. (2002) suggested that this could be due to a tendency in female patients to report non-specific distress and, thereby, artificially inflate depression scores. Overall, sensitivity of the BDI tends to be greater than specificity. Berg et al. (1989) suggest that the BDI is sensitive enough to perform well as a screening tool, but should not be used for the diagnosis of depression.

Difficulty with scale completion has also been reported (Aben et al. 2002; House et al. 1991). House et al. (1991) suggested that reduced completion rates could be associated with difficulties in following the forced choice response format.

A single study has examined the use of proxy respondents to complete the BDI (Berg 1989). Caregivers tended to rate individuals with stroke as more depressed than the patients themselves by approximately 4 points and the association between caregiver ratings and patient scores was relatively poor with correlations ranging from 0.37 – 0.43 over the period of 18 months following stroke. Proxy or caregiver ratings of patient depression appeared to be more strongly related to their own feelings of depression than to the patient's own ratings (r=0.60 – 0.61, p<0.001).

Summary – Beck Depression Inventory

Interpretability: The BDI is a well-established measure, with generally accepted cut-off scores for both the presence and severity of depression. No standardized norms are available.
Acceptability: Although the BDI takes only 5 – 10 minutes, problems with completion have been noted within a stroke population (Aben et al. 2002).
Feasibility: The BDI is short and simple to administer requiring no training. There is limited information available regarding its effectiveness when used for evaluation purposes in a longitudinal study.

Table 20.2.1.1 BDI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IC)	+++	+++	+	+	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.2 Behavioral Inattention Test (BIT)

The Behavioral Inattention Test (BIT) is a comprehensive battery designed to screen for unilateral visual neglect and provide information relevant to its treatment. Unilateral visual neglect is a condition characterized by impairment in the ability to respond to stimuli located in space contralateral to a brain lesion. The BIT was developed by Wilson et al. (1987) to provide an ecologically valid assessment of everyday skills relevant to this condition. As such, the test gives therapists and clinicians a detailed description of patients' capabilities and provides a useful framework upon which to base rehabilitation interventions.

The BIT is divided into two major sections, each comprised of its own set of subtests. The BIT conventional section (BITC) consists of 6 conventional tests of visual neglect: Line crossing, letter cancellation, star cancellation, figure and shape copying, line bisection, and representational drawing. The BIT behavioral section (BITB) consists of 9 behavioral tasks: Pre-scanning, phone dialing, menu reading, article reading, telling and setting the time, coin sorting, address and sentence copying, map navigation, and card sorting. Parallel versions of the test have been created to minimize practice effects upon re-testing. Each version is comprised of the 6 conventional and 9 behavioural subtests.

Tables 20.2.2.1 and 20.2.2.2 contain brief descriptions of each of the subtests that make up the BITC and the BITB, respectively:

Table 20.2.2.1 BITC – Conventional Section Test Descriptions

BITC Subtest	Test Description	Scoring
Line Crossing	Patients are required to detect and cross out all target lines on a page. When administering this test, the examiner demonstrates the nature of the task to the patient by crossing out two of four lines located in a central column, and then instructing them to cross out all lines they can see on the page.	The four central lines are not included and neglect is diagnosed if any lines are missed by the patient. A score sheet is provided to notate the nature of the neglect (i.e., contralateral, ipsilateral, or more diverse patterns of omission).
Letter Cancellation	Paper and pencil test in which patients are required to scan, locate, and cross out designated targets from a background of distractor letters. The test consists of 5 rows of 34 upper case letters presented on a rectangular page (279 x 210mm). Forty target stimuli ('E' and 'R') are positioned such that each appear in equal number on both sides of the page. Each letter is 6 mm high and positioned 2mm apart from the next.	The maximum score is 40, and a scoring template allows scorer to divide the total array into four columns, two on the left and two on the right. On completion of the task, the total number of omitted target letters is calculated, and the location of the omissions is noted.
Star Cancellation	This test consists of a random array of verbal and non-verbal stimuli. The stimuli are 52 large stars (14mm), 13 randomly positioned letters and 19 short (3-4 letters) words are interspersed with 56 smaller stars (8mm) which comprise the target stimuli. The patient is instructed to cancel all the small stars. Two examples of small stars are pointed out and cancellation of two central stars is demonstrated.	As with the letter cancellation task, the test sheet can be subdivided into columns to calculate the number and location of errors.
Figure and Shape Copying	In this test, the patient is required to copy three separate, simple drawings from the left side of the page. The three drawings (a four pointed star, a cube, and a daisy) are arranged vertically and are clearly indicated to the patient. The second part of the test requires the patient to copy a group of three geometric shapes presented on a separate stimulus sheet. Unlike the previous items, the contents of the page are not pointed out to the patient.	Scoring is based on completeness of each drawing. Neglect is defined as an omission or gross distortion of any major contralesional component of the drawing.
Line Bisection	Patients are required to estimate and indicate the midpoint of a horizontal line. The expectation is that the patient with left neglect will choose a midpoint to the right of true centre. Each patient is presented with three horizontal, 8-inch (204mm) black lines (1mm thick) displayed in a staircase fashion across the page. The extent of each line is clearly pointed out to the patient who is then instructed to mark the centre.	The test is scored by measuring deviations from true midpoint. Deviations to left scored as negative; to the right as positive. Deviation score is calculated using the normative data obtained from the age-matched controls. Each of the three lines is scored out of a maximum of three. Using data from the control sample, score values between 0 and 3 (+ or -) are assigned to the patient's performance.

Representational Drawing	Patient is asked to draw pictures of a clock face, together with the numbers and a setting of the hands; a man or woman; and a simple outline drawing of a butterfly. The task is designed to assess patient's visual imagery independent of direct sensory input. Patients with left sided neglect typically use the right side of the page and their drawings often contain major omissions of features on the left hand side. Drawings of a clock face, the human form and a butterfly have shown themselves clinically to be sensitive tests objects.	Scoring is similar to copying tasks, where neglect is defined as the omission or gross distortion of any major contralesional component of the drawing.

Table 20.2.2.2 BITB – Behavioral Section Test Descriptions

BITB Subtest	Test Description	Scoring
Picture Scanning	Three large photographs (a meal, a wash basin and toiletries, and a large room flanked by various pieces of furniture and hospital aids), each measuring 357 x 278mm are presented one at a time. Each photograph is placed in front of the seated patient who is not permitted to move it. The patient is instructed to name and/or point to the main items in each picture.	Only omissions are scored, though errors of identification also noted. Scoring of this and all other BITB tests is out of a total of nine and is calculated from the total number of omissions recorded.
Telephone Dialing	A telephone with a numbered dial or a push button keyboard is presented. Each number is placed directly in front of the telephone and patient instructed to dial the number sequence presented.	Dialing sequence is recorded, together with number and location of omissions or substitutions.
Menu reading	A menu 'open-out' page (420 x 297mm) containing 18 common food items arranged in 4 adjacent columns (2 on the left and 2 on the right) is presented. The food items are presented in 6mm high letters. Patient is instructed to open the menu and read out all the items. Language-impaired patients are permitted to point to all the words they can see.	Each of 18 items is scored as correct or incorrect, where incorrect responses refer to partial/whole word substitutions or omission.
Article Reading	Three short columns of text are presented, which patients are then instructed to read.	Scoring is based on the percentage of words omitted across all three columns. Word omissions and partial or whole word substitutions are scored as errors.
Telling and Setting the Time	This test has three parts. First, the patient is required to read the time from photographed settings on a digital clock face. Second, the patient is required to read the time from three settings on an analogue clock face. Finally, the patient is instructed to set times on the analogue clock face as they are called out by the examiner.	All three parts are scored according to # of omissions or substitutions made.
Coin Sorting	An array of familiar coins (six denominations, three of each type) is presented. The patient is then instructed to indicate the locations of the coin type called out by the examiner. This task requires selective scanning of the coin array in order to not miss any instance of the named denomination.	Scoring is based on the number of omissions.
Address and	Patient is required to copy an address and a	Score is calculated from the number of letters

Sentence Copying	sentence on separate pages.	omitted or substituted from each side of the page.
Map Navigation	Patient is required to follow and locate spatial points (letters) positioned on a network of pathways located on a sheet of paper. More specifically, after having been shown the junctions of each pathway, patients are instructed to use their fingers to trace out routes (Sequences of letters) called out to them.	Failure to complete any segment of the route sequence incurs a penalty deduction of one point down to a minimum of zero for each trial.
Card Sorting	Sixteen playing cards are presented in a 4 x 4 matrix. Initially, each card is pointed out to the patient, who is then required to point to each of the card types present as the examiner calls them out.	To score, the position and total number of omissions are recorded.

N.B. Information in tables 10 and 11 from Halligan, Cockburn and Wilson (1991)

Aggregate scores for the BITB and the BITC, as well as the total score for the BIT are obtained by adding the subtest scores together. Neglect is diagnosed based on two aspects of patient performance: 1) failure to attend to target stimuli (as evidenced by target omission or incomplete drawing); and 2) relative spatial location of targets omitted (with reference to side of lesion and/or the patient's sagittal midplane). Halligan et al. (1991) established cut-off scores beyond which neglect is diagnosed. The cut-offs were derived from the aggregate of the lowest scores achieved by any control participant on each of the conventional tests, each of the behavioural tests, and for the total test. For the BITC, the BITB, and the total BIT, the cut-offs are 129 out of 146, 67 out of 81 and 196 out of 227, respectively (Menon & Korner-Bitensky 2004).

To score the relative spatial location component (the index of laterality), the number of screening tests that demonstrated an overall lateralized performance is calculated. If half of the tests show lateralized performance and half do not, the index of lateralized performance is then determined by the total number of omissions/errors made on each side. Finally, a severity of neglect score can be calculated based on performance on the 6 BITC tests. This score is determined by the number of conventional tests on which a given patient demonstrates neglect. The severity rating ranges from a score of 1 (mild neglect) to a score of 6 (severe neglect).

The test takes approximately 40 minutes to administer and can be purchased commercially.

Advantages

It has been suggested that single paper-and-pencil tests are insufficient to evaluate hemineglect given the relative variability and complexity of the diagnosis (Azouvi et al. 2002; Lopes et al. 2007). As a comprehensive battery, the BIT provides a more detailed and ecologically valid assessment of patient functioning than individual tests of visual neglect. The BIT was in fact, designed to provide such a description for rehabilitative purposes (Wilson et al. 1987). With this purpose in mind, the authors ensured the test had strong face validity by selecting test items with the help of psychologists and occupational therapists familiar with everyday problems faced by visual inattention patients. However, it is important to note that the target patient population itself was not included in the item selection process, as has been done with other outcome measures, such as the Chedoke Arm and Hand Activity Inventory.

Moreover, the BIT utilizes the strengths of the conventional and behavioural sections to arrive at a comprehensive description of patient function. Whereas the conventional subtests are used to screen for and provide a foundational assessment of visual neglect, behavioural subtests specifically assess skills relevant to rehabilitation and re-integration into the community. As such, this tool is beneficial in helping therapists target the tasks that should be given particular attention during treatment. Other advantages of the BIT include the provision of 2 parallel forms of the test, which allow for re-testing with minimal concern for practice effects, and the fact that the behavioural measures allow for performance to be evaluated irrespective of theoretical orientation. As well, the test has established cut-off values (Wilson et al. 1987), which have been used in more recent studies (Jehkonen et al. 2000).

The BITB appears to be a useful predictive tool, which could aid post-stroke neuropsychological examinations in determining prognosis. Jehkonen et al. (2000) found that the BITB section of the test was the single most powerful predictor of poor functional outcome at 3, 6 and 12 month follow-ups, accounting for 73%, 64%, and 61% of total variance in the Frenchay Activities Index (FAI) at each of these intervals. This standing was held in comparison to a number of alternative predictive variables including age, hemiparesis, and the BITC.

Limitations

The BIT is more time consuming and more expensive in both cost and material than the individual conventional or behavioural tests from which it is composed. However, an 11-minute shortened version of the test was created to provide a more convenient bedside assessment tool. This version may lose some of the sensitivity of the full-length test because it consists of only 3 conventional subtests and 5 behavioural subtests. However, in time-constrained situations, this may be an acceptable sacrifice. This version of the BIT has some evidence of reliability, validity and responsiveness to clinical change (Menon & Korner-Bitensky 2004; Stone et al. 1994). The BIT short form is still considerably longer and more expensive in cost and material than most non-battery tests of neglect.

Summary – Behavioral Inattention Test

Interpretability: The BIT is a comprehensive battery used to screen for unilateral visual neglect and to provide information relevant to its treatment. Cut-offs published by the test creators (129 out of 146 for BITC, 67 out of 81 for BITB, 196 out of 227 for Total test) have been used in more recent research (Jehkonen et al. 2000).
Acceptability: Test administration is lengthy at 40 minutes and requires a number of skills (e.g., reading, writing, visual memory, holding a pencil) to complete. Thus, the BIT is more taxing on participants than individual tests of visual neglect. An 11-minute shortened version is available for more convenient bedside use.
Feasibility: This test requires considerably more time to administer than individual tests of neglect. The BIT can be purchased commercially.

20.2.2.3 BIT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++ (IO) + (IC)	+++	+++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.3 Canadian Neurological Scale (CNS)

The Canadian Neurological Scale (CNS) is a standardized neurological assessment of stroke patients who are either alert or drowsy. The CNS was intended as a simple tool to be used in the evaluation and monitoring of neurological status of stroke patients during the acute period post stroke (Cote et al.

1986). Test items were chosen based on a literature review and on the clinical experience of the scale authors (Cote et al. 1986).

The CNS is a simple clinical evaluation of mentation (level of consciousness, orientation and speech) and motor function (face, arm and leg). Motor function evaluations are separated into sections A1 and A2. A1 is administered if the patient is able to understand and follow instructions. A2 is administered in the presence of comprehension deficits (Cote et al. 1989, 1986). Each motor item is rated for severity and each rating is weighted "according to the relative importance of a particular neurologic deficit," (Cote et al. 1989). Scores from each section are summed to provide a total score out of a possible 11.5. Lower scores are representative of increasing severity.

Assessment using the CNS requires approximately 5 – 10 minutes to complete (Cote et al. 1989, 1986).

Advantages

The CNS does not need to be completed by a neurologist. The CNS was designed so that it could be completed by trained healthcare professionals, not only neurologists. It is a short and simple assessment that may be applied at intervals to monitor change and predict patient outcomes (Anemaet 2002; Cote et al. 1986). It has been demonstrated that the CNS is a valid predictor of outcomes such as length of stay, death and dependency. Furthermore, the Thai version of the CNS has been shown to be reliable and valid (Charoenpong 2013).

Limitations

Assessment using the CNS is focused on limb weakness over other possible neurological impairments (Cuspineda et al. 2003; Muir et al. 1996).

Summary – Canadian Neurological Scale

Interpretability: A simple, straightforward assessment of neurological status. Results from the CNS can be used in a simple formula, along with patient age, to predict outcome (4-month probability of disability or death) (Fiorelli et al. 1995).
Acceptability: The CNS is short and simple. Patient burden associated with its use should be minimal.
Feasibility: The CNS does not need to be administered by a neurologist. It may be used both prospectively and retrospectively. It is available for use free of charge.

Table 20.2.3.1 Evaluation Summary CNS

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	++ (IO) +++ (IC)	++	+++	+	+	N/a

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.4 Clock Drawing Test (CDT)

The Clock Drawing Test (CDT) has been in use since approximately 1986 (McDowell & Newell 1996). The CDT provides a quick assessment of visuospatial and praxis abilities and may reflect both attention and executive dysfunction (Adunsky et al. 2002; McDowell & Newell 1996; Suhr et al. 1998).

In its most basic form, the CDT is a simple task completion test requiring the individual to draw a clock face, place the numbers on the clock and draw hand pointing to a given time. The individual may be presented with a pre-drawn circle and need only place the numbers and hands on the clock face or the clock may be entirely self-generated. The test is very simple to administer taking approximately 1 – 2 minutes to complete (Ruchinskas & Curyto 2003). There are numerous systems by which to score the individuals efforts in completing the test. In general, they evaluate errors and/or distortions in the form of omissions of numbers and errors in their placement such as perseverations, transpositions and spacing (McDowell & Newell 1996). Scoring systems may be simple or complex, quantitative or qualitative in nature.

Advantages

The CDT is an extremely brief and very simple tool that can be used to supplement other cognitive assessments (McDowell & Newell 1996; Ruchinskas & Curyto 2003; Suhr & Grace 1999). Performance on the CDT is more related to functions subserved by the right hemisphere (Suhr et al. 1998) and when used with other assessments may help to create a more complete picture of cognitive function. While there are many possible procedures associated with the administration and scoring of the CDT, the psychometric properties of all the various systems seem quite consistent and all forms have been shown to correlate strongly with other cognitive measures (McDowell & Newell 1996; Ruchinskas & Curyto 2003; Scanlan et al. 2002).

While the multiplicity of scoring systems has a number of associated disadvantages, it also provides a degree of flexibility to the CDT. For instance, simple quantitative systems might be sufficient to discriminate presence versus absence of cognitive impairment for the purposes of initial screening (Lorentz et al. 2002), while a more complex, qualitative system would yield additional information. It has been demonstrated that different scoring methods are better suited to different subject groups (Heinik et al. 2004; Richardson & Glass 2002). For example, patients with multi-infarct dementia are more likely to make errors in time-setting than in number-spacing and greater levels of cognitive impairment are reflected by scales that place more weight on that feature (Richardson & Glass 2002). The CLOX variation designed to discriminate between executive and non-executive elements of cognitive impairment (Royall et al. 1998), may be of particular use in the assessment of individuals with stroke; however, this requires further evaluation.

Limitations

As is the case with many other neuropsychological screening measures, CDT is influenced by increasing age, level of education and the presence of depression (Lorentz et al. 2002; Lourenco et al. 2008; Ruchinskas & Curyto 2003), although the degree to which these variables have an effect is dependent upon the scoring system used (McDowell & Newell 1996). Clock drawing can also be affected by other conditions prevalent in rehabilitation settings such as visual neglect, hemiparesis and motor dyscoordination (Ruchinskas & Curyto 2003). Given its focus on right hemisphere function, it might best be used as a supplement to another test rather than as an independent assessment (McDowell & Newell 1996).

In the identification of cognitive impairment (mild through dementia), reported sensitivity is often low for a variety of scoring methods (Can et al. 2012; Ehreke et al. 2011; Lee et al. 2008; Lourenco et al. 2008; McDowell & Newell 1996). Reported AUC values in recent studies have been low to adequate and appear consistent across evaluated scoring methods (Lee et al. 2008; Lourenco et al. 2008; Nokleby et al. 2008). Although the CDT has been used to identify the presence of specific deficits in visuospatial function and neglect, it should be used with caution. In a recent report examining the 7-minute screen, Manos and Sunderland scoring methods, sensitivity for identification of impairment in visuospatial function ranged from 55 to 68% while sensitivity ranged from 44-74% depending upon the scoring method and cut-off score used (Nokleby et al. 2008). Values for sensitivity and specificity for identification of attention and neglect were 55% and 42-49% respectively. No single scoring method appeared to yield superior results. It should be noted that none of these most recent results were obtained from a group of individuals with stroke.

The number of available scoring systems has made it difficult to develop normative databases, which could be stratified for age and level of education (Ruchinskas & Curyto 2003). Additionally, the variability in scoring methods decreases the facility with which one might compare results between studies or patient groups.

Summary – Clock Drawing Test

Interpretability: No normative values are available. Given the multiplicity of scoring procedures, comparison across groups or studies is difficult. No single system has been agreed upon as standard.
Acceptability: The test is very short and simple. It is a nonverbal task and may be less threatening to patients than a series of grade-school type questions.
Feasibility: The CDT is inexpensive and highly portable. It can be administered in situations in which longer tests would be impossible or inconvenient. Even the most complex administration and scoring system requires approximately 2 minutes. It can be used by individuals with little or no training or experience in cognitive assessment.

Table 20.2.4.1 CDT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) ++ (IO)	+++	++	n/a	n/a	n/a

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.5 Frenchay Aphasia Screening Test (FAST)

First published in 1987 (Enderby et al. 1987a,1987b), the FAST was created to provide healthcare professionals working with patients who might have aphasia a quick and simple method to identify and gauge language deficit. The FAST was intended to be used as a screening device to identify those patients having communication difficulties who should be referred for a more detailed evaluation performed by a speech and language pathologist (Enderby & Crow 1996; Enderby et al. 1987a, 1987b).

The FAST assesses language in 4 major areas: comprehension, verbal expression, reading and writing. Testing is focused around a single, double-sided stimulus card depicting a riverside scene on one side and geometric shapes on the other and five written sentences. All instructions given to the respondent are of graded length and difficulty. Points are awarded based on the correctness or completeness of response. Scores from each test area are summed to provide a total score out of 30. Ten points are available for each of comprehension and verbal expression; five for each of reading and writing. It is possible to reduce administration time by administering only the first two sections of the test (comprehension and expression) for a total combined score of 20. The classification sensitivity of this shortened version of the FAST is reported to be similar to that reported for the complete assessment (Enderby et al. 1987b). Age stratified norms are available for the total test and for administration of only the comprehension and expression subsections. Reported administration time ranges from 3 to 10 minutes (Enderby & Crow 1996; Spreen & Risser 2002).

Advantages

One of the best known and most thoroughly evaluated screening measures, the FAST is both quick and simple to administer. Administration of the comprehension and expression subtests alone provides an option for an abbreviated screening. This could be most useful for patients who are unable to tolerate longer testing procedures. The FAST has been reported to be reliable when used during both the acute and post acute periods and shows good concurrent validity when evaluated against assessments of both impairment and function (Al-Khawaja et al. 1996; Enderby et al. 1987a). In addition to identifying the presence of language deficits, FAST scores have been used as a way to provide a quick snapshot of change over time (Enderby et al. 1987b). While repeated administration of the FAST demonstrated significant change in the expected direction, the responsiveness of the FAST to change has not been evaluated in more detail.

Limitations

While use of the FAST has been reported to have good classification sensitivity, the specificity of the FAST appears to be adversely affected by the presence of visual field deficits, visual neglect or inattention, illiteracy, deafness, poor concentration or confusion (Al-Khawaja et al. 1996; Enderby et al. 1987b; Gibson et al. 1991). O'Neill et al. (1990) reported lower specificity associated with FAST than with clinical examination suggesting that administration of the screening test provides no real advantage over the careful examination of an experienced clinician.

A significant inverse relationship between age and FAST score has been reported (O'Neill et al. 1990). Although stratified cut-offs and normative data are available for both the complete and shortened versions of the FAST for three age groups; ≤ 60 years, 61 – 70 years and ≥71 years, this is based on the assessment of a small sample (n=123) of normal individuals aged 21 – 81+ (Enderby et al. 1987b; Spreen & Risser 2002). As the representation of the very old within the normative sample was limited, it has been recommended that test scores be interpreted with caution and the cut-off point signifying the presence of language difficulties in this group be lowered to avoid the incorrect classification of very elderly subjects (O'Neill et al. 1990).

Summary – Frenchay Aphasia Screening Test

Interpretability: Age-stratified normative data is available, based on the assessment of 123 individuals aged 20 to 81+. In interpreting results among the elderly, it should be noted that, of these individuals, only 10 were over the age of 81 and 21 were between the ages of 71 and 80.
Acceptability: The FAST is short and simple, requiring less than 10 minutes to administer. It may be well suited for use among individuals who are unable to tolerate long or complex testing procedures.
Feasibility: The FAST is simple to administer even during a bedside evaluation. Test materials are simple and portable.

Table 20.2.5.1 FAST Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IO)	+	+++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.6 Fugl-Meyer Assessment of Motor Recovery after Stroke (FMA)

The Fugl-Meyer Assessment is a disease-specific impairment index designed to assess motor function, balance, sensation qualities and joint function in hemiplegic post-stroke patients (Fugl-Meyer et al. 1975; Gladstone et al. 2002).

The scale comprises five domains; motor function (in the upper and lower extremities), sensory function, balance (both standing and sitting), joint range of motion and joint pain. Items in the motor domain were derived from Twitchell's 1951 description of the natural history of motor recovery following stroke and incorporates Brunnstrom's stages of motor recovery (Gladstone et al. 2002). Items are intended to assess recovery within the context of the motor system. Functional tasks are not incorporated into the evaluation (Chae et al. 2003).

Scale items are scored on the basis of ability to complete the item using a 3-point ordinal scale where 0=cannot perform, 1=performs partially and 2= performs fully. The total possible scale score is 226. Points are divided among the domains as follows: 100 for motor function (66 upper & 34 lower extremity), 24 for sensation (light touch and position sense), 14 points for balance (6 sitting & 8 standing), 44 for joint range of motion & 44 for joint pain. Classifications for impairment severity have been proposed based on FMA scores (Duncan et al. 1994; Fugl-Meyer 1980).

It is not uncommon for the sections of the FMA to be administered separately. However, it should take approximately 30 – 45 minutes to administer the total FMA. Assessments are completed by direct observation on a one-to-one basis and should be performed by a trained physical therapist (Gladstone et al. 2002).

Advantages

The Fugl-Meyer assessment is widely used and internationally accepted. The motor assessment is grounded in well-defined, observable stages of motor recovery (Gladstone et al. 2002). The FMA has been used as the gold standard against which the validity of other scales is assessed.

The total assessment may be administered in whole or in part, though the motor sections are the most thoroughly studied and most often used. Joint pain and sensation are more subjective in nature and are used less frequently (Gladstone et al. 2002). The ability to use subsections independently according to purpose increase the flexibility and feasibility of the measure. Hiengkaew et al. (2012) found that the FMA lower extremity (LE) subscales are a reliable measure to detect postural balance and lower limb movements. Page et al. (2012) found that FMA is also reliable in assessing upper extremity function specifically wrist stability and mobility. A computerized adaptive testing system has been developed which allows efficient and reliable assessment of motor function through the FMA (Hou et al. 2012).

The assessment, administered in its entirety, is quite lengthy. In order to increase clinical usefulness, Hsieh et al. developed a 12-item short form based on the upper and lower extremity subscales of the FMA (Hsieh et al. 2007). Items were retained on the basis of representativeness of Brunnstrom staging and item difficulty assessed via Rasch analysis. Similarly, Crow et al. (2008) proposed a shortened method of administration for the upper and lower extremity portions of the FMA. Using Guttman analysis the authors determined that scale items in these two sections fulfill the statistical criteria for a valid hierarchy. Therefore, test administration may begin at a stage considered appropriate to the observed level of patient recovery. If a patient is awarded the maximum score for an entire stage, all items in previous stages may also be awarded a full score. Likewise, when the individual being tested fails to score for all of the scale items in a given stage, assignment of a score of 0 points for any remaining untested, more advanced, items. This method of assessment could represent a substantial reducion in the time required to perform the test. Full guidelines for hierarchical testing procedures are provided by Crow et al. (2008).

Limitations

Though a trained therapist should be able to administer the test in approximately 30 – 45 minutes, it may take considerably longer. Average reported times for administration of motor, sensation and balance range from 34 to 110 minutes with a mean time of 58 minutes (SD=16.6)(Malouin et al. 1994). The scales' relative complexity and length may make it less amenable to use in clinical practice (Poole & Whitney 2001) and may be associated with substantial patient burden, particularly in individuals experiencing difficulties with fatigue or endurance.

Van der Lee et al. (2001) suggested that, as an assessment of recovery within the context of the motor system, the FMA may separate motor recovery from functional recovery and, therefore, may not be responsive to functional improvements in chronic populations. However, significant associations of moderate strength between FMA-UE scores and scales that assess functional limitations in the upper extremity, such as the ARAT and WMFT, have been reported in groups of stroke patients during subacute and chronic phases (Hsieh et al. 2009; Lin et al. 2010, 2009; Wei et al. 2011).

The reliability and validity of the balance section (particularly sitting balance, see chart above) of the FMA has been shown to be questionable. Revisions to the scoring of the parachute items within the balance scale (Hsueh et al. 2001; Mao et al. 2002) appear to have resulted in an increase in reliability. However, further testing of the modification is required. Assessment of somatosensory impairment using the sensation subscale has also been criticized for lack of face validity, low construct and predictive validity in addition to poor responsiveness as evidenced by large ceiling effects and weak to moderate effect sizes (Lin et al. 2004).

Subsequent to principal components and Rasch analyses, it has been suggested that the three items measuring reflex (biceps reflex, triceps reflex, normal reflex activity) do not make a significant contribution to the assessment of upper extremity impairment (Woodbury et al. 2007). In addition, the item-difficulty hierarchy of a 30-item assessment (reflex items removed) produced by Rasch analysis appears better suited to understanding the progression of recovery in the upper extremity following stroke (Woodbury et al. 2007) and may be used to inform both short and longer term rehabilitation goals (Velozo & Woodbury 2011). Further analysis demonstrated that the item-difficulty hierarchy of these 30 items was stable over time and, therefore, provides a longitudinally valid assessment of upper extremity function (Woodbury et al. 2008).

Summary – Fugl-Meyer Assessment of Motor Recovery after Stroke

Interpretability: The interpretability of the FMA is enhanced by the scale's strong foundation in well-defined stages of motor recovery. It is widely used and internationally accepted. Classifications of severity of motor impairment by FMA score have been proposed by several sources (Duncan et al. 1994; Fugl-Meyer 1980; Fugl-Meyer et al. 1975). A clinically important difference of 5.25 has been suggested for the FMA-UE based on ratings of change in overall UE function in a group of individuals with chronic stroke and mild to moderate impairment (Page et al. 2012).
Acceptability: Administration of the entire test can be a lengthy process, however, when the motor scale is administered on its own, it takes approximately 20 minutes. As the test is scored via direct observation, it cannot be used with proxy respondents.
Feasibility: The FMA should be administered by a trained physical or occupational therapist. It requires no specialized equipment and can be administered across a variety of settings and has been tested for use in longitudinal assessments.

Table 20.2.6.1 Fugl-Meyer Assessment Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++(IO) ++ (IC –*balance)*	+++	+++ *(but note problems with balance & sensation subsections)*	+++	++ +++ (FMA-UE) + (FMA-S)	+ (FMA-S)

NOTE: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)

20.2.7 General Health Questionnaire – 28 (GHQ-28)

The General Health Questionnaire (GHQ) is a screening tool developed to detect possible cases of psychiatric disorders (McDowell & Newell 1996) and has been noted as “one of the most widely used questionnaires to screen for psychiatric morbidity” (Andersen et al. 2002). This self-administered questionnaire is not intended to be diagnostic, rather it serves to identify those who may require further psychiatric evaluation (McDowell & Newell 1996). Its aim is to uncover two main classes of problems: the inability to execute normal healthy functions and the manifestation of new distressing phenomena (Goldberg & Hillier 1979). The GHQ is concerned with four aspects of distress: depression, anxiety, social impairment, and hypochondriasis (McDowell & Newell 1996). The instrument is geared to detect deviations from ‘usual state’ by inquiring about the presence and magnitude of symptoms as compared to what is normal for that individual (McDowell & Newell 1996). Thus, the GHQ was not designed to detect long-standing phenomena (chronic illnesses) that have become ‘usual’ to the individual (Richard et al. 2004).

The GHQ-28 is one of several scaled variations of the original 60-item questionnaire. Based on a factor analysis of 523 completed GHQ-60 questionnaires, four 7-item subscales were created; somatic symptoms (A), anxiety and insomnia (B), social dysfunction (C) and severe depression (D) (Goldberg & Hillier 1979). Each subscale is scored separately to provide a profile of scores on 4 subscales. It was intended that this version be used in situations where it may be more helpful to have separate scores for each symptom area as opposed to a single severity score (Goldberg & Hillier 1979). The GHQ-28 has been recommended for detecting morbidity in posttraumatic clinical and research settings (Andersen et al. 2002).

The self-report questionnaire consists of 28 questions each representing a particular symptom. Respondents rate each question using the options provided (“better than usual”, “same as usual”, “worse than usual” and “much worse than usual”). Three different scoring methods can be used for any GHQ derivation. These are described in Table 20.13. Item scores for each subscale are summed. Subscale scores may be summed to provide a score out of 28 (for the GHQ and CGHQ scoring methods). Goldberg and Hillier (1979) claim that the conventional scoring method provides just as good if not better results than the Likert method, therefore they recommend this simpler method when using the GHQ for screening purposes. Regarding the GHQ and CGHQ scoring methods, results have been mixed as to which is most appropriate, however Richard et al. (2004) found that the choice of scoring method does lead to different individuals being labeled psychologically distressed; they conclude that it would be most advantageous to use both methods simultaneously and recognize all individuals that scored positive according to either system. This version of the GHQ takes approximately 3 to 4 minutes to complete, thus it is a relatively quick assessment (McDowell & Newell 1996).

Advantages

The GHQ-28 is a simple questionnaire to administer and score and it requires less time and energy from the patient than the original version, which is especially important for a physically or mentally ill population. Low refusal rates suggest that the questionnaire is not difficult for most individuals to complete.

The GHQ-28 provides useful subscores – unlike the other versions of the GHQ – so it may be possible to get a more accurate indication of the possible psychopathology (Kilic et al. 1997) or to identify certain mood disorders (Aylard et al. 1987; Lobo et al. 1988).

Table 20.2.7.1 Scoring methods used for the GHQ-28*

GHQ - conventional	Dichotomous system in which each symptom is rated as absent or present. The first 2 response options are scored as 0, the last 2 as 1.
Likert scoring	Assigns weight to each response based on symptom frequency. Responses are scored as 0,1,2,3.
Corrected GHQ	As for the GHQ method but, for items that indicate an illness or health problem, the response "same as usual" receives a score of 1 rather than 0. Scoring for other items remains unchanged.

**as described in McDowell and Newell (1996)*

Rabins and Brooks (1981) suggested that the total GHQ score can be used as a measure of severity; however, one must be cautious when making these interpretations as the intention of the test is to screen, not to make diagnostic implications. Lobo et al. (1988) and Rabins and Brooks (1981) have suggested that the total GHQ score can be used as a measure of severity. Lastly, Goldberg et al. (1997) found no significant differences in classification validity across gender, age, language or educational level, which suggests that the use of the GHQ-28 may be appropriate in many populations. Lincoln et al. (2003) comment that because the GHQ-28 provides an indication of "psychological distress" rather than depression, it may be more sensitive to the issues faced by the stroke population.

Limitations

Most psychometric evaluations of the GHQ-28 have been limited to sensitivity and specificity calculations and determination of construct validity. Very little information is available regarding the reliability of the measure. The GHQ has been translated into many languages including Italian, Cambodian, Mexican-Spanish, Japanese and Chinese (McDowell & Newell 1996). However, according to Kilic et al., reliability figures have been found to be higher in English-speaking countries, suggesting that issues related to translation and semantics may influence the reliability of the instrument (Kilic et al. 1997).

While the GHQ has been tested in many different populations, it has not been validated very well in the stroke population where it is frequently used. A common criticism of the GHQ, that is quite pertinent to stroke patients, is that it tends to miss the influence of chronic illness (O'Rourke et al. 1998) or confuse physical illness with psychiatric disturbance (Lykouras et al. 1996). Individuals suffering from a chronic illness may choose the option "same as usual" or "no more than usual" because their condition has remained the same for some time, not because the symptom is absent, thus they receive a negative score on that item (Benjamin et al. 1982). Furthermore, due to items on the somatic subscale, those with physical illnesses may score high on the GHQ which results in a misclassification of these individuals as possibly having a psychiatric disorder (Lykouras et al. 1996). The Corrected GHQ scoring method was proposed by Goodchild and Duncan-Jones (1985) to try to improve the GHQ's ability to detect chronic illness.

There has been some confusion surrounding the construct that is actually being measured by the GHQ; it has been described as a measure of psychiatric morbidity (Andersen et al. 2002), emotional morbidity (Lobo et al. 1988), psychological distress (Lincoln et al. 2003), non-psychotic mental illness (Burvill & Knuiman 1983) and psychiatric disturbance (Koeter 1992), which are all constructs that are difficult to define precisely. Also, while an advantage of the GHQ-28 is the fact that it provides subscores, it is important to realize that correlation can be considerable between the scales, so it is not appropriate to assume that they are distinct measures (Werneke et al. 2000).

The GHQ is a tool that attempts to separate those who probably do not have a psychiatric disorder from those who might have a psychiatric disorder; a score does not suggest a particular diagnosis, but expresses the likelihood of being a psychiatric case (McDowell & Newell 1996). Optimal threshold scores vary across studies, which can be affected by the 'gold standard' used for validation, the prevalence of disorder in the population and the population demographics, among other things (Furukawa et al. 2001). Many studies have found that using 4, 5 or 6 positive answers as the criteria for 'caseness' (using the traditional scoring method) results in adequate classification validity. Goldberg et al. claim that the mean GHQ score provides a rough estimate of the optimal threshold whereas Willmott et al. (2004) believe it is the median GHQ score that guides this estimate (Goldberg et al. 1998). However Furukawa et al. (2001) suggest using stratum-specific likelihood ratios (SSLRs) to interpret scores instead of the best threshold approach; nonograms – to aid in the computation of post-test probabilities – are provided in their study and online at http://www.epbcenter.com.

Summary – General Health Questionnaire – 28

Interpretability: Caution must be exercised in the interpretation of GHQ scores. The intention of the assessment is to screen for, not diagnose, psychiatric disturbance. While the cut-off of 5/6 is commonly used, it has not been validated as most appropriate in a stroke population. The sole study evaluating the use of the GHQ-28 as a screening tool for depression after stroke recommended the use of 11/12 for this purpose.

Acceptability: Most of the studies reported a very low refusal rate, suggesting that the instrument is acceptable to patients. The 28-item version takes half the time that the original version takes to complete, which may be more appropriate for a physically ill population. Assessment by proxy would not be acceptable for this instrument.

Feasibility: The GHQ is an inexpensive instrument that is simple to administer and score, especially if using a dichotomous scoring method. It is common practice to have the questionnaire filled out while the patient is in the waiting room, which makes it an efficient process for patient and clinician.

Table 20.2.7.2 GHQ-28 Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (IC)	+++	+++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.8 Geriatric Depression Scale (GDS)

The Geriatric Depression Scale was developed in 1982 by Brink and Yesavage. It was initially designed as a screening test to detect depression in elderly individuals and was intended to be short, simple and easy to use in primary care settings (McDowell & Newell 1996). The GDS is a self-rating scale comprised of 30 items selected from a pool of 100 items selected by researchers and clinicians for their validity in distinguishing groups of elderly, depressed people from the general population (McDowell & Newell

1996). Questions require simple yes/no answers and were intended to be both non-threatening and age-appropriate (Stiles & McGarrahan 1998).

The respondent is to provide responses to each question with reference to the past week. One point is given for each “yes” response and the number of points is summed to provide a single score. Scores from 0 to 10 are considered normal, while scores 11 indicate the presence of depression. Depression can be further categorized into mild (11 - 20) and moderate-severe (21 – 30) depression (McDowell & Newell 1996). The test requires approximately 8 – 10 minutes to complete in self-administered format (McDowell & Newell 1996). Oral administration by an examiner, however, might be more inclusive of a wider range of individual abilities (Stiles & McGarrahan 1998; van Marwijk et al. 1995).

Given the number of questions and length of time to administer, it has been suggested that the use of the GDS as a screening tool is impractical in primary care settings (van Marwijk et al. 1995). Many shorter versions of the GDS have been developed to address this potential difficulty. The 15-item version, developed by Sheikh and Yesavage (1986) is the most commonly used short form. The response and scoring format were retained from the original version. Scores of 0 – 4 are considered normal, while scores of 5 – 9 indicate the presence of mild depression and scores of 10 – 15 indicate the presence of moderate to severe depression (McDowell & Newell 1996). It requires approximately 5 – 7 minutes to administer. One, three, four, five and ten item versions of the Geriatric Depression Scale have also been evaluated for use in screening for the presence of depression (Almeida & Almeida 1999; MacNeill & Lichtenberg 2000; Rinaldi et al. 2003; van Marwijk et al. 1995).

Advantages

The GDS focuses on affective aspects of depression rather than somatic components, which may not be useful indicators of depression in the elderly. When used as a screening tool, it performs as well as some longer, interview-based assessments but requires much less time and training to administer.

Limitations

In general, the GDS has been found to have better specificity and sensitivity among higher functioning, community dwelling subjects (Stiles & McGarrahan 1998). Reports of its ability to screen for depression when used with cognitively impaired individuals have been varied possibly due to the emphasis placed upon short-term memory and personal insight by the self-report format of the GDS. In one instance, the GDS was reported to perform no better than chance in screening for depression among the cognitively impaired elderly (Burke et al. 1989). It has been suggested that the GDS should not be used with patients who have more than a moderate cognitive impairment (Kafonek 1989; McDowell & Newell 1996; McGivney et al. 1994; Stiles & McGarrahan 1998).

Although oral administration may include individuals with a wider range of abilities, among those with higher levels of cognitive ability, the oral method of administration may result in the endorsement of fewer items when compared to the written method of administration (Cannon et al. 2002). The need to provide an answer aloud may discourage some respondents from providing an answer they may consider embarrassing (Williams et al. 2005).

Gender may have an effect on the ability of the GDS to correctly classify individuals. The GDS has been reported to be more accurate in classifying women as more depressed than men. In the case of male respondents, there tend to be more false negatives (Stiles & McGarrahan 1998).

While many of the shortened versions of the GDS have been found to be highly correlated with the original, the short forms tend to have higher negative predictive values suggesting that the short forms might be best suited to screening out or excluding possible cases (Almeida & Almeida 1999; van Marwijk et al. 1995).

Summary – Geriatric Depression Scale

Interpretability: Currently, there is no standardized format for administration and many different short-forms comprised of different sets of question making comparisons difficult between studies or groups.
Acceptability: The items were developed specifically for an elderly population. The yes/no response format is easy to understand and familiar. Shorter versions are available to attenuate potiential problems of attention and fatigue. The GDS has been evaluated for use with proxy respondents.
Feasibility: The GDS is easy to administer and requires no additional training. It is not suited for use with patients who are cognitively impaired. The 30-item version may be too long to be of practical use in primary care settings.

Table 20.2.8.1 Geriatric Depression Scale Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IC)	+++	+++	n/a	n/a	n/a

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.9 Hospital Anxiety and Depression Scale (HADS)

The Hospital Anxiety and Depression Scale (HADS) is a bi-dimensional scale developed specifically to identify cases of depression and anxiety disorders among physically ill patients (Bjelland et al. 2002; Flint & Rifat 2002; Herrmann 1997; Zigmond & Snaith 1983). As such, the scale was intended to detect both depression and anxiety without the possible confounding influence of somatic symptomatology that could be attributed to physical illness rather than psychological states. Items that address somatic symptomatology of depression, such as fatigue, weight loss or headache are not included.

The total HADS consists of 14 items, which can be divided into two subscales of seven items each: the anxiety subscale (HADS-A) and the depression subscale (HADS-D). Anxiety items reflect the state of generalized anxiety while most (5 of 7) items on the depression subscale focus on the concept of anhedonia (Flint & Rifat 2002; Roberts et al. 2001). The respondent rates each item on a 4-point scale ranging from 0 (absence) – 3 (extreme presence). Five of the 14 items are coded in reverse. Scores are derived by summing responses for each of the two subscales or for the scale as a whole.

The total scale score is out of 42 or 21 for each of the subscales. Higher scores indicate greater levels of anxiety or depression. The total HADS score may be regarded as a global measure of psychological distress (Johnston et al. 2000; Roberts et al. 2001). Examination of sensitivity and specificity in individuals with remitted, not fully remitted and current major depressive episodes revealed ranges of scores associated with four categories of severity or depressive states (Hung et al. 2012). Scores ranging from 0-7 may be interpreted as normal (or full remission), 8-10 as mild depression (or partial remission), 11-14 as moderate (or lower than average severity for a major depressive episode) and 15-21 as severe (or higher than average severity for a major depressive episode) (Hung et al. 2012).

The test can be completed in approximately 2 – 6 minutes and can be scored in approximately one minute, with practice (Herrmann 1997; Visser et al. 1995). No training is required to score or administer the test. Although the test is freely available, commercial use requires permission and/or purchase of the test questionnaires (from: www.nfer-nelson.co.uk).

Advantages

The HADS is simple to administer and score and requires no specialized psychiatric training to use. It is widely used and has been translated into a wide variety of languages (Pais-Ribeiro et al. 2007) (from:http://shop.nfer-nelson.co.uk/icat/hospitalanxietyanddepress). Administration of the HADS appears to be well tolerated by medical patients who may be quite unwell (Herrmann 1997; Johnston et al. 2000). In addition, evaluation of telephone administration suggest no significant difference in results obtained via telephone interview when compared to face-to-face administration in group of individuals with stroke (Hoffmann et al. 2010).

Total scale scores may be indicative of psychological distress rather than depression per se (Johnston et al. 2000; Roberts et al. 2001). However, total scale scores have been reported to be similarly sensitive and specific in screening for the possible presence of depression as the depression subscale scores alone (Aben et al. 2002). This may be a reflection of the moderately strong correlation that exists between the two scales, despite its confirmed 2-factor structure (Bjelland et al. 2002; Flint & Rifat 2002; Helvik et al. 2011; Johnston et al. 2000; Marinus et al. 2002; Roberts et al. 2001).

Limitations

One item, “I feel as if I am slowed down”, has been identified as problematic (Flint & Rifat 2002; Helvik et al. 2011; Johnston et al. 2000). It does not belong definitively to either subscale, and in fact, may be interpreted as a somatic symptom. Elderly patients, in particular, may endorse this item if they interpret “slowed down” as representative of the physical slowing attributable to age or physical ailments (Flint & Rifat 2002).

While exclusion of somatic items may be effective in preventing inflated scores among the physically ill, it may also represent a reduction in the face validity of the scale (Marinus et al. 2002). As Marinus et al. pointed out, five of the nine criteria for depression included in the DSM-IV reflect somatic symptomatology (Marinus et al. 2002). None are represented on the HADS. The HADS assessment of depression focuses on the core symptoms of mood and anhedonia only. By way of contrast, the BDI incorporates 6 of 9 DSM-III criteria for depression (Beck et al. 1988).

Summary – Hospital Anxiety and Depression Scale

Interpretability: No norms are available in English. Percentiles and t-scores are available for the German version. No standardization for age or gender has been performed and cutoff points used are not particularly well established.

Acceptability: The scale is quick and easy to use. It has been reported to be well-tolerated by patients who may be quite unwell.

Feasibility: The HADS is simple to use and score. No specialized training is required to administer the scale.

Table 20.16 HADS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) ++ (IO) ++ (IC)	+++	++	+	+	+++

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.10 Line Bisection Test (LBT)

The Line Bisection Test (LBT) is a screening tool for unilateral neglect (ULN). Occurring more frequently in response to right hemisphere injury, this condition is characterized by failure to respond to stimuli located in extrapersonal space contralateral to the lesion site (Ferber & Karnath 2001). The LBT has been in use for over 70 years. However, it was only more recently that Schenkenberg et al. (1980) formally validated the measure.

During the LBT, patients are required to mark, in pencil, the centre-points on a series of horizontal lines presented on a sheet of paper. The LBT is scored by measuring patients' average deviation (in centimeters or millimeters) from the true centre-point of the line. Most testers utilize a formula that divides the deviation by half the length of the line and then multiplies this quotient by 100 to yield a percentage. ULN is diagnosed when the markings deviate, on average, from the true-centre-point beyond a pre-determined cut-off value. It is important to note that there seems to be no standard for this value within the literature. Typically it is defined as the lowest score of any control in given study.

The test takes under 5 minutes to administer and requires no specialized training for the tester.

Advantages.

The LBT is a versatile test in that can be used as part of the behavioural inattention test battery for improved sensitivity, or on its own as a more convenient, bedside screen for unilateral spatial neglect. Used as the latter, the test is economical in both time and cost, taking roughly 5 minutes to complete and requiring only a pencil and the test paper as materials. There is also a virtual reality version of the test available; however, it has demonstrated only moderate agreement with the conventional LBT (Fordell et al. 2011).

Limitations.

The LBT seems unable to discriminate between unilateral neglect and visual field defects, such as hemianopia. This complication arises from the fact that the LBT measures a set of cognitive processes (i.e., correct perception of the size of a single stimulus) that are also impaired in visual field defects. The finding that hemianopic patients without neglect consistently make errors by bisecting lines on the side contralateral to their lesion is well established (Ferber & Karnath 2001). Thus, a positive score on the LBT can only be taken as a certain indicator of ULN once the confounding role of these related disorders has been ruled out.

Another source of criticism towards the LBT has come from Ferber and Karnath who argue that the cognitive skills assessed by this test are correlated with spatial neglect, but not fundamentally associated with it (Ferber & Karnath 2001). In their research, Ferber and Karnath (2001) compared the sensitivity of the LBT to that of several cancellation tests in a sample of 35 individuals with well-defined spatial neglect. They found that the LBT missed 40% of the cases, while letter cancellation and bells tests missed only 6%. Coupled with a number of studies that have found double dissociations where impairment is found in cancellation tests but not the LBT or vice versa (Ferro & Kertesz 1984; Halligan et al. 1991; Marshall & Halligan 1995), the authors interpret their findings as evidence that LBT performance is not fundamentally related to spatial neglect. In light of this, they recommend that LBT results should be treated with caution in clinical settings and suggest that cancellation tests may be more helpful tools in detecting spatial neglect.

Further evidence for this argument has come from studies specifically comparing performance on the LBT with that on cancellation tests. These studies have found either weak correlations or no correlation at all between the tests (Binder et al. 1992; Ferber & Karnath 2001). Finally, a factor analysis conducted on a battery of neglect tests found line bisection to be a factor on its own, which was not included in the factor containing letter or symbol cancellation (McGlinchey-Berroth 1991).

Summary – Line Bisection Test

Interpretability: The LBT is a simple, quantitative screening tool for unilateral neglect. Test administration is problematic in terms of standardization, as there is a lack of consistency in the literature with respect to both method and scoring of the test. Specifically, the length of lines, the number of lines and the means of determining a cut-off all tend to differ.

Acceptability: The test is brief and represents little burden to the patient.

Feasibility: The LBT is simple to administer and does not require specialized training. The only materials required for the test are a pencil and the test paper.

Table 20.2.10.1 LBT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR)	++	++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.11 Mini-Mental State Examination (MMSE)

The Mini-Mental State Examination was developed as a brief screening tool to provide a quantitative assessment of cognitive impairment and to record cognitive changes over time (Folstein et al. 1975). While the tool's original application was the detection of dementia within a psychiatric setting, its use has become widespread.

The MMSE consists of 11 simple questions or tasks. Typically, these are grouped into 7 cognitive domains; orientation to time, orientation to place, registration of three words, attention and calculation, recall of 3 words, language, and visual construction. Administration by a trained interviewer takes approximately 10 minutes. The test yields a total score of 30 and provides a picture of the subjects present cognitive performance based on direct observation of completion of test items/tasks. A score of 23/24 is the generally accepted cutoff point indicating the presence of cognitive impairment (Dick et al. 1984). Levels of impairment have also been classified as none (24-30); mild (18-24) and severe (0-17) (Tombaugh & McIntyre 1992).

An expanded version of the MMSE, the modified mini-mental state examination (3MS) was developed by Teng and Chui increasing the content, number and difficulty of items included in the assessment (Teng & Chui 1987). The score of the 3MS ranges from 0 – 100 with a standardized cut-off point of 79/80 for the presence of cognitive impairment. This expanded assessment takes approximately 5 minutes more to administer than the original MMSE.

Advantages

The Mini-mental State Examination is brief, inexpensive and simple to administer. Its widespread use and accepted cut-off scores increase its interpretability.

Limitations

It has been suggested that the MMSE may attempt to assess too many functions in one brief test. An individual's performance on individual items or within a single domain may be more useful than interpretation of a single score (Tombaugh & McIntyre 1992; Wade 1992). However, when used to screen for visual or verbal memory problems or for problems in orientation or attention, it is not possible to identify acceptable cut-off scores (Blake et al. 2002).

Perhaps the greatest limitation of the MMSE is its low reported levels of sensitivity particularly among individuals with mild cognitive impairment (de Koning et al. 1998; Tombaugh & McIntyre 1992) and in patients with right-sided lesions within a general neurological patient population (Dick et al. 1984) and within a stroke population (Blake et al. 2002; Nys et al. 2005; Suhr & Grace 1999). A single study by Tang et al. (2005) suggested that, as a screening instrument for dementia, it may perform with acceptable levels of sensitivity and specificity among patients with lacunar infarcts and using an adjusted cut-off score of 18/19. It has been suggested that the low level of sensitivity associated with use of the MMSE derives from the emphasis placed on language items and a paucity of visual-spatial items (de Koning et al. 1998; Grace et al. 1995; Suhr & Grace 1999). Various solutions have been proposed to the problem of the MMSE's poor sensitivity including the use of age-specific norms (Bleecker et al. 1988) and the addition of a clock-drawing task to the test (Suhr & Grace 1999). Clock-drawing tests themselves have been assessed as acceptable to patients, easily scored and less affected by education, age and other non-dementia variables than other very brief measures of cognitive impairment and would have little effect on the simplicity and accessibility of the test (Lorentz et al. 2002).

MMSE scores have been shown to be affected by age, level of education and sociocultural background (Bleecker et al. 1988; Lorentz et al. 2002; Tombaugh & McIntyre 1992). These variables may introduce bias leading to the misclassification of individuals. Improved classification sensitivity and specificity have been reported when scores are adjusted for these recognized confounders. In a group of stroke patients, Godefroy et al. (2011) reported sensitivity of 70% and specificity of 97% based on adjusted scores using a cut-off of ≤24. It should be noted that not all studies have demonstrated bias associated with age or education (Agrell & Dehlin 2000) and concern has been expressed that the need to make adjustments for these biases may limit the general utility of the MMSE (Lorentz et al. 2002). Bour et al. reported good classification sensitivity/specificity for cognitive impairment and dementia post stroke, with no adjustments for age or education (Bour et al. 2010). In addition, MMSE scores were predictive of cognitive impairment and dementia on follow-up.

Summary – Mini Mental State Examination

Interpretability: The MMSE is widely used and has generally accepted cutoff scores indicative of the presence of cognitive impairment. Documented age and education effects have led to the development of stratified norms (Crum et al. 1993).
Acceptability: The test is brief requiring approximately 10 minutes to complete. It may be affected by such patient variables as age, level of education and sociocultural background. As it is administered via direct observation of task completion, it is not suitable for use with a proxy respondent.
Feasibility: The test requires no specialized equipment and little time, making it inexpensive and portable. A survey conducted by Lorentz et al. (2002) revealed participant physicians found the MMSE too lengthy and unable to contribute much useful information.

Table 20.2.11.1 MMSE Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) ++ (IO) ++ (IC)	+++	++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.12 Modified Ashworth Scale (MAS)

The Ashworth scale was originally developed to assess the efficacy of an anti-spastic drug in patients suffering from multiple sclerosis (Ashworth 1964). The scale is used to assign a subjective rating of the amount of resistance or tone perceived by the examiner as a limb is moved through its full range of motion.

The original Ashworth scale consisted of 5 grades from 0 – 4. In 1987, Bohannon and Smith added one grade (1+) and revised the wording of the scale (see below) in an attempt to make the scale more sensitive (Bohannon & Smith 1987; Gregson et al. 2000; Pandyan et al. 1999). Changes to wording incorporated approximations of how much resistance was perceived and at what point during the motion resistance was felt (Damiano et al. 2002).

Table 20.2.12.1 Modified Ashworth Scale for grading spasticity

Grade	Description
0	No increase in muscle tone.
1	Slight increase in muscle tone, manifested by a catch and release, or by minimal resistance at the end of range of motion when the affected part(s) is moved in flexion or extension.
1^+	Slight increase in muscle tone, manifested by a catch followed by minimal resistance throughout the remainder (less than half) of the range of movement (ROM).
2	More marked increase in muscle tone through most of ROM, but affected part(s) easily moved.
3	Considerable increase in muscle tone, passive movement difficult.
4	Affected part(s) rigid in flexion or extension.

Ref: Bohannon and Smith (1987)

A graded rating of spasticity is made from 0 – 4, using the guidelines appearing in the above table to describe the resistance perceived while moving a limb passively about a joint, through its full range of motion, for one second (Pandyan et al. 1999, 2001).

Advantages

The modified Ashworth scale has gained widespread clinical acceptance. It is routinely used to assess spasticity and indeed, is the current clinical standard (van Wijck et al. 2001).

Limitations

There remains some question as to whether the Ashworth scale is a valid measure of spasticity. It has been suggested that the scale, in either form, is a descriptive assessment of resistance to passive movement (RTPM), and as such, reflects only an aspect of spasticity rather than providing a comprehensive measurement (Pandyan et al. 1999, 2001) while Damiano et al. (2002) found Ashworth scores to be more closely related to measurements of stiffness than to magnitude of resistance. Patrick and Ada (2006) suggested that the Ashworth Scale makes no distinction between spasticity and contracture and, in fact is counfounded by contracture. Pandyan et al. (2003) suggest that even taken as a measure of resistance to passive movement the Ashworth scale lacks sensitivity in that grades 1, 1+ and 2 are not discriminative of change. As such, the authors recommend merging these 3 levels into one.

In studies of post stroke patients, the most common ratings reported are 0, 1 & 1+ (Blackburn et al. 2002; Pandyan et al. 1999, 2001) and the highest levels of inter-observer and intra-observer agreement are noted among patients with a 0 rating. In a 1999 review, Pandyan et al. noted that the reduction of reliability in the Modified Ashworth Scale centers on disagreements around 1 and 1+ ratings (Pandyan et al. 1999). The greater degree of discrimination introduced to the scale by Bohannon and Smith may be accompanied by a reduction in the scale's reliability (Bohannon & Smith 1987; Damiano et al. 2002; Haas et al. 1996). In addition, Naghdi et al. (2008) reported that the ordinal relationship between 1 and 1+ ratings was lost when scores were compared to the Hslp/Mslp ratio (a neurophysiological measure). Ansari et al. (2006) have proposed a modified version of the MAS in which the problematic 1+ rating is eliminated. Evaluations of the MMAS in small patient samples suggest adequate to excellent interobserver reliability (κ=0.63 – 0.89) when the MMAS is used in the assessment of wrist and elbow flexors and knee extensors (Ansari et al. 2009, 2008; Ghotbi et al. 2011; Kaya et al. 2011; Naghdi et al. 2007). Further study of this latest revision to the MMAS using larger groups of patients is required to determine whether elimination of the 1+ rating has resulted in improved ordinal relationships between scores.

No standardized testing procedures or guidelines for the use of the scale exist. Given the ambiguity of wording used within the scale and the inherently subjective nature of the rating, development of standard procedure for assessment of spasticity using the Ashworth scale may contribute to increased levels of reliability (Gregson et al. 1999, 2000). However, standardized guidelines may not be an adequate solution. Blackburn et al. (2002) reported poor levels of interrater reliability despite the use of written guidelines. In this study, the assessors had not been trained specifically in the use of the scale suggesting that guidelines need to be accompanied by training of test administrators to achieve improved reliability (Blackburn et al. 2002).

Reliability of the MAS is dependent upon the muscle being assessed. In general, the MAS may be best suited to assessments of the elbow, wrist and knee flexors (Gregson et al. 2000; Pandyan et al. 1999). Assessments of ankle plantarflexors often demonstrate low levels of reliability (Gregson et al. 2000; Haas et al. 1996; Pandyan et al. 1999). Given the reported variability in reliability, it would not be advisable to combine scores from individual muscle assessments to provide a rating of global spasticity for a given patient. Such summation would mask unreliability arising from individual scores (Pandyan et al. 1999). In addition, Ansari et al. (2006) suggest that repeated stretching may introduce variability and make reliable grading of spasticity more difficult. Although for the purposes of their own study, the authors used three passive stretches for each rating, they suggest that clinicians should use only one (Ansari et al. 2006).

Summary – Modified Ashworth Scale

Interpretability: The original Ashworth and Modified Ashworth scales are the primary clinical measures of tone. Despite lower levels of reliability, they are widely used and accepted. Ambiguity of wording and lack of standardized procedures limit the scales' usefulness for comparison across studies as well as reliability.
Acceptability: While testing should be relatively brief, manipulation of the affected limb/joint may be uncomfortable for patients.
Feasibility: No specialized equipment is required, however, training of test administrators and standardization of test procedures is essential to the reliability of the MAS.

Table 20.2.12.2 Modified Ashworth Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	++(TR) ++(IO)	+	++	+	++	N/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.13 Montreal Cognitive Assessment (MoCA)

The Montreal Cognitive Assessment (MoCA) is a brief screening tool designed to detect mild forms of cognitive impairment not captured by other cognitive screening instruments. Administration of the MoCA involves completion of several independent tasks used to assess the following six domains: memory, visuospatial ability, executive functioning, attention and concentration, language, and orientation (see Table 20.20). The MoCA yields a total score out of 30 with scores of 26 or lower indicating the presence of cognitive impairment.

Table 20.2.13.1 Task description and scoring procedure for the MoCA*

Domain	Task Descriptions	Scoring
Memory	The patient is required to repeat five words (face, velvet, church, daisy, and red) during two learning trials and a scored delayed recall trial (approximately five minutes following the learning trials).	One-point is awarded for each correctly remembered word during the delayed recall trial for a total of 5 points. No points are awarded if the patient requires queing.
Visuospatial	The patient is required to draw a clock and copy a three-dimensional figure (a cube)	A total of three-points are awarded for correctly drawn clocks, with contour, numbering, and hand positioning each worth one-point. For the cube copy, one-point is awarded for cubes that have the correct number and positioning of lines.
Executive functioning	Executive functioning is assessed with an alternation task (trailing a line from 1 to A, 2-B, etc.) and a two-item verbal abstraction task (identify the similarity between two word pairs).	One-point is awarded for correctly completing the alternation task with no errors. Two-points are awarded for the abstraction task, one for each of the word pairs.
Attention and concentration	This domain is evaluated with a sustained attention task (identify when an ‘A’ is read in a list of letters), a serial subtraction task (serial 7 subtraction from 100), and a forward and backward digit repetition.	For the sustained attention task, one-point is awarded as long as no more than two errors are made. A maximum of two-points can be awarded for correct repetition of the two digit sequences. For the serial subtraction task, three-points are awarded for 4-5 correct subtractions, two-points for 2 or 3 correct subtractions, and one-point for 1 correct subtraction.
Language	Language is assessed with a naming task of low-familiarity animals (lion, rhino, and camel), repetition of two complex sentences, and a phonetic fluency task (patients must name as many words that begin with the letter F as they can in one minute)	One-point is awarded for each correctly identified animal, for a total of three-points. One-point is also awarded for each correctly repeated sentence, for a total of two-points. For the fluency task, one-point is awarded if more than ten “F” words are identified in the allotted time.
Orientation	The patient is required to identify the date (month, year day) as well as their current location (place and city).	A maximum of six-points can be awarded for this domain, with one-point for correct identification of the date, month, year, day,

		place, and city.

** from Nasreddine et al. 2005*

Advantages

The MoCA can be used to detect mild forms of cognitive impairment in patients that score in the normal range on other assessment measures (Nasreddine et al. 2005). For example, Pendlebury et al. (2010) administered both the MoCA and the MMSE to 413 patients following a stroke or TIA and reported that 58% of those who scored in the normal range on the MMSE (≥27) scored in the cognitively impaired range on the MoCA. Similarly, MacKenzie et al. (2011) reported that of 20 patients with TIA or mild stroke 90% of patients scored in the normal range (≥26) of the MMSE, while only 45% scored in the normal range on the MoCA (≥26).

The MoCA is brief, available free of charge, has been translated into more than 30 languages and requires little training to administer. Pendlebury et al. (2013) found that the telephone assessement of cognition using the MoCA was feasible and reliable. However, it was limited in its ability to examine visuoexecutive and complex language tasks compared to the traditional face to face assessment.

Limitations

The validity of the MoCA has not been thoroughly tested; in particular, there is limited information regarding its use in the post-stroke population. Some concerns have also been noted regarding the cut-off scores recommended by the scales authors. Specifically, using the recommended cut-off score, the specificity of the MoCA has been found to be much lower was reported in the original validation study (Luis et al. 2009; Smith et al. 2007). Consequently, Luis et al. (2009) suggested that the sensitivity and specificity of the MoCA are optimized when a lower cut-off score of ≤ 23 for the identification of impairment is used. However in a recent study in a population of stroke patients, Godefroy et al. reported an optimal cut-off score of ≤24 and an associated sensitivity and specificity of 70% and 97% respectively (based on scores adjusted for age and education) (Godefroy et al. 2011). Also in a group of stroke patients, Dong et al. (2010) identified an even lower optimal cut-off of ≤21with sensitivity and specificity of 90.3% and 76.8%. It should be noted that Dong et al. (2010) were using the Singaporean version of the MoCA. In Godefroy et al. (2011) and Dong et al. (2010), the optimal cut-off for the MMSE was identified at ≤24. In both cases, the MMSE demonstrated slightly lower sensitivity but greater specificity than the MoCA, when using the identified optimal cut-off and adjusted scoring.

Summary – Montreal Cognitive Assessment

Interpretability: Recommended cut-off scores can be used to identify individuals with mild cognitive impairment.
Acceptability: The MoCA is brief, requiring only 10 minutes to complete. Assessment can not be completed by a proxy respondent.
Feasibility: The MoCA is portable, requires no specialized equipment, and is available for use free of charge at www.mocatest.org.

Table 20.2.13.2.– MoCA Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) ++ (IC)	++	+++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;*

20.2.14 Motor-free Visual Perception Test (MVPT)

Originally developed for use with children (Colarusso & Hammill 1972), the Motor-free Visual Perception test (MVPT) measures visual perceptual skills in 5 areas; spatial relations, visual discrimination, figure–ground discrimination, visual closure and visual memory. The test consists of 36 items involving 2 dimensional configurations presented on separate cards or plates. Each plate consists of an example and a multiple choice response set of 4 alternatives (A,B,C,D) from which to choose the item that matches the example. The subject points to or says the letter that corresponds to the desired answer option (Mercier et al. 2001; Su et al. 2000). Standardized guidelines have been developed for the administration and interpretation of the test within an adult population, though the original test plates and manual are still required for administration (Bouska 1982). The test takes approximately 10 -15 minutes to administer.

One point is given for each correct response. Scores range from 0 to 36. In addition to summary scores, the time to complete each item is noted and an average time per item calculated. The test takes approximately 5 minutes to score (Brown et al. 2003). Normative data (U.S.) is available for adults aged 18 – 80 (Bouska 1982) and normative data specific to older adults (aged 50+) has been proposed (Mercier et al. 2001).

Advantages

The Motor-free Visual Perception Test is a widely used, standardized test of visual perception (Mazer et al. 1998). It is both simple and well tolerated by subjects (Su et al. 2000). Although originally developed for use in paediatric populations, age-specific norms are available for adults allowing for appropriate adjustments for age (Mazer et al. 1998).

Horizontal and vertical presentations are available for use. The vertical version removes unilateral visual neglect as a variable in test performance (Mazer et al. 1998) while maintaining high levels of reliability (Mazer et al. 1998). However, elimination of this variable may not always be desirable, as in a test of driving ability (Mazer et al. 1998).

Limitations

The MVPT provides a global score and, therefore, less information about specific visual dysfunction than a scale providing domain-specific scores (Su et al. 2000).

Summary – Motor-free Visual Perception Test

Interpretability: The MVPT is widely used in many populations. Age-specific norms are available for adults and older adults.

Acceptability: The test is short (15 minutes), simple and it is reported as well tolerated by subjects (Su et al. 2000). The test is administered via direct observation of task completion and is not suited to proxy use.

Feasibility: Administration requires the standardized instructions for administration in an adult population, test plates and manual.

Table 20.2.14.1 MVPT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IC)	++	++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.15 National Institutes of Health Stroke Scale (NIHSS)

The NIHSS is a measure of the severity of symptoms associated with cerebral infarcts and is used as a quantitative measure of neurological deficit post stroke. It is widely used and can be administered rapidly following acute admission (Anemaet 2002; Schlegel et al. 2004).

The NIHSS is a composite scale derived from items appearing on the Toronto Stroke Scake, the Oxbury Initial Severity Scale, the Cincinnati Stroke Scale and the Edinburgh-2 Coma Scale (Brott et al. 1989). Additional items were selected based on the clinical expertise of investigators from the NINDS stroke treatment studies (Brott et al. 1989). In all, the NIHSS consists of 15 items used to assess severity of impairment in LOC, ability to respond to questions and obey simple commands, papillary response, deviation of gaze, extent of hemianopsia, facial palsy, resistance to gravity in the weaker limb, plantar reflexes, limb ataxia, sensory loss, visual neglect, dysarthria and aphasia severity (Anemaet 2002; Brott et al. 1989; Heinemann et al. 1997; Schlegel et al. 2004). Items are graded on a 3 or 4 point ordinal scale on which 0 represents no impairment (Brott et al. 1989; Heinemann et al. 1997). Total scores range from 0 – 42. Higher scores reflect greater severity. Stroke severity may be stratified on the basis of NIHSS scores as follows: >25 = very sever, 15 – 24 = severe, 5 – 14 = mild to moderately severe and 1 – 5 = mild impairment (Anemaet 2002; Brott et al. 1989).

Brott et al. (1989) reported a mean administration time of 6.6 minutes over 48 examinations using the NIHSS.

Advantages

Administration of the NIHSS is both quick and simple. Like the CNS, use of the NIHSS is not restricted to neurologists. Reliable use of the NIHSS has been reported when used by both non-neurologist physicians and experienced nursing staff (Brott et al. 1989; Goldstein & Samsa 1997; Josephson et al. 2006). Furthermore, Kerr et al. (2012) found that NIHSS was sensitive to change as early as after 7 days post stroke. Modified versions of the NIHSS, including a shortened version (Lyden et al. 2009) and a plain English adaptation (Dancer et al. 2009), have demonstrated excellent reliability and strong concurrent validity with the original scale. Demaerschalk et al. (2012) found conducting NIHSS assessment through real time video smartphones had excellent reliability.

Certification in the use of the NIHSS is required for participation in many clinical trials and is recommended to maintain reliable assessment practices. A training and certification DVD was produced in 2006 and is available from several professional bodies including the American Academy of Neurology, the American Heart Association and the National Stroke Association (Lyden et al. 2009). A recent study has demonstrated that, for users from North America in particular, the DVD is a valid and reliable tool for training and certification for individual, group and website users (Lyden et al. 2009).

Limitations

Good reliability is dependent upon the use of trained raters and standardized application of the rating scale (Schmulling et al. 1998). Training using videotapes has been shown to be effective in achieving moderate to excellent reliability (Lyden et al. 2009). However, once trained and certified, repeated use and re-certification may not necessarily result in improved reliability (Josephson et al. 2006).

Poor agreement for the item “limb ataxia” has been reported repeatedly (Dewey et al. 1999; Goldstein et al. 1989; Millis et al. 2007; Schmulling et al. 1998). Some research has demonstrated via factor analysis that this item did not correlate well with any of the identified scale factors and it has been recommended that this item be considered for elimination (Dewey et al. 1999; Lyden et al. 1999; Millis et al. 2007). Based on results of factor analysis, Lyden et al. (1999; 2001) proposed a scale revision that eliminated this item as well as several other that had demonstrated poor item loadings on identified factors. Zandieh et al. (2012) however, reported 4 factors as a result of principal components analysis rather than the more commonly reported 2 factors. In that solution, the ataxia item along with visual field were associated with a single factor that the authors suggest may reflect deficits associated with posterior circulation strokes.

Many scale items are not testable in patients that have experienced severe stroke (Muir et al. 1996). Based on Brott et al.'s original summary of testability and incidence of impairment for each item, Heinemann et al. (1997) suggest that many appear to have limited utility. Some have a high proportion of patients rated as normal of the first testing while other have a high proportion of patients listed as untestable (e.g. limb ataxia).

The NIHSS may favour assessment of left hemisphere strokes; 7 of 42 possible points are related to language function while only 2 points describe neglect functions (Meyer et al. 2002; Woo et al. 1999). In the proposed revision by Lyden et al. (2001) the dysarthria item has been removed. Meyer et al. (2002) suggest that this may help to decrease the lateralization bias of the assessment. However, sebsequent analysis has demonstrated that 14/15 scale items (ataxia item excepted) function differently when used to assess patients with left vs. right hemisphere lesions (Millis et al. 2007). In this study, Rasch analysis revealed varying person and item separation statistics as well as rank item orders across lesion location groups (right vs. left). The authors suggest that interpretation of information gathered from the administration of the NIHSS might be enhanced if the total score were supplemented by the Rasch-transformed score corresponding to the side of lesion (Millis et al. 2007).

When used for retrospective evaluation, scoring is difficult. Lower reliability and item completion rates have been reported than for the CNS (Anemaet 2002; Bushnell et al. 2001). When used for this purpose, ratings should be based on evaluation reports from a neurologist (Bushnell et al. 2001).

Summary – NIHSS

Interpretability: The NIHSS is a widely used rating tool that provides a quantitative measure of neurological deficit post stroke. Using the NIHSS, stroke severity may be classified as very severe, severe, mild to moderately severe and mild.

Acceptability: The assessment may be completed in approximately 6 minutes and should represent little patient burden.

Feasibility: While the assessment need not be completed by a neurologist, training and standardized procedures are recommended to maintain scale reliability. The scale is freely available for use. Use of the NIHSS for retrospective evaluation is less reliable than the CNS and should be based on evaluations performed and reported by a neurologist.

Table 20.2.15.1 Evaluation Summary NIHSS

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	++ (TR) ++ (IO) + (IC)	+++	+++	+	+	+ (lg. % score normal or are untestable)

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;varied (re. floor/ceiling effects; mixed results)*

20.2.16 Orpington Prognostic Scale (OPS)

The Orpington Prognostic Scale (Kalra & Crome 1993) is a simple, objective, bedside evaluation, which provides a clinically derived baseline assessment of stroke severity that can be used as a predictor of outcome in elderly stroke patients (Kalra et al. 1994). The assessment includes measures of motor deficit (arm), proprioception, balance and cognition. It is based on an earlier prognostic tool, the Edinburgh Prognostic Score (Prescott et al. 1982) but adds an assessment of cognitive dysfunction (Kalra & Crome 1993). The Orpington Prognositic Scale is presented in Table 20.2.16.1.

Table 20.2.16.1 – Orpington Prognostic Scale

Clinical Features	Score
A. Motor deficit in arm *(Lying supping, patient flexes shoulder to 90 °and is given resistance)* MRC grade 5 (Normal power) MRC grade 4 (Diminished power) MRC grade 3 (Movement against gravity) MRC Grade 1 – 2 (Movement with gravity eliminated or trace) MRC Grade 0 (No movement)	 0.0 0.4 0.8 1.2 1.6
B. Proprioception (eyes closed) *(Locates affected thumb)* Accurately Slight difficulty Finds thumb via arm Unable to find thumb	 0.0 0.4 0.8 1.2
C. Balance Walks 10 feet without help Maintains standing position Maintains sitting position No sitting balance	 0.0 0.4 0.8 1.2
D. Cognition Based on administration of Hodkinson's Mental Test Mental test score 10 Mental test score 8-9 Mental test score 5-7 Mental test score 0-4	 0.0 0.4 0.8 1.2
Hodkinson's Mental Test *(Score one point for each question answered correctly)* Age of patient Time (to the nearest hour) Address given for recall at the end of the test (42 West Street) Name of hospital Year Date of birth of patient Month Years of First World War Name of the Monarch **Count backwards from 20 to 1**	
Total Score = 1.6 + motor + proprioception + balance + cognition	

Reference: Kalra and Crome. 1993; www.strokecenter.org

OPS scores range from 1.6 to 6.8 such that higher scores indicate greater deficit (Kalra & Crome 1993; Kalra et al. 1994; Lai et al. 1998). Deficits can be categorized as mild to moderate (scores <3.2), moderate to moderately severe (scores 3.2 – 5.2) and severe or major (scores >5.2) (Kalra & Crome 1993; Lai et al. 1998). In their initial study, Kalra and Crome (1993) reported that patients with scores of less than 3.2 tended to have mild to moderate deficits and were discharged home within 3 weeks of admission whereas patients scoring in excess of 5.2 tended to have severe deficits and require long-term care.

It has been estimated that administration of the OPS required less than 5 minutes (Lai et al. 1998; Studenski et al. 2001). It is simple to use and does not require extensive training to administer. Instructions for administration have been provided (Kalra et al. 1994).

Advantages

OPS scores may assist in the appropriate allocation of stroke unit resources by identifying patients most, and least, likely to benefit from rehabilitation (Kalra & Crome 1993). The OPS can be used to predict a number of functional and patient-centred outcomes post stroke such as community mobility or independence in personal care, medication administration, meal preparation and upper limb recovery 6 months post stroke (Lai et al. 1998; Meldrum et al. 2004). Given that the predictive ability of OPS scores extends beyond discharge from specialized stroke rehabilitation, they may also help to target community based resources and rehabilitation more effectively, based on predicted long-term needs of stroke patients.

Use of OPS scores also permits the identification of a middle-group of patients with moderate deficits (Kalra et al. 1994; Pittock et al. 2003). Prognosis in these patients may be determined more by extrinsic factors, including rehabilitation quality, availability and intensity, than in patients with either mild or severe deficits (Kalra et al. 1994).

Limitations

The OPS score was intended for use with regard to rehabilitation and the appropriate targeting of therapy resources and should not be used for acute prognosis (Kalra et al. 1994). The scale should not be administered until consciousness level and neurological condition have stabilized. Kalra et al. reported that assessment 2 weeks after the stroke event is optimal with regard to predictive ability (Kalra et al. 1994). However, several studies have demonstrated significant predictive ability of OPS scores obtained within 14 days of the stroke event (Lai et al. 1998; Shoemaker et al. 2006; Studenski et al. 2001), although in one study patients assessed earlier than 3 days post stroke were excluded due to unstable neurologic condition (Studenski et al. 2001). Most recently, Pittock et al. (2003) reported that OPS scores obtained at 48 hours following stroke were strongly predictive of length of hospital stay and place of residence at 6 months. OPS scores obtained at 48 hours and at 2 weeks were also predictive of functional ability and/or dependence at 6 months and 2 years following the stroke event. Although the 2-week OPS scores were more strongly correlated with outcomes at 6 months, the difference was minimal. The authors suggest that the benefit derived from this improvement in association is outweighed by the benefit of earlier stratification of patients.

Kalra et al. reported that the predictive values for dependence and discharge destination were not as strong in the middle group of patients (OPS 3 – 5, 2 weeks post stroke) as for patients with mild or severe deficits (Kalra et al. 1994). The authors suggested that this could be due to the greater influence of factors extrinsic to the stroke deficit (intensity and quality of rehabilitation, presence of a competent caregiver, family support, personality and motivation of the patient, availability of community support systems) on rehabilitation outcome in this group (Kalra et al. 1994). However, Wright et al. (2004) reported that neither the NIHSS nor the OPS was very good at predicting discharge disposition for patients with severe stroke for the same reasons as those given by Kalra et al. (1994) above.

While the predictive validity of the OPS has been reported in several studies, there is little or no information available with regard to any other of its measurement properties.

Summary – Orpington Prognostic Scale

Interpretability: Accepted categorizations of the severity of stroke-related deficit have significant predictive value with regard to discharge destination and a variety of functional outcomes.
Acceptability: A simple, objective bedside examination that requires less than 5 minutes to administer. It has not been tested for administration by proxy.
Feasibility: The OPS does not require extensive training or special equipment. It is a simple, brief clinical examination portable to any patient setting.

Table 20.2.16.2 Evaluation Summary Orpington Prognostic Scale

Reliability		**Validity**		**Responsiveness**		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IO)	++	++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.2.17 Stroke Rehabilitation Assessment of Movement (STREAM)

The Stroke Rehabilitation Assessment of Movement (STREAM) is an assessment tool that was developed to provide a quick and simple means to evaluate motor functioning post stroke (Daley et al. 1999). It was originally designed to fit within routine, clinical assessment conducted by physiotherapists (Daley et al. 1999). Initial test versions were subject to expert review by two panels comprising 20 physiotherapists, which resulted in a process of item testing, evaluation and reduction to create the final 30-item version of the scale.

The STREAM contains 30 items divided equally into 3 subscales: 1) voluntary motor ability of the upper extremity, 2) voluntary motor ability of the lower extremity and 3) basic mobility. The test begins with the participant in supine position, progressing to a seated position and ending in an upright, standing position (Ward et al. 2011). Items on the upper and lower extremity subscales are scored on a 3-point ordinal scale ranging from 0 (unable to perform the test movement through any appreciable range including flicker or slight movement) to 2 (able to complete the movement in a manner that is comparable to the unaffected side). Items on the basic mobility subscale (where mobility is defined as the level of independence in the activity) are scored on a 4-point ordinal scale, ranging from 0 (unable to perform the test activity through any appreciable range, i.e. minimal active participation) to 3 (able to complete the activity independently with a grossly normal movement pattern, without the use of an aid (Ahmed et al. 2003; Daley et al. 1999). Total raw scores for the STREAM range from 0-70 (20 for each of the upper and lower extremity subscales and 30 for the mobility subscale, respectively) (Daley et al. 1999). Total and subscale scores may be converted to a percentage score, and taken as an average, to accommodate missing scores on some items (Ahmed et al. 2003; Daley et al. 1999).

The test takes approximately 15 minutes to administer (time range from 0-30 minutes) (Ahmed et al. 2003). The STREAM assessment requires no equipment other than a pencil/paper. No previous training is required for test administration (Rehab Measures 2010). The STREAM itself is purposefully designed to be fast and simple to administer (Wang et al. 2002).

Advantages

The STREAM provides an assessment of voluntary movement that includes the testing of amplitude, gross quality and independence in mobility, while maintaining simplicity and objectivity (Daley et al. 1999). The simple scoring systems and standardized testing instructions as well as the progression of assessment items from supine to standing and from low to high level in terms of ability contribute to the reliability, and rapidity, of assessment (Daley et al. 1999). A 15-item, simplified version of the STREAM or S-STREAM has also been developed based on the results of a Rasch analysis of the original scale (Hsueh et al. 2006).

The STREAM can be used in the assessment of individuals who have experienced severe stroke. Ahmed et al. reported relatively low completion rates on other commonly used functional measures (21% on the Barthel Index and 26% on the TUG), whereas all participants could complete assessment with the STREAM (Ahmed et al. 2003). Assessment can be completed within the first few days of the stroke event to provide information used in the prediction of discharge destination, length of stay or functional potential at 3 months post stroke (Ahmed et al. 2003; Ward et al. 2011). In addition minimal clinically important difference values of 2.2, 1.9, and 4.8 points for the upper extremity, lower extremity and mobility subscales of the STREAM, respectively, have been reported, based upon ratings of perceived change in function made by a group of 81 individuals with stroke (Hsieh et al. 2008).

Limitations

The STREAM may offer a restricted range of assessment. At admission to rehabilitation large floor effects have been reported, as have large ceiling effects for assessments at the time of discharge (Hsueh et al. 2008). However, the shorter, Rasch-modelled, S-STREAM, may provide an improved range of assessment. In the same study, Hsueh et al. (2008) also reported that the S-STREAM demonstrated no significant floor or ceiling effects at either admission to or discharge from rehabilitation. In addition, S-STREAM appeared more sensitive to change over time (S-STREAM SRM=1.19, 1.14 and 1.26 vs. STREAM SRM=0.78, 0.84 and 0.95 for the upper extremity, lower extremity and mobility subscales, respectively) than the 30-item STREAM.

Ahmed et al. (2003) noted that scores may be affected by both age and the presence of cognitive impairment.

Summary - STREAM

Interpretability: Scoring system is simple, based on ability versus inability to perform simple voluntary movement and basic mobility items. Scores may be influenced by age and cognition. MCID values have been reported for each of the STREAM subscales.
Acceptability: Test administration is short and can be completed by individuals with severe stroke.
Feasibility: The assessment is brief, and simple to administer. No training or specialized equipment is required.

Table 20.2.17.1 Evaluation Summary STREAM

Reliability		Validity		Responsiveness		
Rigor	**Results**	**Rigor**	**Results**	**Rigor**	**Results**	**Floor/ceiling**
++	+++(TR) +++(IC)	++	+++	+	+++	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;*

20.3 Activity/Disability Outcome Measures

This section corresponds to the second level or category of the ICF classification system. While keeping in mind that the fit of a given instrument within a single ICF category is rarely perfect, measures appearing in this section focus primarily on the identification or assessment of limitations in activity.

20.3.1 Action Research Arm Test (ARAT)

The Action Research Arm Test (ARAT) is an observer-rated, performance-based assessment of upper extremity function and dexterity (Hsueh et al. 2002). The test was developed by Lyle using a sample of 20 patients with hemiplegia, secondary to cortical injury arising from stroke and forms of brain injury and was derived from the Upper Extremity Function test (UEFT) (Carroll 1965; Lyle 1981). The UEFT is a much longer, more complex assessment containing redundant items and requiring approximately one hour to administer (Lyle 1981).

While the UEFT has 33 items grouped into 6 categories, the ARAT has only 19 items, which are grouped into 4 subsets. Subsets include: grasp (6 items), grip (4 items), pinch (6 items) and gross movement (3 items). All items are rated on a 4-point ordinal scale ranging from 0 to 3 where 0 represents no movement possible and 3 represents normal performance of the task.

Within each subset, the first item is the most difficult and the second is the easiest. The remainder of the items are ordered by ascending difficulty. Successful completion of a particular task or item implies that subsequent, easier tasks can also be successfully completed. For each subset, the most difficult task is attempted first, and, if successful (i.e. 3 points awarded), full points for that subsection are awarded. If the item is not completed successfully (i.e. <3 points were awarded), the next (easiest) item is attempted. If the patient receives a score of 0 on the easiest item, no points are awarded for that subsection and no further items are attempted. If the patient receives a score greater than 0, all remaining items within the subset are assessed.

Summation of scores yields a total score between 0 and 57. Performance time is not recorded. If all 19 items are completed the test takes a maximum of 20 minutes to complete, although it was completed within 8 minutes in at least one study (De Weerdt 1985). With the exception of the testing table(Lyle 1981), items required for the test can be obtained easily and include a chair, woodblocks, a cricket ball, a sharpening stone, two different sizes of alloy tubes, a washer and a bolt, two glasses, a marble and a 6 mm ball bearing.

Advantages

The ARAT is a relatively short and simple measure of upper limb function that provides assessment of a variety of tasks over a range of complexity. The test covers most aspects of arm function, including proximal control and dexterity. Given the emphasis placed on functional task items, ARAT scores may be predictive of improvement in ADL or IADL outcomes (Li et al. 2012). No formal training is required to administer the test. Since the scoring of the ARAT is based on a hierarchical Guttman scale, the testing can be completely quickly on higher functioning patients. Evaluations have demonstrated excellent test retest and interrater reliability. Standardized guidelines for administration are available (Platz et al. 2005; Yozbatiran et al. 2008).

Limitations

In more impaired individuals, testing time can extend to 20 minutes or more. Test administration requires a fairly long list of materials. Significant floor and ceiling effects have been identified. In patients with severe impairments or near normal function, the scale may not be sensitive enough to detect changes in performance (van der Lee et al. 2002). It has been suggested that the ARAT may be most appropriate for use in the assessment of patients with moderate to severe hemiparesis since the test allocates points to be awarded for movement of the arm and hand even though the patient may not be able to pick up items required within the testing environment (Chanubol et al. 2012).

Analysis of scale construction (Mokken analysis) has demonstrated that the 19 items appearing on the ARAT are evaluating a single construct and, therefore, the ARAT is a unidimensional measure (Koh et al. 2006; Nijland et al. 2010; van der Lee et al. 2002). Given these findings, item scores should be summed to provide a single overall score representing upper extremity function, rather than using item responses in 4 subscales (Koh et al. 2006). In addition, as the measure did not fit Rasch model expections, Koh et al. (2006) suggested that raw ARAT scores are not suitable for transformation to interval data and should be treated as ordinal level data only. In contrast to previous work, Chen et al. (2012) item fit anlasyses (infit statistics) suggest that two items on the gross subscale, “place hand behind head” and “place hand on top of hand”, revealed a poor fit. These items might reflect a different aspect of upper extremity (UE) motor function as they involve upward flexion with a larger degree of forearm flexion and a smaller degree of forearm pronation compared to the other subscale items.

A disordering of the ARAT threshold measure has been found (Chen et al. 2012), indicating that the original 4-point scale does not differentiate stroke patients (with mild-to-moderate UE motor dysfunction) effectively with redunacy in the 0- and 1- point scale categories. Chen et al. (2012) recommend using a revised rating category (3-point ordinal scale), that combines scoring categories 1 and 0. The revised rating categories are labeled as 1, can perform no part of the test or partially perform the test within 60 seconds; 2, completed test but takes an abnormally long time (5-60s) or has great difficulty; and 3, performs test normally within 5 seconds. The revised 3-point scale supports the decision rule for ARAT administration where within each subscale once a patient scores 3 in the first item the remaining items are skipped and scored 3, an if a patient scores 0 on the second item the remaining items are skipped and scored a 0.

Summary – Action Research Arm Test

Interpretability: As a Guttman scale, level of performance is easily understood and compared.
Acceptability: Not appropriate for use with proxy; minimal burden for patients.
Feasibility: An extensive collection of items and a specialized table are required. Testing must be carried out in a formal setting. There is no cost to the test but the original guidelines for administration contain limited detail. Standardized guidelines for administration have been
proposed by Yozbatiran et al. (2008).

Table 20.3.1.1 ARAT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++(TR) +++(IO)	++	+++	++	+++	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.2 Barthel Index (BI)

The Barthel Index of Activities of Daily Living (BI) has been in use since 1955 (Mahoney 1965). It was originally intended as a simple index of independence by which to quantify the ability of a patient with a neuromuscular or musculoskeletal disorder to care for him/herself (regardless of particular diagnostic designations). It is, perhaps, the most widely used measure of functional disability.

The BI is very simple, consisting of 10 common activities of daily living (ADL) activities, administered through direct observation. These are assessed for independence/dependence and scored via an arbitrary weighting system (originally applied to reflect nursing care and social acceptability). Eight of the ten items represent activities related to personal care; the remaining 2 are related to mobility. The index yields a total score out of 100 – the higher the score, the greater the degree of functional independence (McDowell & Newell 1996). The BI can take as little as 2 – 5 minutes to complete by self-report and up to 20 minutes by direct observation (Finch et al. 2002). It does not require training to administer, however studies have reported equovical reliability for the BI when administered by trained versus untrained personel. One study has shown the BI to be equally reliable when administered by skilled and unskilled individuals (Collin et al. 1988), where as a systematic review by Duffy et al. (2013) demonstrated a strong trend for improved reliability for raters who have been trained in the BIs application and administration

Advantages

The clearest advantage of the BI is its simplicity and ease of administration – in all of its forms. Its reliance on information collected during functional examination enhances its convenience and cost effectiveness in longitudinal assessment. Its established, widespread use provides a high degree of familiarity and interpretability. It has been used across a variety of settings without a significant decrease in reliability or validity.

Minimal clinically important differences (MCID) have been identified for the BI when used within a stroke population (Hsieh et al. 2007). Hsieh et al. (2007) reported that that a mean BI change score of 1.85 corresponded to patient ratings of minimally important change (a little better to somewhat better) while, using an alternative method based upon the standard error of measurement (SEM), the calculated MCID was 1.45. Use of 1.85 points as the MCID includes patients' subjective perception of change and exceeds the measurement error of the instrument (Hsieh et al. 2007). It should be noted that, as no individual included in the Hsieh et al. study reported deterioration over time, this estimate of MCID is applicable to improvement only (Hsieh et al. 2007).

Limitations

Perhaps the most common criticism of the BI is its relative insensitivity and lack of comprehensiveness particularly as is reflected in large reported ceiling and floor effects. In contrast to additional ADL and instrumental activities of daily living (IADL) measures (i.e. the Frenchay Activites Inedex and Nottingham Extended ADL Scale), the BI is found to have a significantly higher percentage of stroke patients that score a maximum value (100) (Sarker et al. 2012). Duncan et al. (1997) demonstrated that, among patients recovering from mild stroke or TIA who scored 100 on the BI, there continue to be deficits in health status suggesting that the BI is not sensitive to change among the least impaired stroke survivors. However, Wade and Collin (1988) point out that while the BI may not be able to detect change within an individual who is independent, it is able to detect when a patient requires assistance. This distinction may, the authors point out, have more significance to clinical practice than to research.

In addition to the criticisms regarding lack of responsiveness and significant ceiling/floor effects, problems have been noted with regard to dichotomization typical to use with the BI. Because it is frequently used as a dichotomous index, it attracts further criticism for its imprecision (McDowell & Newell 1996). The dichotomization of scales reduces outcome information and may limit a scale's ability to detect a significant shift in disability (Duncan et al. 2000).

Although Granger (1977) proposed a 60/61 split as the threshold of dependence/independence, this has not been adopted as a standardized cut-off and, indeed, there seems little agreement regarding classifications derived from the BI score. Quinn et al. (2011) identified the most common cut-off to define "good outcome" as >95 points. In addition, the proliferation of scale modifications and alternate scoring methodologies has not served to clarify the confusion that surrounds the definition of independence or "good outcome". There may be as many as 4 scales, described as the BI but including modifications such as deleting or addition of items, changes to item definitions, re-ordering items and scoring variations in current use (Quinn et al. 2011). The modified Barthel developed by Collin and Wade (1988) is perhaps the most common of these. This version maintains content that appears to be equivalent to the original scale, but provides a revision to scoring resulting in a total scale score of 0-20. In the case of the 20-point version, ≥19/20 has been used to signify independence (Kwakkel et al. 2011; Stroke Unit Trialist's Collaboration 2007).

Kwon et al. (2004) recently attempted to use the Modified Rankin Scale as a reference to translate BI scores into level of disability and determined that BI scores could be categorized in terms of 4 MRS levels (MRS (0,1,2), MRS 3, MRS 4 and MRS5). Uyttenboogaart et al. (2005) examined cut-off scores for the BI corresponding to categories of disability represented by the Modified Rankin Scale. The authors reported that a cut-off BI score of 95 corresponded to MRS 1 with sensitivity of 85.6% and specificity of 91.7%. MRS2 and MRS3 similarly corresponded to cut-off BI scores of 90 (sensitivity = 90.7%, sensitivity 88.1%) and 75 (sensitivity = 95.7%, specificity 88.5%). While the authors recommend that these values, along with the corresponding MRS scores, be used as the basis for dichotomizing outcome as favourable versus unfavourable, there is, as yet, no apparent consensus for categorization of BI scores, whether in terms of dichotomization for functional dependence or translation to level of disability, and, therefore, comparison of outcomes across trials is difficult and does not favour any sort of meta-analytic approach (Duncan et al. 2000; Roberts & Counsell 1998; Sulter et al. 1999).

Summary – Barthel Index

Interpretability: The degree of familiarity of the BI contributes to its interpretability. However, there is a lack of agreement regarding threshold for independence/dependence and several different scoring systems are used making comparisons across groups/studies more difficult. There are no norms available for comparison.

Acceptability: The BI has been evaluated for both self-report and use with proxy respondents in addition to direct observation. Both self-report and interview formats generally take less time to complete than the original (direct observation) and may serve to reduce patient burden.

Feasibility: The BI is simple to administer and requires no training. It has been developed in many forms that can be administered in many situations and seems suited for longitudinal assessment.

Table 20.3.2.1 BI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IO) +++ (IC)	+++	+++	+++	++	Varied

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.3 Berg Balance Scale (BBS)

The Berg Balance Scale (BBS) provides a quantitative assessment of balance in older adults (Berg 1989). It was intended for use in monitoring the clinical status of patients or effectiveness of treatment interventions over time (Berg et al. 1995).

The scale consists of 14 items requiring subjects to maintain positions or complete movement tasks of varying levels of difficulty. All items are common to everyday life. Administration of the scale is completed via direct observation of task completion. It takes 10 – 15 minutes and requires a ruler, a stopwatch, chair, step or stool, room to turn 360°(Berg et al. 1995; Juneja et al. 1998). Items receive a score of 0-4 based on ability to meet the specific time and distance requirements of the test. A score of zero represents an inability to complete the item and a score of 4 represents the ability to complete the task independently. It is generally accepted that scores of less than 45 are indicative of balance impairment (Berg et al. 1992; Zwick et al. 2000).

Advantages

The BBS measures a number of different aspects of balance, both static and dynamic, and does so with relatively little equipment or space required (Nakamura 1998; Whitney et al. 1998; Zwick et al. 2000). No specialized training is required to be able to administer the BBS (Nakamura 1998). The high levels of reliability reported by Berg et al. (1995) were achieved when the individuals administering the test had no specific training in the administration of the scale. Based on the examination of absolute reliability, values for minimum detectable change of 6.68 (Liaw et al. 2008) to 6.9 (Stevenson 2001) based on a 95% confidence interval have been reported.

Wee et al. (1999) suggested that the BBS may be particularly well suited for use in acute stroke rehabilitation, as the majority patients do not obtain maximum scores on admission to rehabilitation.

The BBS takes somewhat longer than other balance measures to administer (Chou et al. 2006; Whitney et al. 1998) and may suffer from some item redundancy given its extraordinarily high levels of internal consistency. Chou et al. (2006) developed a 7-item version with a revised 3-level response format (Wang et al. 2004). Results obtained via this new short form agree significantly with those obtained using the original BBS (ICC = 0.99) (Chou et al. 2006). In addition, the new version appears to be both valid and, with the exception of a significant floor effect (>40%), responsive. As Chou et al. (2006) point out, the floor effect may, in part, be attributed to the removal of the simplest item on the scale (unsupported sitting).

Limitations

The BBS may not be suitable for the evaluation of active, elderly persons, as the items included are not sufficiently challenging for this group (Berg 1989; Nakamura 1998; Zwick et al. 2000). The BBS may suffer from decreased sensitivity in early stages post stroke among severely affected patients as the scale includes only one item relating to balance in the sitting position (Mao et al. 2002).

No common interpretation exists for BBS scores, their relationship to mobility status, and the use of mobility aides (Wee et al. 2003). The rating scales associated with each item, while numerically identical, have different operational definitions for each number or score; a score of 2, for example, is defined differently and has a different associated level of difficulty from item to item (Kornetti et al. 2004). There is also no common score associated with successful item completion (Kornetti et al. 2004). Use of an overall score that adds together ratings with different meanings having no common reference point may not be appropriate as interpretation is difficult and very little functional information is provided about the individual patient (Kornetti et al. 2004).

A recent item-fit analysis (Rasch analysis) identified two BBS items as misfits (item-13, stand with one foot in front; item-14, stand on one leg) (Straube et al. 2013). The authors note that the self-selection nature of these BBS items (participant being able to choose the lower extremity, impaired or unimpaired, to perform each task) may allow patients with low balance ability to score high on these items when the unimpaired lower extremity is test. Conversely, patients with high balance ability score low on these items when the impaired lower extremity is tested. More standardized instruction regarding lower extremity (impaired vs. unimpaired) should be used to perform each item task in order to help improve item fit. Kornetti et al. (2004) performed a Rasch analysis of the BBS and revealed that some item ratings were not used at all or were underutilized, and others were unable to distinguish between individuals with different levels of ability. Collapsing rating scales to eliminate infrequently endorsed categories and creating a common pass/fail point for each item resulted in changes to the ordering of item difficulty, reduced tendencies for ceiling effects and an improved functional definition of the 45/56 cut-off point (Kornetti et al. 2004). An additional study utilizing Rasch anaslysis also indicated the need to modify the BBS (La Porta et al. 2012). La Porta et al. (2012) suggest a modified model score, as 11-items showed disordering thresholds, in addition to the deletion of two scale items (items 2 and 3 assessing sitting and standing balance). Following such modifications, an analysis of differential item factoring (DIF) showed invariance for the patient factors (sex, age, days since lesion, and etiology).

While earlier studies found no relationship between BBS scores and age, Steffen et al. (2002) reported a trend toward declining performance with increasing age for both men and women. The authors provided age and gender-related performance data based on a small sample of community-dwelling, independent elderly people and recommended that further data be gathered from larger samples in order to create age and gender stratified norms for reference purposes.

Summary – Berg Balance Scale

Interpretability: While the reliability and validity of the scale are excellent, there are no common standards for the interpretation of BBS scores though there is an accepted cutoff point for the presence of balance impairment.

Acceptability: This direct observation test would not be suited for severely affected patients as it assesses only one item relative to balance while sitting. Active individuals would find it too simple. The scale is not suited for use by proxy.

Feasibility: The BBS requires no specialized training to administer and relatively little equipment or space.

Table 20.3.3.1 BBS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++(IO) +++ (IC)	+++	+++	+++	+++	Varied

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.4 Box and Block Test (BBT)

The Box and Block Test (BBT) is a performance-based measure of gross manual dexterity originally developed by A. Jean Ayres and Patricia Holser Buehler for use in the assessment of adults with cerebral palsy (Mathiowetz et al. 1985). In 1957, the test was revised and copyrighted in its current format (Cromwell 1976; Mathiowetz et al. 1985).

Test respondents are seated at a table, facing a rectangular box that is divided into two square compartments of equal dimension by means of a partition. One hundred and fifty, 2.5 cm, coloured, wooden cubes or blocks are placed in one compartment or the other. The respondent is instructed to move as many blocks as possible, one at a time, from one compartment to the other for a period of 60 seconds. Standardized dimensions for the test materials and procedures for test administration and scoring have been provided by Mathiowetz et al. (1985).

To administer the test, the examiner is seated opposite the respondent in order to observe test performance. The BBT is scored by counting the number of blocks carried over the partition from one compartment to the other during the one-minute trial period. The patient's hand must cross over the partition in order for a point to be given, and blocks that drop or bounce out of the second compartment onto the floor is still rewarded with a point. Multiple blocks carried over at the same time, count as a single point. Higher scores on the test indicate better gross manual dexterity. Norms have been established for various populations including healthy elderly individuals (Desrosiers et al. 1994), healthy adults (Mathiowetz et al. 1985), adults with neuromuscular involvement (Cromwell 1976) and healthy 7, 8 and 9 year old children (Mathiowetz et al. 1985).

Administration takes approximately 5 minutes. The BBT is easy to administer and does not require highly specialized training. The test is readily available for purchase and can be obtained from a variety of online sources.

Advantages

The BBT is a popular measure of gross manual dexterity that is both quick and simple to administer. The simplicity of the performance task and the seated administration position may make the test more accessible to a wider range of individuals. Standardized administration and scoring procedures are available (Mathiowetz et al. 1985). Moreover, established, age and gender-stratified norms are available for a variety of populations, thereby increasing the interpretability of test results.

BBT scores have been found to be predictive of physical health as measured by the Medical Outcomes Study 36-Item Short form Questionnaire (SF-36) (Higgins et al. 2005; McEwan 1995). McEwan (1995) demonstrated that an increase of 7 blocks on the BBT was associated with a change of 2 units in the Physical Component Summary Score of the SF-36, an amount of change considered to be clinically relevant. Thus, the BBT may have utility as a prognostic indicator of physical health.

Figures for clinically significant change in BBT performance have been reported in stroke populations, with improvements of four to five blocks (Carey et al. 2002) and eight blocks(Kimberley & Lojovich 2004) considered clinically important. However, the aforementioned studies did not evaluate minimal detectable differences in scores and different designs were used (Svensson & Hager-Ross 2006).

Limitations

As an assessment of upper extremity function, the BBT does not provide assessment of a range or variety of tasks. As such, use of the BBT may be associated with substantial floor effects in some patient groups, given that patients must have sufficient arm movement, strength and grip function in order to transport blocks (Chanubol et al. 2012). No points are awarded for partial arm movement or movement of the arm against gravity. It has been recommended that the BBT, therefore, might be most appropriate for individuals with mild to moderate hemiparesis and moderate weakness (Chanubol et al. 2012).

The BBT is noisy to administer and could be distracting to other patients in a busy clinic (Mathiowetz et al. 1985).

Summary – Box and Block Test

Interpretability: Age-stratified norms have been established on various populations including healthy elderly individuals.
Acceptability: The test is brief at approximately 5 minutes, including instruction and pre-test trials, and represents little patient burden.
Feasibility: The BBT is easy to administer and does not require highly specialized training. Little equipment is required. There is a cost associated with purchase of the test.

Table 20.3.4.1 BBT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++ (IO)	++	+++	++	++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.5 Chedoke-McMaster Stroke Assessment Scale (CMSA)

The Chedoke-McMaster Stroke Assessment Scale (CMSA) is a 2-part assessment consisting of a physical impairment inventory and a disability inventory. The impairment inventory is intended to classify patients according to stage of motor recovery while the disability inventory assesses change in physical function (Gowland et al. 1993).

The scale's impairment inventory has 6 dimensions; shoulder pain, postural control, arm movements, hand movements, leg movements, and foot movements. Each dimension (with the exception of 'shoulder pain' whose rating scale is unique) is rated on a 7-point scale corresponding to Brunnstrom's 7 stages of motor recovery (where 1=flaccid paralysis & 7= normal). The maximum total score for physical impairment is 42. The disability inventory consists of a gross motor index (10 items) and a walking index (5 items). With the exception of a 2-minute walking test, items are scored according to the same 7-point scale used in the Functional Independence Test (FIM) where 1 represents total assistance and 7 represents total independence. The walking test item receives a score of either 0 or 2. Overall, the disability inventory has a maximum score of 100: 70 from the gross motor index, 30 from the walking index. Assessments are completed by direct observations.

Instructions on administration, scoring and interpretation are required to perform the CMSA (Gowland 1995). In addition to the manual, administration of the test requires a mat or bed and a chair. It takes approximately 1 hour to complete (Cole & Basmajian 1994; Poole & Whitney 2001).

Advantages

The Chedoke-McMaster Stroke Assessment was designed for use in conjunction with the FIM and uses the same rating method for its disability inventory. This may provide improved interpretability by using a consistent concept of independence, while improving sensitivity to small physical changes (Gowland et al. 1993). In a review of motor function assessments, Poole and Whitney concluded that, by comparison, the CMSA is comprehensive and has been well studied for reliability and validity (Poole & Whitney 2001).

Limitations

One must order the manual in order to administer the CMSA. The relative complexity and length of administration may make the CMSA less useful for application in a clinical practice setting (Poole & Whitney 2001).

The upper extremity tasks included on the test are not functional and, except for items related to transfer and gait, the CMSA is primarily a measure of motor impairment. It is recommended that measures of motor impairment be accompanied by a measure of functional disability such as the BI or FIM (Poole & Whitney 2001). The analysis of Valach et al. (2003) would seem to support this recommendation. Regression analysis revealed that although as few as 3 items of the CMSA disability index could be used to predict BI scores, there was still a large portion of unexplained variance. In addition, the BI-derived factors of eating/drinking and bowel/bladder incontinence were shown to add information not covered by the Chedoke-McMaster assessment (Valach et al. 2003).

Summary – Chedoke McMaster Stroke Assessment

Interpretability: The use of Brunnstrom staging and FIM scoring increase interpretability and facilitate comparisons across groups of stoke patients. However, the assessment might best be regarded as a measure of motor impairment (Poole & Whitney 2001; Valach et al. 2003).
Acceptability: The CMSA is a long test. It is not suited to proxy use.
Feasibility: Requires little equipment but is fairly lengthy and complex to administer. It has been tested for use in longitudinal assessment.

Table 20.3.5.1 CMSA Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IO) +++ (IC)	+	+++	+	+++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.6 Chedoke Arm and Hand Activity Inventory (CAHAI)

The Chedoke Arm and Hand Activity Inventory (CAHAI) is a relatively new measure for assessing functional upper-limb recovery in stroke survivors. The measure was developed by Barreca et al. (2004) to provide a valid, clinically relevant means of assessment for the recovering paretic limb. The 5 main objectives of the test are: 1) to discriminate between different categories of upper limb dysfunction; 2) to predict anticipated functional recovery in the paretic upper limb; 3) to quantify the amount of change in upper limb function; 4) to determine the importance of that change to stroke survivors; and 5) to serve as a guide to treatment. Moreover, the CAHAI was developed as a complimentary measure to the Chedoke-McMaster Stroke Assessment (CMSA), a well-established stroke measure that classifies arm and hand impairment into 7 stages.

Test items consist of 13 real-life functional tasks intended to reflect: 1) the domains deemed important by survivors of stroke; 2) bilateral activities; 3) non-gender specific items; 4) the full range of normative movements, pinches, and grasps; and 5) the various stages of motor recovery post-stroke. All 13 items are scored using a 7-point quantitative scale. Total scores are obtained by summing the item scores and thus can range from 13 to 91. Higher scores indicate greater ability.

The test takes 25 minutes to administer and requires easily obtained, transportable, and inexpensive materials. Training is recommended for administration (Barreca et al. 2005).

Advantages

A major advantage of the CAHAI is its ecological validity. Working closely with stroke survivors, test items/skills were specifically selected to be meaningful and relevant to a stroke population. Being ecologically valid is important because it ensures that the test highlights tasks that should be given special attention during treatment and thus helps to inform the rehabilitation process.
The CAHAI is a well-constructed test that was designed to be compatible with World Health Organization (WHO) guidelines as well as the CMSA. The WHO disability domain for a client specific model describes specific criteria that are relevant to disability. These include personal care, dressing, feeding, mobility, communication and recreation (Barreca et al. 2004). Items for the CAHAI were purposefully generated to meet these criteria. In terms the CMSA, the compatibility of the CAHAI is advantageous because it means that researchers and clinicians have the option of utilizing the CAHAI as part of a comprehensive assessment package that targets general motor and functional recovery post stroke.

The CAHAI covers a wide range of functions not assessed by other measures of paretic-upper limb dysfunction. These include normative upper-limb movements of manipulation, reach and grasp, non-gender-specific tasks, degree of motor recovery, and bilateral tasks (Barreca et al. 2004). Additionally, the test was designed to be applicable across different settings and may be used in the hospital, at home, or in an outpatient unit.

Psychometrically, the CAHAI has demonstrated strong validity and reliability (Barreca et al. 2005; Barreca et al. 2006a, 2006b). In addition, the CAHAI has demonstrated responsiveness to change over time and a value for minimal detectable change has been reported. Three shortened versions of the CAHAI were created for more efficient data collection. Evaluations of the CAHAI-9, CAHAI-8 and the CAHAI-7 (9, 8 and 7 items, respectively) have demonstrated measurement characteristics comparable to the parent scale while reducing the time required for administration (Barreca et al. 2006a, 2006b). A reliable and valid German translation is available for the CAHAI-7-8-9 (Schuster et al. 2010).

Limitations

While the CAHAI appears to be a promising measure of upper-limb function, there has been relatively little third-party evaluation of the scale’s measurement properties. Further research is required.

Summary – Chedoke Arm and Hand Activity Inventory

Interpretability: The CAHAI is designed to measure recovering upper-limb function in stroke survivors. At this point, no norms are available for scoring.
Acceptability: The test takes a moderate amount of time to administer at 25 minutes. However, during piloting, there were no complaints amongst stroke patients with respect to fatigue (Barreca et al. 2004). As well, three shortened versions of the test have been created for quicker administration.

Feasibility: The test requires easily obtained, transportable, and inexpensive materials. It was designed to be flexible in terms location of administration and may be utilized across different settings (e.g., hospital, home, outpatient unit). Specialized training is recommended for administration (Barreca et al. 2005).

Table 20.3.6.1 – CAHAI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) ++(IC)	+	+++	+	++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;*

20.3.7 Clinical Outcome Variables (COVS)

The Clinical Outcomes Variables scale (COVS) was published as a tool designed to be used by physiotherapists in the assessment of functional mobility status in order to identify treatment goals and initiate treatment protocols (Eng et al. 2002; Hajek et al. 1997; Seaby & Torrance 1989). The 13-items comprising the COVS were selected in such as way as to be representative of outcomes associated with a regular physiotherapy caseload within the general rehabilitation population (Finch et al. 2002; Seaby & Torrance 1989). The concept of environmental barriers and the ability to negotiate within the environment is incorporated into the test items (Seaby & Torrance 1989), which include assessment of transfer abilities to and from bed and from the floor as well as wheelchair skill (Low Choy et al. 2002).

Each item or functional task has its own 7-point rating scale based on the Patient Evaluation Conference System (PECS) (Harvey & Jellinek 1981) with 1 representing the worst possible outcome and 7 the best possible outcome (i.e. the highest amount of function). Items can be considered individually or summed to provide a composite score ranging from 13 – 91. Items can also be summed in various combinations to provide assessments of ambulation (4 items), mobility in bed (2 items), transfers (2 items) and arm function (2 items) (Seaby & Torrance 1989).

The COVS is usually administered by a trained physiotherapist and may be completed as part of a routine physical therapy assessment. A full assessment takes approximately 15 – 45 minutes to complete. One can purchase the test directly from the Institute for Rehabilitation Research and Development at (*www.rehab.on.ca/irrd/covs*). Written training guidelines, a training video, database software and detailed rating guides are also available (Finch et al. 2002).

Advantages

The COVS provides detail in areas of mobility not assessed by global functional assessments such as the FIM (Low Choy et al. 2002; R. 2002). It monitors motor tasks retrained by physiotherapists and includes both the use of assistive devices and the ability to negotiate environmental barriers. Overall, it has demonstrated good reliability as well as strong construct and predictive validity. Examinations of longitudinal validity have demonstrated that the COVS is sensitive to change over time. The COVS was designed to be performed as part of a routine physiotherapy assessment which may offset the potential for increased patient burden associated with its length (Huijbregts 1996).

Limitations

Administration of the COVS requires a fairly lengthy list of equipment (stopwatch, plastic mug, penny & slotted can or pincushion and straight pins, an exercise mat, ramp with a 1 – 12 inch rise, and a 6-inch platform) and a substantial amount of time. There is an ongoing need for further validation of the COVS, which is relatively widely used.

Summary – Clinical Outcome Variables Scale (COVS)

Interpretability: Items are all based on functional mobility tasks. Factor analysis has confirmed (Hajek et al. 1997) that the scale is a unidimensional assessment making interpretation of scores relatively simple. In addition the scale incorporates the concepts of environmental barriers and the use of assistive devices.
Acceptability: The test, while quite lengthy on its own, can be incorporated into a routine physiotherapy assessment, which may reduce the patient burden associated with a long assessment process.
Feasibility: There is additional cost associated with the purchase of the test itself and any supplementary materials required. Physiotherapists should be trained prior to administration and/or scoring in order to achieve the levels of reliability reported. Although the equipment list is long, many of the items (with the exception of those required to simulate outdoor settings) are easily obtainable.

Table 20.3.7.1 COVS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IO) ++ (IC)	++	+++	++	+++	++

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.8 Functional Ambulation Categories (FAC)

The Functional Ambulation Categories (FAC) is a measure developed at Massachusetts General Hospital to rate the ambulation ability of patients undergoing physical therapy (Holden et al. 1984). This 6-point scale assesses ambulation status by determining how much human support the patient requires to walk, regardless of whether or not they use a personal assistive device (Holden et al. 1984). The FAC is an extensively used outcome measure in the rehabilitation setting alongside conditions that have detrimental effects on walking ability including hemiplegia (Hesse et al. 1994; Holden et al. 1986, 1984), multiple sclerosis (Holden et al. 1986, 1984), stroke (Brock et al. 2002; Collen et al. 1990; Cunha et al. 2002; da Cunha et al. 2002; Lord et al. 2004; Simondson et al. 2003; Stevenson 1999) and cerebral palsy (Schindl et al. 2000). Wade (1992) suggests that the best use of the FAC is not for the measurement of actual disability but for measuring progress in active rehabilitation.

To use the FAC, an assessor (usually a physiotherapist) asks the subject various questions and briefly observes their walking ability to provide a rating from 0 to 5 (Collen et al. 1990). If the subject scores 0 they are a non-functional ambulator (cannot walk); a score of 1, 2, or 3 denotes a dependent ambulator who requires assistance from another person in the form of: continuous manual contact (1), continuous or intermittent manual contact (2), or verbal supervision/guarding (3); a score of 4 or 5 describes an independent ambulator who can walk freely on: level surfaces only (4) or any surface (5 = maximum score) (Holden et al. 1984).

The FAC is readily available (Holden et al. 1986, 1984; Wade 1992). There is no equipment required for the administration of this scale and the classification is explained in thorough detail especially if using the description provided by Holden et al. (1986; 1984).

Advantages

The FAC is a simple scale to administer and requires no special training or equipment (Collen et al. 1990). This scale has been shown to be a discriminatory measure among individuals with higher-level mobility function (Lord et al. 2004).

Limitations

The FAC may lack responsiveness, especially if using it to distinguish between groups at lower levels of functioning (Collen et al. 1990; Lord et al. 2004) and large ceiling effects have been reported. However, a study (Mehrholz et al. 2007) has reported moderate to large effect sizes when the FAC was used to evaluate change in ambulation over a period of 6 months. Given that this study included only individuals who were non-ambulatory at baseline, responsiveness could be somewhat over-estimated. Future research is required to determine whether the assessment tool is equally responsive in higher-functioning individuals.

Summary – Functional Ambulation Categories

Interpretability: FAC scores should be interpreted with caution given the reduced responsiveness among individuals with lower levels of function and the large reported ceiling effects associated with its use. A rating on the FAC should be construed as a description of a subject's walking ability only (Collen et al. 1990).

Acceptability: Administration of the FAC is simple, requiring only brief questioning and observation, thereby creating little patient burden.

Feasibility: The FAC is quick and easy to use and the scale can be obtained at no cost. Also, there is no equipment that needs to accompany administration of the scale, which makes it a virtually free assessment tool. No formal training is required to administer the FAC but the user
should be familiar with the scale prior to its use

Table 20.3.8.1 Functional Ambulation Categories Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+ (TR) +++ (IO)	++	+++	+	+++	+

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.9 Functional Independence Measure (FIM)

Developed in 1987, in part as a response to criticism of the Barthel Index, the FIM was intended to address issues of sensitivity and comprehensiveness as well as provide a uniform measurement system for disability for use in the medical remuneration system in the United States (McDowell & Newell 1996). Rather than independence or dependence, the FIM assesses physical and cognitive disability in terms of burden of care – that is, the FIM score is intended to represent the burden of caring for that individual.

The FIM is a composite measure consisting of 18 items assessing 6 areas of function (self-care, sphincter control, mobility, locomotion, communication and social cognition). These fall into 2 basic domains; physical (13 items) and cognitive (5 items). The 13 physical items are based on those found on the Barthel Index, while the cognitive items are intended to assess social interaction, problem-solving and memory. The physical items are collectively referred to as the motor-FIM while the remaining 5 items are referred to as the cognitive-FIM.

Each item is scored on a 7-point Likert scale indicative of the amount of assistance required to perform each item (1=total assistance, 7 = total independence). A simple summed score of 18 – 126 is obtained where 18 represents complete dependence/total assistance and 126 represents complete independence. Subscale scores for the physical and cognitive domains may also be used and may yield more useful information than combining them into a single FIM score (Linacre et al. 1994).

Administration of the FIM requires training and certification. The most common approach to administration is direct observation. The FIM takes approximately 30 minutes to administer and score. The developers of the FIM further recommend that the rating be derived by consensus opinion of a multi-disciplinary team after a period of observation.

Advantages

The Functional Independence Measure has been found to be as effective as such lengthy measures as the Sickness Impact Profile (SIP) in predicting burden of care following stroke and therefore, just as useful in determining the amount of physical assistance a person might need at home following a stroke. To its advantage, the FIM is far less lengthy and represents a smaller burden to the patient than the SIP, which requires the subject to complete the lengthy questionnaire (Granger et al. 1993).

In clinical assessment, the greater number of items and wider choice of responses per item may yield more detailed information on an individual basis than assessments with fewer items and response options (Hobart et al. 2001). Minimal clinically important differences (MCID) have been identified for the FIM when used within a stroke population (Beninato et al. 2006). Based upon ratings of clinical change made by physicians shortly following discharge from stroke rehabilitation, Beninato et al. (2006) determined that 22, 17 and 3 were the change scores for the total FIM, motor FIM and cognitive FIM, respectively, which best separated those patients who had demonstrated clinically important change from those who had not.

Limitations

The reliability of the FIM is dependent upon the individual conducting the assessment. Training and education in administration of the test is a pre-requisite for good levels of inter-rater reliability (Cavanagh et al. 2000). Length of time and amount of training required to arrive at a consensus score, as recommended by the developers of the FIM, may have significant implications for the practical application of the FIM in clinical practice.

The use of a single summed raw score may be misleading as it gives the appearance of a continuous scale. Steps between scores, however, are not equal in terms of level of difficulty and cannot provide more than ordinal level information (Linacre et al. 1994). Kidd et al. (1995) suggested that one use the summed scores as though on an interval level scale while the individual items remain ordinal. However, based on Rasch-based analyses of the FIM scale, the motor-FIM alone appears to be a unidimensional assessment that fulfills model expectations, without deletion of items (Lundgren & Tennant 2011).

In an evaluation of responsiveness, FIM, motor FIM and the BI were all found to have similar effect sizes. The total-FIM was reported to exhibit no ceiling effect -- 0% as compared to the BI's 7% (van der Putten et al. 1999). This would suggest that the FIM might have no real advantage in terms of responsiveness to change despite having more items and a more precise scoring range for each item.

Identification of MCID for the FIM may increase the interpretability of FIM scores and FIM change scores; however, it should be noted that the external criterion around which these figures were developed were retrospective physician ratings of change. Patient, caregiver or family assessments were not included in the ratings of important change. In addition, retrospective ratings could be subject to recall bias. The authors also demonstrated that the MCID was influenced by the FIM scores at admission such that patients with lower admission FIM required greater change scores in order to demonstrated significant change and identification of patients with clinically important change became more difficult to identify accurately as FIM admission scores increased.

Summary – Functional Independence Measure

Interpretability: The FIM has been well studied for its validity and reliability. It is widely used and has one scoring system increasing the opportunity for comparison. It is important to remember, when interpreting FIM scores, that it is an ordinal not continuous level scale.
Acceptability: Modes of administration include interview. The FIM has also been studied for use by proxy respondents.
Feasibility: Training and education of persons to administer the FIM may represent significant cost. Use of interview formats may make the FIM more feasible for longitudinal assessment.

Table 20.3.9.1 FIM Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IO) +++ (IC)	+++	++	+++	++	++

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.9.1 Barthel Index vs. the Functional Independence Measure

The Functional Independence Measure (FIM) was developed, in part, to create a means of assessment that would be less restrictive and more responsive to clinically significant change than the Barthel Index. Therefore, direct comparisons of the two have arisen on a number of occasions.

Both scales have undergone extensive scrutiny in terms of reliability and validity. It is generally accepted that both are strongly reliable and valid measures of functional disability in stroke populations (see descriptions of the individual measures). Hobart et al. (2001) suggest that, in terms of reliability, there appears to be no particular advantage to choosing one scale over the other. Similarly, they find that the BI and the motor-FIM (the FIM's 13 physical subscale items) have comparable convergent and discriminant construct validity. Overall, they appear to be psychometrically similar measures of motor disability (Gosman-Hedstrom 2000; Mao et al. 2002).

Kidd et al. (1995) suggest that the inclusion of items related to communication and cognition as well as the ranking of 7 levels of severity for each item make the FIM more sensitive and inclusive. However, the contribution of the cognitive subscale to the scale as a whole is questionable as it has been shown to have less reliability and responsiveness than either the motor FIM or the total FIM (Ottenbacher et al. 1996; van der Putten et al. 1999). Gosman-Hedstrom and Svensson (2000) suggest that although the FIM is more inclusive than the BI, it does not appear to be more discriminative of change within the individual in a clinical setting when assessed at the level of the scale items.

Responsiveness, or the ability of an instrument to detect clinically significant change over time, is identified as an important criterion to assess in the selection of an outcome measure. The BI has often been criticized for the limited range of disability within which it is able to detect change as evidenced by significant ceiling effects. In studies focusing on the responsiveness of the 2 scales, little to no difference is found in comparisons of the BI, the motor-FIM and the total FIM when used within a population of stroke patients (Hobart & Thompson 2001; Hsueh et al. 2002; van der Putten et al. 1999; Wallace et al. 2002). In a study of MS and stroke patients (that did not include any severely disabled individuals), van der Putten et al. (1999) reported a 7% ceiling effect for the BI, while the total FIM showed no ceiling effect at all (1% for motor-FIM). Hsueh et al. (2002)reported a substantially larger floor effect for admission BI scores than for admission motor FIM scores (18.2% vs 5.8%) in a similar diagnostic population, which did include more severely disabled patients.

In spite of this perceived limitation to the spectrum of detectable change with the BI, both studies (Hsueh et al. 2002; van der Putten et al. 1999) reported significant and comparable change scores for both outcome measures. Wallace et al. (2002) found that the BI & motor FIM exhibited similar responsiveness to change in a population comprised of individuals recovering from stroke. As Wallace et al. (2002) point out, their study – like the others cited here – focus on the responsiveness of the measures to improvement – that is, to unidirectional change only. The ability of the measures to assess decline as well as improvement is not addressed.

Given the demonstrated similarity between these 2 measures, choosing which to use will be dictated by the purpose for which the instrument is to be used and may focus on issues of appropriateness or practicality rather than psychometric properties.

20.3.9.2 CIHI - National Rehabilitation Reporting System

The Canadian Institute for Health Information launched a project in 1999 in order to develop national indicators and outcome reports for adult inpatient rehabilitation services. The purpose in creating the reporting system was to collect & analyze data from adult rehabilitation facilities, provide support for multiple levels of managerial decision-making, facilitate comparisons between regions and support related research and analysis.

The National Rehabilitation Reporting System data elements include the Functional Independence Measure (FIM) as well as 12 CIHI items developed to contribute to the cognitive domain of the FIM. The CIHI pilot project reports the data set as having strong reliability and validity as well as being sensitive to change in functional status. The database of the NRS contains data collected at the time of admission and discharge from participating adult, inpatient, rehabilitation facilities from across Canada. Currently, the MOHLTC mandates the participation of all facilities having designated adult, inpatient rehabilitation beds.

Resource: Canadian Institute for Health Information. Online at: www.cihi.ca.

20.3.10 Frenchay Activities Index (FAI)

The Frenchay Activities Index (FAI) is a measure of instrumental activities of daily living (IADL) for use with patients recovering from stroke. The Index provides an assessment of a broad range of activities associated with everyday life. The items included on the FAI move beyond the scope of ADL scales, which tend to focus on issues related to self-care and mobility (Holbrook & Skilbeck 1983). It was intended to give a moreobjective measure of actual activities undertaken in the subject's recent past (Wade et al. 1985).

The FAI contains 15 items or activities that can be separated into 3 factors; domestic chores, leisure/work and outdoor activities. The frequency with which each item or activity is undertaken over the past 3 or 6 months (depending on the nature of the activity) is assigned a score of 1 – 4 where a score of 1 is indicative of the lowest level of activity. The scale provides a summed score from 15 – 60. A modified 0-3 scoring system introduced by Wade et al. (1985) yields a score of 0 – 45. More recently, Lin et al. (2012) have also suggested a modified 0-3 scoring system to reduce the observed redundancy (disordered thresholds) found with the original 4-point scale. Administered in an interview format (with or without the patient's family), the FAI takes approximately 5 minutes to complete (Segal & Schall 1994; Wade et al. 1985).

The FAI was developed in the 1980's. There has be criticism that the FAI items need to be modified inorder to better represent IADL of the 21st century (Wendel et al. 2013). A modified (extended) Swedish version of the FAI has been created (Wendel et al. 2013) to better represent current out-of-home activities and modes of transportation. This extended version of the FAI also includes the addition of three new response scales for each FAI item assessing the frequency changes, self-reported cause for change, and satisfaction with activity performance, and improves the descriptive and evaluative information collected using the FAI. High inter-rateer agreement has been reported for the extended FAI but additional testing is required particuallry in other contexts (Wendel et al. 2013).

Advantages

The brevity and simplicity of the FAI make it easy to use in a clinical setting (Wade 1992). The FAI seems to be suitable for use with proxy respondents so is inclusive of cognitively impaired stroke survivors. The scale is based on behaviour. Its emphasis on frequency rather than quality of activity may reduce elements of subjectivity, which undermine the reliability of proxy assessment (Segal & Schall 1994).

It has been suggested that domestic, lifestyle, leisure and social activities should be included in assessments of the consequences of stroke (Sveen et al. 1999). Pedersen et al. (1997) demonstrated that the FAI provides different information about ADL function than that obtained on the BI and may represent the next steps along the ADL continuum in terms of item difficulty. A more comprehensive ADL assessment may be obtained by using both assessment tools.

Limitations

In chronic stroke patients the smallest real difference (SRD; the smallest change that indicates real improvement or deterioration for an individual), appears to be quite large (a 6.7 change score) (Lu et al. 2012). It is cautioned that individuals using the FAI keep this SRD value in mind when detecting real change in individual patients (Lu et al. 2012).

The original FAI is reported to be multidimensional, consisting of three factors: domestic chores, leisure/work, and outdoor activities. Most research indicates the FAI to be multidimensional but construct validation studies have produced varying results regarding factor structure (ranging from 2 to 4 factors). Recent dimensionality of the FAI suggests that the items can be divided into two factors (domestic chores and work/leisure) for stroke patients with mild to moderate upper extremity impairment (Lin et al. 2012).

The original authors warned that gender may have some influence on FAI scores; they recommended male and female scores be considered separately (Holbrook & Skilbeck 1983). Sveen et al. (1999)

reported that men had significantly higher scores in outdoor activities while there was a trend toward women having higher domestic activity scores, perhaps based on conventional, gender-based activity patterns. Similarly, Han et al. (2006) in a study of Japanese elderly, demonstrated lower performance by males on items corresponding to "domestic chores" and greater performance on "work and leisure" items. Wade et al. (1985) did not find the same gender bias, but did note different patterns of activity and prevalence of male versus female activity on some items. These patterns changed following stroke. Within the overall score, however, there may be a balance of gender dominance (Appelros 2007; Wade et al. 1985). Appelros (2007) also reported no difference between male and female respondents for total FAI scores, although, there were significant between-gender differences noted on individual items similar to those noted previously.

Other factors in addition to gender may influence FAI scores. Age may significantly impact FAI scores, such that younger age is associated with better scores (Appelros 2007; Han et al. 2006). Appelros (2007) reported that, on regression analysis, age was significantly associated with FAI scores one year post stroke such that each year increase in age was associated with a decrease on the FAI of 0.57 points. Wu et al. (2011) suggested that the activities represented on the FAI are of limited scope and are not necessarily of importance to the respondent. Items such as "gainful work" or "telephone use" may be less relevant to the older, often retired, individuals who have experienced stroke. Significant differential item functioning for two tasks has also been found in relation to time since stroke. Chronic patients (≥ 12 months) were more involved in hobby/sport and car/bus travel in contrast to non-chronic patients (onset < 12 months) (Lin et al. 2012).

Despite good overall reliability, considerable variability in strength of agreement at the level of individual scale item scores has been reported both for test retest and inter-observer reliability (Green & Young 2001; Piercy et al. 2000; Wade et al. 1985). This may be due, in part, to the lack of specific criteria or guidelines for scoring items and reliance upon the discretion or interpretation of the individual administering the test (Piercy et al. 2000; Post & de Witte 2003). In contrast to other measures of ADL and IADL (BI and Nottingham Extended ADL Scale), a floor effect has been found with the FAI, where a significantly large number of pateints (19%) score the minimum 0 value (Sarker et al. 2012).

While the FAI has been assessed for use by proxy with good overall results, there is less agreement between proxy and patient assessments at the item level (Tooth et al. 2003; Wyller et al. 1996). In addition, there are a number of reported biases that should be kept in mind when considering the use FAI scores obtained via proxy. In a study by Tooth et al. (2003) it was reported that patients tended to score themselves as performing activities more frequently than proxy respondents especially in meal preparation, heavy housework, social outings, driving and home maintenance. In addition, male proxy respondents and respondents who were friends or relatives (rather than spouses) tended to give higher ratings, particularly in the area of domestic activities (Tooth et al. 2003). This response pattern may be explained by the reduced amount of exposure to patient activities on the part of a friend and/or by traditional gender differences in activity patterns (Tooth et al. 2003; Wade et al. 1985).

Summary – Frenchay Activities Index

Interpretability: The lack of standard guidelines for administration and reliance on the interpretation of the individual administrator reduces interpretability and comparison across studies.
Acceptability: Short, simple and encourages participation of significant others or family members. It is suited to use with proxy respondents.
Feasibility: Simple to administer and requires no training or special equipment. It has been used for longitudinal assessment.

Table 20.3.10.1 FAI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	++ (TR) ++ (IO) +++(IC)	+++	+++	+	++	+++

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.11 Modified Rankin Handicap Scale (MRS)

Originally developed in 1957, the Rankin scale is a global outcomes rating scale for patients post-stroke (Rankin 1957). The scale assigned a subjective grade from 1 – 5 based on level of independence with reference to pre-stroke activities rather than on observed performance of specific tasks. By referring to pre-stroke levels of independence, previously existing limitations are taken into account and discounted in the final rating.

An original Rankin score of 1 indicated no significant disability and 5 the most severe level of disability. van Swieten et al. (1988) expanded the ranking system to include 0; no symptoms (see below). Criticism that the Rankin scale focused on disability rather than handicap lead to suggestions that the scale be further modified by introducing changes to the wording of items to include “lifestyle” and replacing “disability” with “handicap”. The conventional method of administration for the Rankin Scale is via a guided interview process.

Table 20.3.11.1 Modified Rankin Handicap Scale

Rankin Grade	Description
0	No symptoms
1	No significant disability despite symptoms; able to carry out all usual duties and activities
2	Slight disability: unable to carry out all previous activities but able to look after own affairs without assistance.
3	Moderate disability: requiring some help, but able to walk without assistance
4	Moderately severe disability: unable to walk without assistance, and unable to attend to own bodily needs without assistance.
5	Severe disability: bedridden, incontinent, and requiring constant nursing care and attention.

(ref: van Swieten et al. 1988)

Advantages

The Modified Rankin Scale is an extremely simple, time efficient measure with well-studied reliability used to categorize level of functional outcome. As such, it is feasible for use large centers or in large trials (de Haan et al. 1993; Wade 1992). De Haan et al. (1993) suggest that scale scores may lend themselves to dichotomization (0-3 = mild to moderate disability & 4-5 = severe disability) for purposes of comparison in evaluating the effectiveness of an intervention.

Methods have been evaluated for administration of the mRS via telephone interview. Janssen et al. (2010) reported significant agreement between the results of telephone and face-to-face administration (k_w=0.71).

Limitations

The subjective nature of the score and lack of clear criteria by which to assign grades may diminish the reliability of the scale. It is suggested that using BI scores to generate Rankin grades could improve reliability (Wolfe et al. 1991). The categories within the scale have been criticized as being broad and poorly defined, left open to the interpretation of the individual rater (Wilson et al. 2002). In addition, the use of the term "without assistance" is problematic. There is no indication as to whether this might include the assistance of assistive devices or environmental modifications or other compensatory techniques that may enable the stroke survivor to improve the performance of daily activities (New & Buchbinder 2006).

The reported inter-rater reliability of the mRS is often somewhat low, particularly in studies with larger sample sizes (Quinn et al. 2009). A structured interview format for the administration of the Modified Rankin Scale is available. Use of the structured interview has been associated with significant improvements in interobserver reliability (Banks & Marotta 2007; Wilson et al. 2002, 2005). In addition, a recent guided interview and accompanying questionnaire in Japanese has been published (Shinohara et al. 2006). Quinn et al. (2007) describe the development of a training and certification package for the MRS. Based on certification assessment data, use of this standardized training procedure is associated with improved interobserver reliability, particularly among those raters who have passed their certification attempts (Quinn et al. 2008). Most recently, Saver et al. described the development of the Rankin Focused Assessment (RFA) tool that may be used to derive a mRS grade (Saver et al. 2010). The tool provides specific, operationalized criteria to distinguish between grade levels and allows the rater to indicate which functional difficulties were used in assigned a given score (Saver et al. 2010). As for other standardized assessment tools, use of the RFA was associated with improved inter-observer reliability.

Although the scale might be suitable for dichotomized groupings, there is no standardized or consistent point at which this is done (New & Buchbinder 2006; Sulter et al. 1999) suggesting a lack of consensus regarding favourable vs. poor outcome in terms of Rankin score.

The use of dichotomization to classify global outcome may be associated with a loss of information with regard to benefits derived any rehabilitation intervention. Lai and Duncan (2001) reported that 62% of patients included in their study experienced recovery represented by a shift of one or more Rankin grades in the first 3 months following stroke. If these shifts were between grades 1 and 0 or between 4 and 5, for instance, no change would be reported using a dichotomized system of outcome where favourable outcome was defined as MRS = 0, 1 and 2 and unfavourable as MRS = 3, 4 or 5. Lai and Duncan (2001) further demonstrated significant differences in physical and social functioning between Rankin grades of 0/1, 2,3, and 4 ($p<0.05$) as well as differences in the Barthel Index scores for patients with Rankin scores of 3, 4, and 5 ($p<0.05$). These benefits, associated with a transition in Rankin grades, would not be captured adequately by simple dichotomization of outcome. It is suggested that transition in Rankin grades might be more appropriate in the assessment of intervention benefit (Lai & Duncan 2001).

Summary – Rankin Handicap Scale

Interpretability: Very simple tool, useful for the categorization according to functional disability. It is easily understood and lends itself to dichotomization. However, there is no standardized point for this to be done thereby limiting comparisons. Use of the structured interview may increase reliability.

Acceptability: Administration of the Rankin by structured interview takes approximately 15 minutes. It has not been assessed for use with proxy respondents.

Feasibility: The MRS is time efficient and requires no special tools or training. Although it has been used to compare the effectiveness of interventions, there is no agreed upon dichotomization point by which to assess favorable vs. poor outcomes.

Table 20.3.11.2 MRS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) ++ (IO)	++	+++	+	+	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.12 Motor Assessment Scale (MAS)

The Motor Assessment Scale (MAS) was developed to provide valid and reliable means of assessing everyday motor function following stroke (Carr et al. 1985). The MAS is based on a task-oriented approach to evaluation that assesses performance of functional tasks rather than isolated patterns of movement (Malouin et al. 1994).

The MAS is comprised of 8 items corresponding to 8 areas of motor function (supine to side lying, supine to sitting over the edge of a bed, balanced sitting, sitting to standing, walking, upper-arm function, hand movements and advanced hand activities). Also included is a single item, general tonus, intended to provide an estimation of muscle tone on the affected side (Carr et al. 1985). Each item, with the exception of general tonus, is assessed using a 7-point hierarchy of functional criteria. Performance of each criterion is associated with a score ranging from 0 (most simple) to 6 (most complex) (Carr et al. 1985; Malouin et al. 1994; Poole & Whitney 1988; Sabari et al. 2005). Patients perform each task 3 times and the best of the three performances is recorded.

The general tone item is evaluated through observation and handling during the assessment. It is scored such that a score of 4 represents optimal function while scores greater or less than 4 are indicative of degrees of hypertonus and hypotonus, respectively (Carr et al. 1985). Item scores, excluding general tonus, may be summed to provide an overall score out of a possible 48 points (Malouin et al. 1994).

The scale is available from Carr et al. as are the criteria for grading each item and a list of general rules and equipment for the administration of the MAS (Carr et al. 1985). While Carr et al. (1985) suggested that administration of the MAS requires approximately 15 minutes, subsequent studies report administration times ranging from 15 to 60 minutes (Malouin et al. 1994; Poole & Whitney 1988).

Advantages

The MAS provides a brief and simple means by which to evaluate the performance of motor tasks following stroke. General rules for administration are provided along with a list of required equipment. Equipment required is commonly available in a variety of settings and includes a stopwatch, 8 jellybeans, a rubber ball, a stool, comb, spoon, pen, teacups, water and a table. However, a short instruction and practice period, including practice assessment on at least 6 patients, is recommended prior to using the test in a formal setting (Carr et al. 1985).

The MAS has been used as a tool to differientiate different groups of stroke patients. A Rasch-based scoring approach for the upperlimb subscale of the MAS [UL-MAS; includes three test items: (1) upper arm function, (2) hand movements, and (3) advanced hand activities], as opposed to the conventional summative score, can improve precision for discriminating between patient groups [patients who score in the upper (fourth) or lower (first) quartile versus patients who score in the second or third (middle) quartiles] at admission and discharge (Khan et al. 2013).

Limitations

Reports suggest that the item "general tonus" is difficult to assess reliably. The scoring criteria provided by the authors gives no guidance regarding the testing of tone, where it should be tested or how to score the item when tone varies between the arm, leg and trunk (Poole & Whitney 1988). This item is often omitted from the scale and reports using the MAS or about the MAS may not include it (Loewen & Anderson 1990; Malouin et al. 1994)

Items are assessed using a 7-point hierarchy of performance of motor activities. For each item, successful completion of a higher-level criterion implies that the individual would be able to meet all criteria corresponding to lower scores as well (Sabari et al. 2005). While this might serve to reduce the amount of time required for administration and increase interpretability (patients' with the same score can perform the same tasks), it is based on the assumption of an appropriate hierarchy of functions. The hierarchy of behavioural criteria has been examined for the items used to assess function in the upper limbs (items 6, 7, & 8) but not for the remaining items of the MAS.

Poole and Whitney (1988) and Malouin et al. (1994) both noted problems in the scoring hierarchy associated with the advanced hand activities item. In each case, it was reported that individuals who could complete the most difficult task (holding a comb and combing hair at the back of head) were unable to complete a lesser criterion (drawing horizontal lines). Sabari et al. (2005) used Rasch analysis to examine the validity of the scoring hierarchies for the upper arm function, hand movements and advanced hand activities items. Of these three items, only the upper arm function item demonstrated an appropriate hierarchy in terms of task difficulty. For each of the other items, substantial discrepancies in task order were identified as well as multiple tasks within each item of the same level of difficulty. In addition, substantial floor effects were identified for all items and ceiling effects for the upper arm function and hand movements items (Sabari et al. 2005). The authors make suggestions for the deletion and addition of criteria in order to improve the task hierarchy and alleviate the floor and ceiling effects. However, Miller et al. also used Rasch analysis to examine the UL subscales (MAS 6,7,8 – upper arm, hand movements and advanced hand activities) (Miller et al. 2010). Contrary to the results reported by Sabari et al. (2005) and Miller et al. (2010) found the test item hierarchy in the upper arm and hand movements subscales to be valid. The authors did demonstrate significant differential item functioning associated with age for a single item (#72 – radial deviation of the wrist) such that this task was easier to perform for individuals under the age of 65. It is recommended that the use of the upper limb items as a separate scale be approached with caution pending further investigation of the scoring hierarchy within these subscales (Hsueh & Hsieh 2002; Lannin 2004).

Summary – Motor Assessment Scale

Interpretability: Scores reflect a task-oriented approach to assessment. Use of a task hierarchy within items enhances interpretability; however, the validity of the task hierarchies used requires further study.
Acceptability: The test is relatively simple and brief to administer. Assessment by proxy is not appropriate as evaluation is performance-based.
Feasibility: The MAS is freely available in Carr et al. A period of instruction and practice assessment is recommended prior to formal use in a clinical or research setting (Carr et al. 1985). While the list of equipment required for administration is relatively long, items are commonly available.

Table 20.3.12.1 MAS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++(IO)	+++	++	+	+	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.13 Nine-hole Peg Test (NHPT)

The Nine Hole Peg Test (NHPT) is a timed, quantitative measure of fine manual dexterity. It is also a component of the National Multiple Sclerosis Society's Multiple Sclerosis Functional Composite (MSFC). The MFSC is a multi-dimensional quantitative measure that evaluates three dimensions (ambulation/leg function, arm/hand function, and cognition) in multiple sclerosis. The NHPT was developed by Kellor et al. (1971) and standardized by Mathiowetz et al. (1985). Mathiowetz et al. (1985) also published clinical norms for this instrument.

During this test the patient is seated at a table with a container holding 9 pegs and a wood or plastic block with 9 empty holes. While being timed, the patient is required to take the 9 pegs out of the container, one at a time, and place them into the empty holes in the block as quickly as possible. Once all of the holes are filled, the patient is required to remove each of the pegs, one at a time, and place them back into the container as quickly as possible. Total time required to complete the task is recorded. The test is run twice consecutively for the dominant hand and then twice consecutively for the non-dominant hand (Procedure from Multiple Sclerosis Society Website, *http://www.nationalmssociety.org/MUCS_9hole.asp*).

Test score is an average of the 4 trials. The two trials for each hand are averaged and then converted to the reciprocals of the mean times. These two reciprocals are then averaged. This score can be used individually or as part of the MSFC composite score. Lower scores indicate better fine manual dexterity. Norms for the NHPT have been published for both genders, as well as hand dominance in adults spanning 20 to 75+ years of age (Mathiowetz et al. 1985), and for both genders as well as hand dominance in children aged 4 to 19 (Poole 2005; Smith & Hong 2000; Yim 2003).

Administration time varies depending on the skill of the patient. However, the test typically takes 10 minutes or less. Training is required for administration and several commercial versions of the test are available for purchase. Major companies marketing these are Smith & Nephew Rehabilitation Division, Sammons Preston, S&S Worldwide, and North Coast Medical.

Advantages

Psychometrically, the NHPT has demonstrated good reliability and validity in adult as well as paediatric populations. Norms for age, gender, and hand dominance have been established, allowing for clarity of interpretation when testing for pathology in a clinical setting. However, original norms published by Mathiowetz et al. (1985) may not transfer directly to the more readily-used, commercial versions of the test. Another advantage with the NHPT is its flexibility, as it may be used on its own or as a component of the Multiple Sclerosis Functional Composite. Finally, the test is quick and easy to administer.

Limitations

The NHPT is susceptible to practice effects. Cohen & Marino (2000) demonstrated improved performance from test to retest. This effect tends to plateau after multiple administrations and researchers have therefore suggested administering the test several times to arrive at an accurate assessment of patient function.

Although norms as well as standardized procedure for the NHPT have been published for some time (Mathiowetz et al. 1985), numerous commercial versions, each with varying material and design, compromised the use of these norms (Davis et al. 1999). Because the commercial versions differ from the original used by Mathiowetz et al. (1985) the norms generated from that study do not transfer over well. For example, Davis et al. (1999) compared performance speed on the version used by Mathiowetz and colleagues to performance speed on the Smith and Nephew Rehabilitation Division version. On a sample of 32 patients between 21 and 72 years of age, the authors found significant differences in the time it took to complete the different versions of the test and concluded that the norms established by Mathiowetz et al. (1985) were not transferable to this version of the NHPT. As it is likely that similar results would be found with other commercial versions of the test, Davis et al. (1999) warned that extreme caution should be taken when interpreting original norms for the NHPT while using commercially available versions of the test. The authors also stressed that research was needed to develop norms consistent with commercially available versions. Fortunately, said norms have since begun to surface for adults as well as children (Oxford et al. 2003; Poole 2005). Further normative research for the various commercial versions of the test would be useful.

In addition, the generalizability of published normative values to the stroke population is questionable. Many individuals who experience stroke are elderly but few people 75 years of age or older participated in the normative studies for the NHPT (Kellor et al. 1971; Mathiowetz et al. 1985). Nonetheless, all mean values were greater than 20 seconds for healthy males age 60 and over, and greater than 18 seconds for healthy females age 60 and over (Mathiowetz et al. 1985). Wade (1992) maintains that people with normal function usually take 18 seconds to complete the task (if timing how long it takes to place the pegs only) and Heller et al. (1987) also used this as their criteria for “normal”. However, when using the test within an elderly population, it has been suggested that a completion time of 20 – 25 seconds be considered normal.

Summary – Nine Hole Peg Test

Interpretability: The NHPT is a simple and commonly used quantitative measure of fine manual dexterity. Normative data for adults and children on commercially available versions of the test exist; however, few elderly individuals were included in normative samples. Norms published by Mathiowetz et al. may not be transferable to various commercial versions of the test (Mathiowetz et al. 1985).
Acceptability: At approximately 10 minutes, administration is brief and should represent little patient burden.
Feasibility: Administration is brief and simple. Test materials are limited and easy to transport. Several versions of the test are available commercially and training is required to administer the test.

Table 20.3.13.1 NHPT Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++(TR) +++ (IO)	+++	+++	+	+	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.14 Rivermead Mobility Index (RMI)

The Rivermead Mobility Index (RMI) is an extension of the Rivermead Motor Assessment Gross Function Scale. It was intended as a short, simple way to provide a quantitative assessment of mobility disability focused on fundamental aspects of mobility, that is, aspects of mobility independent of one's social environment (Collen et al. 1991; Wade 1992).

The RMI is a hierarchical scale consisting of 15 items that progress in difficulty from item 1 through 15. Fourteen items are questions about the performance of functional activities assessed by self-report and one activity is assessed by direct observation. All items generate a dichotomous yes/no response. A "yes" response is given a score of 1. The total scale score ranges from 0 – 15 where a score of 0 would indicate complete inability to perform any of the functional activities included in the assessment.

Assessment using the RMI takes approximately 2 – 3 minutes and requires no special equipment or training (Collen et al. 1991; Forlander & Bohannon 1999). It is usually administered by interview of the patient and/or his or her primary caregiver (Hsueh et al. 2003).

Advantages

The RMI is a short and simple assessment requiring no special equipment or training and is easily performed in a variety of settings (Collen et al. 1991; Forlander & Bohannon 1999; Hsieh et al. 2000). Results of psychometric evaluation suggest that the RMI is a reliable instrument to assess and monitor mobility performance over time. The absence of differential item functioning (DIF), according to age, sex, or side of stroke lesion, allows for valid comparisons of RMI scores between subgroups of stroke patients (Roorda et al. 2012).

Level of performance is easily interpreted in a hierarchical (Guttman) scale such as the RMI. Patients with the same scores can accomplish the same things and changes in scores represent comparable changes in ability. It has been suggested that this represents a clear advantage over a summated index in which identical scores may be obtained from various item combinations and do not necessarily reflect the same level of performance (Hsieh et al. 2000). More recently, two independent start-and-stop rules have been formulated for the RMI (Roorda et al. 2012), offering improved interpretation and faster scoring. The first start-and-stop rule outlines to "start with the easiest and stop if the patient is unable to perform 3 consecutive items". The second rule is slightly different and outlines to "start with the most difficult item and stop if the patient is unable to perform 3 consecutive items".

Limitations

Franchignoni et al. (2003) identified potential difficulties in the order of the first 3 scale items while confirming that the RMI meets Guttman scaling criteria. They reported that more patients could perform the third task than either of the preceding 2 items. Given this, the authors suggested caution in interpreting the RMI as a true hierarchical scale.

The RMI reflects only the patient's own ability to move his/her own body. As such, it does not take into consideration increases in mobility achieved through environmental modifications, the use of assistive devices or with help from another person (Collen et al. 1991).

Summary – Rivermead Mobility Index

Interpretability: As a Guttman scale, level of performance as assessed by the RMI is easily understood and compared.

Acceptability: There is little patient burden associated with administration of the RMI. It takes only 3 – 5 minutes to administer and 14 of the 15 items can be completed by self-report with yes or no responses. While the assessment interview may include information provided by a primary caregiver, the use of proxy respondents for the 14 self-report items has not been assessed.
Feasibility: The RMI has been tested for use in longitudinal assessment. It is simple to administer and requires no special equipment or training. It can be used in a variety of institutional and community settings.

Table 20.3.14.1 RMI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++(TR) +++(IO) +++ (IC)	+++	+++	+++	+++	varied

NOTE: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)

20.3.15 Rivermead Motor Assessment (RMA)

The Rivermead Motor Assessment (RMA) was designed to assess the type and quality of movement during the course of recovery from hemiplegia, with the assumption that stroke patients follow a consistent pattern of physical recovery following a stroke (Lincoln & Leadbitter 1979).

The Rivermead Motor Assessment (RMA) requires patients to complete a series of functional movements in three categories: gross function, leg and trunk, and arm (Lincoln & Leadbitter 1979). The tool's items are ordered so that as the patient improves, they can successfully perform progressively more items in the hierarchy (Guttman scale). The RMA comprises 38 items. Each item is scored "1" if the patient can perform the activity or "0" if they cannot. Three tries are allowed per item, and the test is stopped after 3 items have been failed. The scores can range from 0= inability to perform any of the activities, to 38=patient can perform all of the activities. Depending on the patient's degree of motor recovery, the RMA may take up to 40 minutes to complete (Collin & Wade 1990) The RMA should be administered by a trained physiotherapist.

Advantages

By using a scaled assessment, the time spent assessing is directly related to the patient's motor functioning (Lincoln & Leadbitter 1979); that is, the higher the level of function, the less time required for the assessment. In addition, patients with the same score will be able to perform the same activities. While enhancing the interpretability of the scale, this assumption must be made with caution as not all of the sections of the RMA fulfill Guttman scaling criteria (Kurtais et al. 2009).

Given evidence that the gross function section can be self-reported (Sackley & Lincoln 1990), the RMA could serve as an indication of the patient's perceived ability to complete certain mobility activities.

Limitations

One of the most common criticisms of the RMA is that it can be time consuming to complete (Collen et al. 1990). Lincoln and Leadbitter (1979) reported that the RMA may take as long as 45 minutes to complete when assessing an ambulant patient with a recovering arm.

The validity of the RMA as a Guttman scale may be questionable. In two studies of the RMA, Adams et al. (1997a; 1997b) reported that only the gross function scale met the criteria for scaling (CS) and reproducibility (CR) among acute and nonacute patients. Two items included in the arm section were not passed in scale order in both populations and the section failed to meet scaling criteria in the nonacute population only. Problems with the order of items and inadequate CS and CR values, especially in the patient populations over 65 years of age, have also been reported for the leg and trunk section when used in populations of acute and nonacute stroke patients. Additional research has demonstrated that although the leg and trunk and the arm section satisfied Guttman scaling requirements on Mokken analysis, the gross function section did not (Kurtais et al. 2009). In addition, for all 3 sections, the ordering of items did not agree with the hierarchy as proposed originally by the scale authors (Kurtais et al. 2009). It has been suggested that, perhaps, use of the Guttman technique may not be appropriate and an alternative criteria developed to dictate a stop routine for assessment (Adams et al. 1997a, 1997b; Kurtais et al. 2009).

Summary -- Rivermead Motor Assessment

Interpretability: Scores are straightforward, based on ability vs. inability to perform scale items. Interpretability is enhanced by the characteristics of a Guttman scale. However, such interpretations should be treated with caution given the scaling problems inherent in the RMA. The following categories for severity of hemiplegia have been proposed (Endres et al. 1990); 0 – 9 = plegia, 10 – 15 = severe paresis and 15+= mild paresis.
Acceptability: The test can be lengthy, requiring up to 45 minutes to complete. While the gross function section can be self-reported (Collen et al. 1990) the other two sections have not yet been assessed to determine if self-reporting is reliable. Potential safety issues during gross function assessment may be minimized by close supervision of a trained physiotherapist.
Feasibility: Aside from the gross function section that is suitable for self-completion, the rest of the RMA needs to be completed by a physiotherapist (Collen et al. 1990). The physiotherapist does not require any training to administer the RMA, nor does the RMA require any specialized equipment.

Table 20.3.15.1 Evaluation Summary Rivermead Motor Assessment

Reliability		Validity		Responsiveness		
Rigor	**Results**	**Rigor**	**Results**	**Rigor**	**Results**	**Floor/ceiling**
+	++ (TR) + (IO) +++ (IC)	++	++	+	++	Possible floor effect

NOTE: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;

20.3.16 Six-Minute Walk Test (6MWT)

The 6-minute walk test (6MWT) is a submaximal test of functional exercise capacity. A submaximal test refers to one in which patients self-pace their performance, and generally reach a steady state of oxygen uptake and carbon dioxide production as opposed to achieving a maximal work load (Steele 1996). As a measure of exercise capacity, "*the 6MWT evaluates the global and integrated responses of all the systems involved during exercise, including the pulmonary and cardiovascular systems, systemic circulation, peripheral circulation, blood, neuromuscular units, and muscle metabolism,*" (American Thoracic Society 2002). The 6MWT evolved out of the 12-minute walk test (McGavin et al. 1976) and was developed to provide a measure that was less time consuming and better tolerated by patients (Butland et al. 1982). As with other walking tests, the 6MWT has predominantly been used to assess outcomes in individuals with cardiac and pulmonary diseases, particularly chronic obstructive pulmonary disease (COPD). However, there is some literature addressing the use of this test in stroke populations

(Dalgas et al. 2012; Danielsson et al. 2011; Eng et al. 2004; Flansbjer et al. 2005; Fulk et al. 2008; Lam et al. 2006; Liu et al. 2008; Mehrholz et al. 2007; Ng 2011; Ng & Hui-Chan 2005; Perera et al. 2006; Tang et al. 2006; van Bloemendaal et al. 2012).

During the test, participants are asked to cover as much distance as possible while walking on a hard, level surface for a period of 6 minutes. While the American Thoracic Society guidelines (2002) for the 6MWT recommend using a hallway 100 feet in length; some researchers prefer the use of continuous (oval) tracks. Patients choose their own intensity of exercise and are allowed to stop and rest during the test, at their own discretion. Performance on the 6MWT is measured by total distance walked in feet or meters (6MWD) within the 6 minutes. Complimentary to distance, dyspnea, as measured by the modified Borg dyspnea scale, oxygen saturation (S_PO_2), and pulse rate, is often assessed at the start and end of the test. The test may be administered before and after an intervention to determine if the patient has experienced a clinically significant improvement in function. To this end, distance walked pre- and post-intervention is compared to determine if significant change has occurred.

According to the American Thoracic Society, materials required for the test are a countdown timer (or stopwatch), a mechanical lap counter, two small cones to mark turnaround points, a chair that can be easily moved along the walking course, worksheets on a clipboard, a source of oxygen, a sphygmomanometer, a telephone, and an automated electronic defibrillator. Standardized protocol for test procedures including required materials has been published (American Thoracic Society 2002). Technicians should be trained in the administration of the 6MWT.

Advantages

The 6MWT is safe, simple to administer, inexpensive to perform and well-established psychometrically. In a review of functional walking tests, Solway et al. argued that the 6MWT is better tolerated and more reflective of activities of daily living than other walk tests (Solway et al. 2001).

In comparison to the 12-Minute Walk test (12MWT), the 6MWT is advantageous in that it takes less time to administer and is less physically demanding on patients. Further, the shortened length allows for the test to be repeated up to three times a day in most elderly patients. This helps control for outside sources of error variance, thereby providing more reliable and valid test results. Although the American Thoracic Society guidelines (2002) state repetition as unnecessary in its standardized protocol for the 6MWT, many researchers still feel that repeat testing is important to control for practice effects. Additionally, the 6-minute duration of the test represents a period of time that many ambulatory patients can manage without stopping. As such, it may be a more suitable test to assess dyspnea than longer walking tests in which severely limited patients may be required to stop and rest on several occasions.

As a submaximal test of functional capacity, the 6MWT may be more ecologically valid than conventional exercise testing, which focuses on maximal exercise capacity. The self-paced nature of the test is more reflective of the level of functional exercise demanded by most activities of daily living, which are themselves performed at submaximal levels of exertion. Indeed, the 6MWT has been found to correlate well with self-report measures of everyday physical function in numerous studies (Barr et al. 2000; Guyatt et al. 1985, 1991).

In addition to its predominant indication as a measure of response to intervention in cardiopulmonary disease, 6MWT has been used as a one-time measure of functional status of patients, and as a predictor of morbidity and mortality (American Thoracic Society 2002).

Adult and elderly norms for both genders have been published (Curb et al. 2006; Dalgas et al. 2012; Enright & Sherrill 1998; Fulk et al. 2008; Liu et al. 2008; Mehrholz et al. 2007; Miyamoto et al. 2000; Steffen et al. 2002; Stevens et al. 1999; Tang et al. 2006; Troosters et al. 1999; van Bloemendaal et al. 2012; Wevers et al. 2011). However, there is variability in gender and age specific 6MWT found amongst these studies. This variability may be accounted for by the use of differing procedures and sample populations. Age, height, weight, and sex all independently affect 6MWT in healthy adults and should therefore be considered when interpreting test results.

Limitations

It has been established that learning effects occur with the 6MWT. However, the American Thoracic Society (2002) argues that practice tests are not necessary, as learning effects result in only slightly better performance on subsequent tests. Nonetheless, many studies continue to run multiple tests to control for learning effects, which tend to plateau after a second test. When this is done, the longest distance walked (feet or meters) between tests is commonly used as the measure of performance. If multiple tests are run, the American Thoracic Society recommends waiting at least 1 hour between repeat administrations. Moreover, encouragement has been found to significantly improve test performance (Guyatt et al. 1984) and standardized protocols for timing and content of encouragement have been published (American Thoracic Society 2002). In spite of this, there remains considerable variability in the way encouragement is used by researchers, if it is used at all.

As with all walking tests, the 6MWT is susceptible to effects of learning and motivation. Three major sources of error variance with this test are practice effects, investigator influence (in the form of encouragement), and self-pacing. A study done by Liu et al. (2008) reported a practice effect across repeated trials of the 6MWT in individuals post stroke, a finding which the authors suggest has been demonstrated in healthy elderly adults and in individuals with cardiorespiratory ailments. Other sources of potential error variance include the use of supplemental oxygen and the use of medications during or around the test period. This variance, which is not brought about by actual change in physiological function, can obscure test results and must be controlled through standardized administration and for naïve subjects, multiple tests (although there is debate in the literature about this). While the American Thoracic Society has come up with a standardized, quality assurance format to control error variance (2002), there is still considerable variability in test administration within the literature.

The distance covered during the 6MWT and the number of turns in the course may have an effect on the outcomes of distance covered during the 6MWT. Ng et al. (2011) reported that the distance covered and the number of turns taken while completing the 6MWT were significantly associated with distance covered ($p < 0.05$), such that the greatest distances walked were associated with the fewest number of turns using a 30-meter walkway. For all of walkway lengths evaluated, turning to the affected side versus turning to the unaffected side did not result in a significant difference in the distance covered and the number of turns taken (Ng 2011).

As a submaximal test, the 6MWT is only useful in assessing those with moderate to severe exercise limitation. Individuals with mild cardiopulmonary disease/exercise limitation may not be impaired in their ability to walk and thus fail to demonstrate change or limitation on the test (Steele 1996). Moreover, the submaximal nature of the test means that it "*does not provide specific information on the function of each of the different organs and systems involved in exercise or the mechanism of exercise limitation, as is possible with maximal cardiopulmonary exercise testing*" (American Thoracic Society 2002). Thus, 6MWT is limited in its ability to explain the underlying cause(s) or mechanism(s) of exercise limitation. The information provided by 6MWT should be considered complimentary to rather than a replacement for cardiopulmonary exercise testing. As well, the submaximal nature of the test means that it cannot be used to assess exertion.

While the 6-minute duration of the test is time effective and less demanding on patients than the 12-minute test, it may be a disadvantage when self-pacing is an outcome of interest, particularly in pulmonary rehabilitation (Steele 1996). If this is the case, the longer 12MWT may be a more useful measure than the 6MWT.

Absolute contraindications for the test include: unstable angina during the previous month and myocardial infarction during the previous month. Relative contraindications include: a resting heart rate of more than 120, a systolic blood pressure of more than 180 mm Hg, and a diastolic blood pressure of more than 100 mm Hg (American Thoracic Society 2002).

Summary – Six-Minute Walk Test (6MWT)

Interpretability: The 6MWT is a widely used tool that provides a quantitative measure of submaximal exercise capacity. In spite of a detailed standardized protocol put forth by the ATS (2002), there still exists considerable variability in the administration of this test. Several studies have generated normative data for the 6MWT in healthy, adult samples (Curb et al. 2006; Enright & Sherrill 1998; Miyamoto et al. 2000; Stevens et al. 1999; Troosters et al. 1999; Wevers et al. 2011). However, there is a lack of consensus amongst these studies with respect to 6MWT in healthy adults. This discrepancy may be due to differences in test procedure and/or population investigated. There is consensus that age, height, weight, and sex all independently affect 6MWT in healthy adults and should considered when interpreting results of a single test provided to determine functional status (American Thoracic Society 2002).
Acceptability: The 6MWT is relatively brief and well tolerated by patients, though its use may be complicated by issues of endurance.
Feasibility: The test is brief, inexpensive and simple to administer. However, it does require considerable space to set up, and finding a quiet space where patients will not be distracted may be a challenge. Training is required to administer the test.

Table 20.3.16.1 Evaluation Summary of the 6MWT

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++ (IO)	+++	+++	++	++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;*

20.3.17 Timed "Up & Go" Test (TUG)

An objective measure of basic mobility and balance maneuvers; the timed "up & go" assesses the ability to perform sequential motor tasks relative to walking and turning.

The TUG requires subjects to stand up from a chair, walk a distance of 3 meters, turn around, walk back to the chair and seat themselves. The subject wears regular footwear and is permitted the use of a walking aid if one is required normally. This activity is timed, though the subject is permitted to walk through the test once before the timed session is undertaken. It is administered through direct observation of task completion. The score consists of the time taken to complete the test activity, in seconds.

The TUG is a variation of an earlier test; the "get-up and go" (Mathias et al. 1986) in which the test activity was the same, but not timed. Instead, the test was videotaped and later reviewed by examiners who assigned a rating on a scale from 1 (normal) to 5 (severely abnormal).

Advantages

The Timed "Up & Go" is quick and easy to administer with high inter- and intra-reliabilty, demonstrating consistent and reliable results (Faria et al. 2012). As the test requires no training or specialized equipment (an appropriate chair, a stopwatch or watch with a second hand, and space to walk 3 meters), it can easily be accomplished in community as well as institutional settings. Timed scores are objective and straightforward. Timed assessment is more sensitive to change over time than ordinal measures (Whitney et al. 1998).

Limitations

Rockwood et al. (2000) suggest that the TUG may not be suitable for use among broad, heterogeneous populations. Studies reporting high levels of test retest reliability excluded subjects exhibiting cognitive impairment and, therefore may be more feasible among cognitively intact populations. However, Nordin et al. (2006) reported that, among older individuals with multiple concerns living in residential care (mean MMSE = 18.7, SD = 5.6), the presence of cognitive impairment was not associated with increased variability of scores when verbal cuing was permitted during testing. Rather, the authors suggest that increased variability in TUG performance could be related to frailty and the presence of multiple concerns involving multiple systems.

The TUG is a limited measure addressing relatively few aspects of balance that concentrates primarily on speed rather than quality of performance (Ng 2011). It yields a narrower assessment than more comprehensive balance measures such as the Berg Balance Scale (Whitney et al. 1998). When used in the prediction of falls, it demonstrated lower sensitivity and specificity than the Berg Balance Scale (Andersson et al. 2006). Nordin et al. (2008) demonstrated that, in a group of frail elderly individuals living in a long-term care facility, a score of 15 seconds or less could be used to rule out high risk for falling (negative likelihood ratio = 0.1, 95% CI 0.0 – 0.4). However, TUG scores were not useful in ruling in high risk patients, perhaps due to a non-linear relationship between mobility (assessed by TUG) and risk for falls which may be modified by other factors, both behavioural and environmental (Nordin et al. 2008).

No normative data is available for the TUG, so its primary use has been assessment of change within the individual (Thompson & Medley 1995). Thompson and Medley (1995) reported mean TUG times with and without a cane for 3 age groups of community dwelling seniors (aged 65-69, 70-74, 75-79) and recommended that these times form the basis for standardized mean times. They also noted that while there appeared to be no significant relationship between TUG times and age, there was a tendency for women to perform the test more slowly than men ($p<0.01$), particularly with the use of a cane ($p<0.0001$). Subsequent research has reported a significant ($p<0.001$) age-related decline in TUG scores at discharge from a geriatric day hospital rehabilitation program, while no effect of gender was found (Hershkovitz & Brill 2006).

Siggeirsdottir et al. (2002) reported performance on the TUG to be related directly to chair type ($p<0.001$). Recommendations were made for a standardized chair type with armrests and a seating height of 45 – 47 cm.

Summary – Timed "Up & Go"

Interpretability: Scores are objective and straightforward. Standardized mean times with and without a cane have been suggested for community dwelling men and women in 3 senior age groups.
Acceptability: It is a short, simple activity taking only a few minutes and requiring only basic manoeuvres. Less reliability has been noted among patients with cognitive impairments.
Feasibility: The TUG requires no specialized equipment, training or large amount of time.

Table 20.3.17.1 TUG Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++ (IO)	+++	+++	+	++	+ (*floor – pts unable to complete*)

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.3.18 Wolf Motor Function Test

Originally developed as the Emory Motor Test (Wolf et al. 1989), it was intended to quantify, based on timed performance of tasks, the effect of forced use on upper extremity (UE) function in chronic stroke. Since its initial development, the scale has been modified and renamed the Wolf Motor Function Test (Morris et al. 2001; Wolf et al. 2001, 2005). The WMFT has been used in the study of UE function post stroke, most often in the study of constraint-induced movement therapy (CIMT).

The current version of the WMFT consists of 17 items or tasks. Tasks are arranged in order of complexity and progress from proximal to distal joint involvement (Wolf et al. 2001). Tasks 1 – 6 involve joint-segment movements and tasks 7 – 15, integrative functional movements (Wolf et al. 2001). Tasks are assessed for performance time and quality of movement and function. While each task is timed excessive performance time is typically truncated to 120 seconds. Summary score for performance time assessment is the median time recorded over all tasks (Morris et al. 2001).

Functional scores for the WMFT are derived via the application of a 6 point scale, ranging from 0 (does not attempt with involved arm) to 5 (arm does participate and movement appears normal). Functional ability scale (FAS) scores are expressed as the mean of item scores (Morris et al. 2001), although some have reported using a summed score with a maximum of 75 points (Ang et al. 2006). Performance on the 2 items that assess strength is neither timed nor rated. The patterns of movement assessed by the WMFT range from simple to complex and may be used with individuals demonstrating a range of upper extremity motor function. It provides assessment of both performance time and quality of movement. It should be noted that while the WMFT does provide some assessment of function, over half of the items on the WMFT involve simple limb movements with no clear functional endpoint (Morris et al. 2001)

The WMFT is available free of charge. Although specific equipment is required for the assessment, items are common and easy to obtain. Test administration is fairly lengthy, requiring approximately 30-45 minutes (Bogard et al. 2009). Training is required in order to ensure reliable administration.

Advantages

The WMFT is a stroke-specific scale that is available free of charge and requires commonly-available equipment for administration. The functional ability score of the WMFT (WMFT FA) has shown to be responsive (sensitive indicator of clinical change) in the acute stage of stroke recovery (Edwards et al. 2012). Although administration of the WMFT can take 30-45 min, two streamlined versions (S-WMFT; 6

tasks, rather than 17) exist (Bogard et al. 2009), one for subacte stroke patients and one for chronic stroke patients. Validity and relaiablity research has been conducted for both versions of the S-WMFT (Chen et al. 2012), however the S-WMFT in subacute stroke patients has been found to have a small level of responsiveness (not very sensitive to change) (Fu et al. 2012).

Limitations

Although Wolf et al. (2005) reported a strong correlation between time for task completion and FAS ratings, Richards et al. (2001) reported only a weak association between these two scoring elements suggesting that these may not represent assessments of the same aspects of upper extremity function. While Hsieh et al. (2009) demonstrated moderate associations between total and motor FIM scores and timed performance scores, the relationship between quality of movement (FAS) and FIM scores was substantially weaker. In addition, only timed task completion was predictive of functional outcome as assessed on the FIM. These authors suggest that timed completion and ratings of movement quality may assess different aspects of the underlying construct of motor function in the UE (Hsieh et al. 2009). If use of the WMFT is intended to inform prognosis in terms of recovery of function, or facilitate treatment or discharge planning, the timed task completion may be a more useful assessment.

When interpretating the timed scores, as well as strength-based performance, one should note that performance may be affected by both gender and handedness (Wolf et al. 2006, 2005). It should be noted that in the streamlined versions of the test, Rasch analysis demonstrated no significant differential item functioning on the basis of sex, age or laterality of hemiparesis (Chen et al. 2012).

Information provided regarding the reliability and validity of assessment using the WMFT has been based on ratings made of videotaped testing sessions rather than direct observation. In video-taped assessment, the rater may review and rewind the tape as many times as desired to complete the assessment. This option, of course, is not available in situations involving direct observation. Videotaped assessment adds significant time and expense to any evaluation procedure and may impact the clinical feasibility of the scale. The relationship between videotaped and direct observation has been examined on a single occasion with favourable results; however, a modified version of the current WMFT was used (Whitall et al. 2006). Reported levels of reliability are based on thorough training and practice sessions using videotaped assessment conducted until a minimum level of reliability is achieved (Morris et al. 2001).

Originally developed for use in the assessment of individuals with mild to moderate stroke, significant floor effects have been demonstrated in individuals with lower levels of function. Task completion times are limited to 120 seconds, which may be too short for individuals with moderate to severe stroke (Bogard et al. 2009; Wolf et al. 2005). Although a modified version of the WMFT has been proposed for use with these individuals (Whitall et al. 2006), there is little additional information available at the present time regarding its measurement properties.

Pilot normative data for timed and strength tasks only has been published; however, the sample size was quite small (n=51) and could not accommodate stratification for variables identified as influencing scores (e.g. gender and handedness) (Wolf et al. 2006). The sample, consisting of healthy adults recruited by convenience, was stratified for age by decade (i.e. 40 – 49, 50 – 59, 60 – 69, 70 – 79), which resulted in 4 groups with relatively few individuals in each group.

Although the reported stability of the WMFT appears excellent, Lin et al. (2009) and Fritz et al. (2009) have reported varying estimates of absolute reliability based on a calculation of the minimal detectable change (MDC). The MDC provides an estimate of the smallest detectable difference that might be considered to be true change rather than measurement error. Lin et al. (2009) reported an MDC of 4.36 seconds for the WMFT timed performance based on a 90% CI. That is, should a patient demonstrate a change in performance time of 4.36 seconds or more, one would be 90% confident that this change was real and not attributable to measurement error. Fritz et al. (2009) reported very different MDC values of 0.5 and 0.7 seconds, based on the 90% and 95% confidence intervals, respectively. Although both authors used the same base formula to derive reported MDC values, calculations were conducted differently in each study. Fritz et al. (2009) examined the distribution of scores and determined that, for the timed scores distribution was skewed by the inclusion of a maximum score (121 seconds) for each incomplete task (16% of all timed items). In order to meet assumptions for normality, the timed scores required transformation (natural log – ln). Lin et al. (2009) did not provide information regarding number of patients receiving maximum scores or distribution of timed scores and did not report transformation of data. It should be noted, that the MDC (an indicator of true change) reported by Lin et al. (2009) exceeded the estimated MCID (an indicator of meaningful change) for WMFT performance time. MDC values for the functional ability scale scores did not vary quite so dramatically ranging from 0.1 (Fritz et al. 2009) to 0.37 (Lin et al. 2009). No transformation of data was performed for FAS scores by Fritz et al. (2009) as FAS scores were normally distributed.

The MCID may enhance interpretation of change over time and various estimates for the minimal clinically important difference (MCID) have been reported. Of course, differing estimates of the MCID may be obtained by using different methods of derivation and MCID estimates may vary according to context. Within a group of stroke survivors, Lin et al. (2009) reported an MCID for WMFT-time of 1.64 seconds when using 10 – 15% change on the Fugl-Meyer Assessment (UE) in an anchor-based calculation and 1.37 seconds when using an effect size benchmark (Cohen's effect size of 0.2). For the FAS scores, MCID estimates were 0.33 and 0.14, respectively (Lin et al. 2009). Lang et al. (2008) also reported anchor-based estimates of MCID values for WMFT time and FAS scores. However, rather than using objective ratings obtained from other assessments of the upper extremity such as the Fugl-Meyer Assessment, the authors used subjective ratings of perceived change on which to base their calculations. For performance time, the MCID was estimated to be 19 seconds when assessing the affected dominant UE. An MCID value for the affected nondominant extremity could not be estimated. MCID values calculated for the dominant and nondominant affected extremities were 1.0 and 1.2, respectively. For both time and functional ability, estimates provided by Lang et al. (2008) are far greater than those reported by Lin et al. (2009). Lang et al. (2008) used patient perceived change for their anchor-based calculation. It may be that change scores are not significantly associated with patient-perceived change in that improvement on scale items is of little meaning to the patient. Large changes may be necessary to take on personal meaning for the individual being tested (Lang et al. 2008). When interpreting change over time, one should take the means by which both the reported MDC and MCID estimates were calculated.

Summary – Wolf Motor Function Test

Interpretability: Scores provide an evaluation of upper extremity function based on both performance time and quality of movement. Although pilot normative data has been published, these should be used with caution. Reported MCID and MDC estimates vary substantially.
Acceptability: No reports of patient burden were found, although administration time of 30 minutes may be excessive for more impaired stroke patients.
Feasibility: Although the test itself is free for use, costs may be incurred in the training of individuals who are to administer the test. Clinical feasibility may also be limited by the length of time required for testing and possible requirements for videotaping. There is little evidence regarding the reliability or validity of the scale when used via direct observation.

Table 20.3.18.1 Wolf Motor Function Test Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	+++ (TR) +++ (IO) ++ (IC)	++	+++	++	++	+

__NOTE__: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)

20.4 Participation/Handicap Outcome Measures

The final section corresponds to the third level or category of the ICF classification system. Measures appearing in this section tend to include elements from all domains including those reflective of an individual's involvement in life situations such as social functioning or roles. While these measures have been used to assess health-related quality of life, it is not our intent to define such a construct or its assessment here.

20.4.1 Canadian Occupational Performance Measure (COPM)

The Canadian Occupational Performance Measure is an individualized outcome measure developed by Law et al., in consultation with the Department of National Health and Welfare and the Canadian Association of Occupational Therapists Task Force (Law et al. 1990). The COPM is a generic, client-centred tool, designed to help occupational therapists establish occupational performance goals based on client perceptions of need as well as to assess change in perceived performance and satisfaction with performance over time in areas or activities of personal importance (Law et al. 1990, 1994). Used in conjunction with Occupational Therapy Guidelines for Client-centred Practice, the COPM provides an assessment of the individual's perceived occupational performance in the areas of self-care, productivity and leisure (Finch et al. 2002; Law et al. 1994; McColl et al. 2000).

Administration of the COPM is a 5-step process conducted within a semi-structured interview performed by an occupational therapist (Table 20.4.1.1). The interview focuses on identifying activities that the client wants, needs or is expected to perform (Dedding et al. 2004; Law et al. 1990). Following step 3, patient and therapist create goals for therapeutic intervention. In order to augment understanding of the nature and cause of identified deficits, set short term objectives and plan appropriate interventions, the interviewer may need to supplement information gathered during the COPM interview through other means such as observation, administration of standardized tests, or assessment of patient environments, for example (Law et al. 1990).

Table 20.4.1.1 Administration and Scoring of the COPM*

Step 1: Problem Definition	The therapist conducts an interview of the individual respondent and/or caregiver. Six questions are provided to serve as guidelines for the interview process: for each performance area (self-care, productivity, leisure), the therapist provides examples of activities and asks if the client needs, wants or is expected to perform these activities. If the answer to any of these questions is "yes", the client is asked if he can or does perform these and, if so, whether he is satisfied with his performance. When the client identifies a need to perform an activity along with an inability to perform satisfactorily, this area/activity is designated as a problem.
Step 2: Problem Weighting	Using a scale from 1 (not important) – 10 (extremely important), the respondent rates each identified problem activity in terms of importance.
Step 3: Scoring	The five most important identified problems from step 2 form the scale items. The

	respondent is asked to rate each of these on a scale of 1 – 10 in terms of a) how well they can perform the activity (1 = not able to 10 = able to perform with excellence) and b) how satisfied they are with their present performance (1 = not satisfied to 10 = extremely satisfied). Item ratings of performance and satisfaction are multiplied by their corresponding importance rating to determine baseline scores for each activity (ranging from 0 – 100). Satisfaction & performances scores for all activities may be summed separately and then each divided by the number of rated activities (usually 5). These summary performance and satisfaction scores are used as the basis for comparisons over time.
Step 4: Re-assessment	At an appropriate time following the initial assessment and intervention, as determined by the therapist, the patient and/or caregiver is asked repeat step 3 for each activity included in the individualized COPM.
Step 5: Follow-up	To plan for treatment continuation, follow-up or discharge step 1 is repeated to determine if there are remaining areas of problems or if new problems have emerged.

** Law et al. 1990*

Pilot study data indicated that administration of the COPM interview process required 20 – 40 minutes (Law et al. 1990). However, length of administration may be dependent upon patient cooperation and cognitive ability (Chen et al. 2002). The COPM was designed to be administered by occupational therapists. Training is recommended in order to use the COPM successfully. The COPM manual and instructional/training program is available for purchase from www.caot.ca.

Advantages

Traditional questionnaires or scales usually assess performance on a pre-determined selection of activities, none of which may be important to the individual respondent. The item pool of the COPM is not fixed, rather it is defined by the respondent. Although this may have deleterious effects on the reliability and validity of the instrument (Cup et al. 2003), it is truly focused on the self-perceived problems and needs of individual patients. Therefore, it is helpful in identifying treatment goals and creating treatment plans that are both relevant to the patient and in keeping with his/her own priorities (Carswell et al. 2004; Cup et al. 2003; Law et al. 1990; Ripat et al. 2001; Wressle et al. 2002). Increased patient relevance may translate into enhanced participation or motivation for the individual engaging in the rehabilitation process (Bodiam 1999). Individual patients have provided positive feedback regarding the use of the COPM (Dedding et al. 2004).

Limitations

Use of the COPM requires that the therapist using the tool be comfortable with a client-centred approach to both assessment and practice (Law et al. 1994). The therapist must be willing to create a therapeutic partnership with the client. Both the client and therapist may need time and prior exposure or intervention to establish the necessary relationship for the COPM process to be successful (Law et al. 1990; Waters 1995). In addition, the interview process is of critical importance both in eliciting relevant information and devising patient-centred therapeutic interventions. However, the interview process is not standardized and both the quality and adequacy of information obtained from interviews may vary considerably between interviewers.

The sole measure of test stability available with regard to the COPM is test-retest reliability, since the individual respondent determines the item pool specific to his/her own situation at the time of the step 1 interview (Carswell et al. 2004). However, given the individualized nature of the item pool and the semi-structured interview format, a somewhat different interview with different results may occur even within conditions of supposed stability. New problems may arise and old ones subside on a daily basis. In addition, perceptions of problems change such that, while the same problems may be identified on 2 occasions, priorities shift and ratings of importance change (Cup et al. 2003; Eyssen et al. 2005). In clinical practice, the resulting decrease in reliability may not pose a problem; however, in a research setting the items included on the outcome measure need to be both reliable (stable) and valid (Cup et al. 2003).

The variable item pool of the COPM also creates difficulties in establishing the validity of the tool. Inherent differences in test contents (items included and the spectrum of possible activities or areas covered within the test) between it and the other measures against which one attempts to validate the COPM may weaken the reported strength of relationships between the COPM and other tools (Chan & Lee 1997; Cup et al. 2003).

Results obtained from the COPM may be dependent upon the ability of the client to both understand the process and have insight into their own situation(s). Patients with cognitive deficits as well as those with lack of insight or communication problems, may not be able to participate in the process effectively (Carswell et al. 2004; Cup et al. 2003; Law et al. 1990; Wressle et al. 2002) and may demand goals that are unattainable or inappropriate, making the process both cumbersome and time consuming (Wressle et al. 2002). The scale authors state that in those instances in which the respondent is unable to identify problem activities, a caregiver or proxy respondent may respond on the patient's behalf. However, the caregiver/proxy may not identify the same deficits or problems as the patient would and may not assign the same importance to problem activities (Law et al. 1990; Law et al. 1994). For example, in the initial pilot study, Law et al. (1990) reported differences in opinion between "clients" and family members with regard to the importance of activities. Unfortunately, no study examining the use of the COPM by proxy or comparing problems identified by patients and caregivers could be identified.

In studies examining the clinical utility of the COPM, patients have reported difficulties with the self-evaluation task, and in translating their problems into a score (Bodiam 1999; Dedding et al. 2004; Wressle et al. 2002). Chen et al. (2002) reported that, when compared to younger respondents, older individuals required more time to complete the assessment, required more explanation and were not familiar with the process of self-rating.

Summary – Canadian Occupational Performance Measure

Interpretability: The COPM may be used as a basis for goal setting and development of appropriate patient interventions. In addition, scores may be generated in order to facilitate comparison over time. However, due to the individualized nature of the scale development of or comparisons to normative values is not appropriate.

Acceptability: Patients have reported feeling more included in the process of their own therapy and rehabilitation goals are more relevant. However, some patients may find the process of self-evaluation and translating problems to a score a difficult one.

Feasibility: Successful, reliable use of the COPM requires training in addition to knowledge about client-centred practice and the theoretical basis of the COPM prior to use. The instrument, along with a manual and instructional program, is available for purchase through the Canadian Association of Occupational Therapists.

Table 20.4.1.2 COPM Measure Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	++	++	++	+	+*(p-values only)*	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.2 EuroQol Quality of Life Scale (EQ5D)

The EuroQol scale (EQ-5D) is a generic index instrument, developed by a multi-country, multi-disciplinary team, used to value and describe health states (Group 1990). The EQ-5D was intended to be brief and simple to administer representing little or no patient burden. It focuses on a core set of generic, health-related quality of life items to provide a broad, generic assessment. The EQ-5D was intended to promote the collection of a common data set for reference purposes or as a complement to other, more comprehensive measures (Brooks 1996; Coons et al. 2000; Group 1990; McDowell & Newell 1996).

The EQ-5D is a self-administered questionnaire, in 2 parts. The first contains a simple descriptive profile of health in five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). In the original EQ-5D, each dimension is represented by 3 statements corresponding to 3 levels (3L) of difficulty with the item – 1 (some problems), 2 (moderate problems) and 3 (extreme problems). Due to concerns with the psychometric properties of the EQ-5D-3L, namely discriminant validity and potential ceiling effects, the EuroQol group developed a modified version of the tool, the EQ-5D-5L. This new version consists of the same 5 dimensions of health, but has expanded the 3 levels of responses to 5 levels (5L) of responses. On this new version, each dimension receives a numerical rating of either 1 (no problems), 2 (slight problems), 3 (moderate problems), 4 (severe problems) or 5 (unable to). The respondent chooses the statement most applicable to him/herself at present within each dimension. These ratings are combined such that each combination of choices creates a 5-digit expression of a health state. Theoretically, there are 3125 such representations possible. By applying scores from a standard set of values, each of these health states can be transformed into a utility value from 0 (worst possible) to 1 (best possible). Standard weights or preferences for the EQ-5D-3L were derived from population data obtained using time trade off techniques (Finch et al. 2002). Values have been elicited for health states in Canada, Denmark, Finland, Germany, Japan, Netherlands, New Zealand, Slovenia, Spain, Sweden, UK, US and Zimbabwe. Equivalent value sets for the EQ-5D-5L have yet to be developed. In the interim, cross-walk value sets based on the on the EQ-5D-3L are available for some countries. In the absence of a geographically appropriate value set, researchers are advised to use a value set that best corresponds to their region (Oemar & Janssen 2013).

Part 2 of the EQ-5D consists of a visual analogue scale (VAS) on which respondents rate their current state of health from 0 (worst imaginable) to 100 (best possible).

While the EQ-5D was originally intended for self-administration, it can also be administered by interview. It takes approximately 2 – 3 minutes to complete and yields 3 types of information; a profile indicating the extent of problems experienced on each of 5 dimensions, a population-weighted health index and a self-rated assessment of current perceived health (Coons et al. 2000). The scale is in the public domain and may be used without cost for the most part. Restrictions on the use of the scale as well as current information and references regarding the EQ-5D are available from the website www.euroqol.org.

Advantages

The EQ-5D is very short and simple. High response rates have been reported; 80% (Dorman et al. 1997); 80% to 86% (Dorman et al. 1998); 92.5% (Barton et al. 2008). Reports of missing data are mixed though are relatively low in all (Dorman et al. 1997; Essink-Bot et al. 1997).

The scale also provides considerable flexibility. Though designed as a self-completed postal instrument, it can also be administered in face-to-face interviews and has been evaluated for use with proxy respondents. In addition, the data can be presented and used in 3 distinct forms; a patient profile in 5 domains based on unweighted responses, a health utility or index and an overall rating of perceived health. The development of the 5L version has also improved the ceiling effects and discrimitaory power of the tool (Janssen et al. 2013).

Limitations

The level of validity reported would suggest that the instrument may not be suitable for use in serial assessments of individual patients. It would be more appropriate for use in study and comparison of groups (Dorman et al. 1997; Essink-Bot et al. 1997).

Brazier et al. (1996) reported missing data rates of 10% when using the EQ-5D in an elderly population (mean age 80.1 years). This observation is supported by Coast et al. (1998) who demonstrated that the ability to self-complete the EQ-5D is directly related to age and cognitive function ($p<0.0001$). The authors also report that the probability of requiring interview administration to complete the scale increases from 11% at age 65 to 73% at age 85. This would increase the costs associated with using the EQ-5D with elderly populations.

While the scale has been assessed for use with proxy respondents post stroke, Dorman et al. (1998) observed that reliability was consistently lower when a proxy respondent completed the questionnaire on the patient’s behalf. Levels of agreement between proxy respondents and patients were acceptable for mobility and self-care; however, the more subjective the domain, the lower the levels of agreement. In the case of depression/anxiety, the agreement was no better than chance among the more severely affected stroke survivors (Dorman et al. 1997). Similarly, Pickard et al. (2004) reported the lowest levels of agreement for pain/discomfort ($k_w = 0.21$) and anxiety/depression ($k_w=0.18$) domains during the subacute phase (post acute, but prior to discharge). However, agreement between patient and proxy appeared to improve over time particularly within these more subjective domains (kw = 0.57 and 0.42 for pain/discomfort and anixiety/depression at 6 months, respectively) (Pickard et al. 2004).

The health state valuations used in the EQ-5D utility were derived from time trade-off techniques. These techniques may be prone to biases and have been shown to elicit lower values for minor and major stroke than standard gamble techniques (Post et al. 2001).

Summary – EuroQol Quality of Life Scale

Interpretability: EQ-5D uses population based utility weights (a set of empirically derived valuations) to provide a standard set of utility values for the 5-digit health state derived from the 5-domain index. These weights are available for a large number of countries and cultures. The health profile may also be considered as an unweighted profile in 5-dimensions and is accompanied by a rating of perceived health status.

Acceptability: Although designed to be short and simple, reports of missing data are mixed. Essink-Bot et al. (1997) report higher rates of missing data for the EQ-5D than for the NHP or SF-36. However, its simplicity and brevity remain an advantage for use with stroke survivors. Barton et al. (2008) reported a 92.5% completion rate for self-report administration in a group of individuals with stroke. It has been evaluated for use with proxy respondents though only the mobility and self-care domains remain reliable.

Feasibility: The EQ-5D is designed as a self-completion questionnaire than may be administered as a postal or telephone survey or in a face-to-face interview. It requires no special training to administer and both the scale itself and supporting information are readily available.

Table 20.4.2.1 EQ-5D Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	++ (TR) ++(IO)	+++	++	+	++	varied

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.3 LIFE-H (Assessment of Life Habits)

The LIFE-H or Assessment of Life Habits is a measure of person-perceived social participation. It assesses two things: 1) performance in the accomplishment of daily activities and social roles, and 2) satisfaction with this performance. It is a generic tool that takes into consideration the individual's subjective perception and was developed to evaluate the social participation of people with disabilities, regardless of the type of underlying impairment. The LIFE-H was developed by Fougeyrollas and Noreau (1998) in accordance with the Disability Creation Process (DCP) model, which came out of the revision process of the International Classification of Impairments, Disabilities and Handicaps (ICIDIH-1) framework. While based on this different model, the LIFE-H touches on most items from the ICF participation dimension (Dijkers et al. 2000).

The most recent version of the test (LIFE-H 3.1 Short Form) consists of 77-items covering 12 categories of DCP life habits, which are divided into a daily activities domain and a social roles domain (Tables 20.4.3.1 and 20.4.3.2, respectively). These domains containing the various life-habit items were formed on the basis of 2 concepts: 1) degree of difficulty when performing a life habit; and 2) the type of assistance required (technical aids, adaptation, and/or human assistance).

Table 20.4.3.1 LIFE-H, daily activities domain, categories and item examples

Category	Item examples
Nutrition	Preparing your meal Eating in restaurants
Fitness	Sleep Participating in physical activities to maintain or improve your health
Personal care	Attending to your personal hygiene Using a bathroom or toilet other than those in your home
Communication	Communicating with another person at home or in the community Written communication
Housing	Maintaining your home Doing major household tasks
Mobility	Getting around on slippery or uneven surfaces Driving a vehicle

Table 20.4.3.2 LIFE-H, social roles domain, categories and item examples

Category	Item examples
Responsibility	Making purchases Taking care of your children
Interpersonal relationships	Maintaining friendships

	Having a sexual relationship
Community life	Getting to public buildings in your community Participating in spiritual or religious practices
Education	Participating in educational activities or vocational training Undertaking vocational training
Work	Holding a paid job Carrying out familial or home-making tasks as your main occupation
Recreation	Participating in sporting or recreational activities Taking part in outdoor activities

Performance on the test is assessed with a 10-point accomplishment scale, where a total score of 0 indicates that the life habit is not accomplished and a score of 9 indicates that the life habit is accomplished without difficulty or help. If a specific life habit is not part of one's lifestyle because of personal choice, the item is marked as not applicable. Moreover, a normalized score for each category can be calculated for each section (daily activities and social roles) and for the LIFE-H as a whole. This procedure considers the variable numbers in each category, as well as the number of non-applicable items for the participant. Additionally, level of satisfaction related to the accomplishment of each life habit can be assessed on a five-point Likert scale ranging from very dissatisfied (1) to very satisfied (5).

Test administration takes approximately 20 to 30 minutes and training is recommended for administration (Gagnon et al. 2006).

Advantages.

In measuring the construct of social participation, the LIFE-H provides additional information that which is provided by measures of functional recovery like the FIM of SMAF. Desrosiers et al. (2002) note that in most studies assessing long-term impact of rehabilitation following a stroke, functional recovery is the main outcome measure. However, "*being able to walk, wash and dress are not the only factors needed to resume a 'normal' life,*" (Desrosiers et al. 2002). Both reintegration into the community and readjustment to life post-stroke involve a number of factors beyond these basic functions of living. While the LIFE-H designates many items to basic functional recovery, it also contains items touching on other significant roles and activities that are fundamentally related to successful community integration and optimal quality of life.

The LIFE-H is a generic tool that has been constructed so that it can be used for people with disabilities regardless of underlying cause of impairment. As such, the measure has thus far been used for several purposes across various populations. These include: 1) the development of a profile of handicap situations in children with cerebral palsy and in an older adult population (Desrosiers et al. 2004; Lepage et al. 1998); 2) to identify the occurrence of potential handicap situations and potential association with personal factors in individuals with spinal cord injury (SCI) (Noreau & Fougeyrollas 2000); and 3) to explore the bio-psycho-social predictors of handicap situations in stroke survivors after discharge from an intensive rehabilitation program (Desrosiers et al. 2002).

Finally, the LIFE-H has been well researched by the test creators and has demonstrated strong psychometric properties to date. However, research from outside sources would further validate the measure.

Limitations.

Administration time is considerable at 20 to 30 minutes. Also, further normative data would aid in the interpretation of test results.

Summary – LIFE-H

Interpretability: The test is designed to assess person-perceived social participation. Gender norms for a healthy, elderly (55-85+) population have been published (Desrosiers et al. 2004). This is a crucial population on which to have such data because it enables clinicians and researchers to distinguish changes in participation from normal aging as opposed to pathological aging.

Acceptability: A relatively simple test that takes between 20 and 30 minutes to administer. The length of time required for administration may be associated with patient burden.

Feasibility: Training is recommended for administration (Gagnon et al. 2006).

Table 20.4.3.3 – Evaluation Summary LIFE-H

Reliability		**Validity**		**Responsiveness**		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
++	++ (TR) ++(IO) ++ (IC)	+	++	n/a	n/a	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver;*

20.4.4 London Handicap Scale (LHS)

The London Handicap Scale was developed to provide an assessment of handicap based on the definition of handicap provided by the World Health Organization in the International Classification of Impairments, Disabilities and Handicaps (ICIDH 1990). As such, the LHS is a measure of "disadvantage for a given individual resulting from ill health that limits or prevents fulfillment of a role that is normal for that individual" (Harwood et al. 1994b). The scale is a "classification questionnaire" based on the descriptive system within the ICIDH and classifies handicap according to disadvantages on six dimensions (mobility, physical independence, occupation, social integration and economic self-sufficiency (Harwood et al. 1994a, 1994b).

Each dimension of the LHS consists of a single question. Responses to each question are provided in the form of 6 descriptive statements representing a 6-point hierarchical scale of perceived disadvantage within that particular dimension ranging from 0 (extreme disadvantage) to 6 (no disadvantage). Statements are presented in terms of what someone is able to within his/her normal environment regardless of human or technical assistance required. Respondents are instructed to select the descriptive statement most representative of his or her situation (Harwood et al. 1994a, 1994b).

The LHS provides a profile of handicap based on the responses within each of the 6 dimensions as well as a weighted total handicap score. This overall weighted score should be interpreted as an estimate of the desirability of the health state described by the respondent's profile (Harwood & Ebrahim 2000, 2000). A matrix of scale weights and simple equation to calculate the overall score is provided. Scale weights were derived through interviews with 79 randomly-selected, community dwelling adults who were asked to evaluate a series of possible health states that could be described by the LHS (Harwood et al. 1994a, 1994b).

The LHS is designed as a self-report questionnaire, though it may be completed by a carer or appropriate informant (Harwood et al. 1994b). It requires no training to administer.

Advantages

The LHS is brief and simple to complete and can be used as a postal questionnaire (Harwood et al. 1994a, 1994b). Although the concept of handicap has been replaced by participation in the more recent ICF, the dimensions of handicap within the LHS remain relevant and can be mapped into the participation domain (Jenkinson et al. 2000; Perenboom & Chorus 2003). The LHS has been translated into several other languages including Dutch (Perenboom & Chorus 2003), Hong Kong Chinese (Lo et al. 2001), Sichuan Chinese (Lo et al. 2007), Swedish (Westergren & Hagell 2006) and Turkish (Kutlay et al. 2011).

Most instruments do not measure participation as it appears within the ICF, but include assessment of body function and/or activity as well. In a study of 11 instruments, the LHS was judged to be one of 2 instruments most closely measuring the construct of participation (Perenboom & Chorus 2003). However, the authors note that while the items appear to be formulated in terms of participation, the descriptive response statements span all of the domains of the ICF, from body function to participation. Response statements that describe body functions are typically associated with greater degrees of restriction in participation (Perenboom & Chorus 2003).

Limitations

The use of the scaled matrix to derive a total score could be viewed as a limitation. Overall, it makes the scale more cumbersome to use and more difficult to interpret (Jenkinson et al. 2000). The original matrix of scale weights was developed from rating provided by only 79 community dwelling individuals. They were subsequently modified to include a further 224 interviews (Jenkinson et al. 2000). It has been demonstrated that a simplified non-weighted scoring scheme based on simple summation provides similar information to the original weighted format (Jenkinson et al. 2000).

As a weighted scale based on the views of a sample drawn from the general population, it does not directly assess changes in perceived handicap within the individual (Harwood et al. 1994b). As such, the authors recommend that the scale be used for group comparisons (e.g. in clinical trials or for observational epidemiology) (Harwood et al. 1994a, 1994b).

The LHS was designed as a measure of handicap or disadvantage due to ill health. It may not be appropriate for use among the general population. Dubuc et al. reported a large ceiling effect when the scale was used to assess handicap in a group of healthy, community dwelling adults (Dubuc et al. 2004).

While use of the LHS is commonly reported within the research literature, relatively little has been published with regard to the reliability, validity or responsiveness of the LHS from sources that do not include at least one of the scale's authors. Further, independent evaluation is required.

Summary – London Handicap Scale

Interpretability: Use of scaling weights make scoring and interpretation more difficult. The LHS total score represents an estimate of the relative desirability of a profile of disadvantage provided by responses in six domains.
Acceptability: The LHS is a simple and very brief self-report measure. The questionnaire may be completed by proxy; however, the effects of completion by proxy on scale reliability have not been tested.
Feasibility: The test requires no training to administer or score. The test is well suited to postal administration.

Table 20.4.4.1 LHS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling

+	+++	+	+++	+	++	+

NOTE: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)

20.4.5 Medical Outcomes Study Short Form 36 (SF-36)

The Medical Outcomes Study Short Form 36 (SF-36) is a generic health survey created to assess health status in the general population as part of the Medical Outcomes Study (Ware & Sherbourne 1992). It is comprised of 36 items drawn from the original 245 items generated by that study (McHorney et al. 1993; Ware & Sherbourne 1992).

Items are organized into 8 dimensions or subscales; physical functioning, role limitations- physical, bodily pain, social functioning, general mental health, role limitations – emotional, vitality, and general health perceptions. It also includes 2 questions intended to estimate change in health status over the past year. These 2 questions remain separate from the 8 subscales and are not scored. With the exception of the general change in health status questions, subjects are asked to respond with reference to the past 4 weeks. An acute version of the SF-36 refers to problems in the past week only (McDowell & Newell 1996).

The recommended scoring system uses a weighted Likert system for each item. Items within subscales are summed to provide a summed score for each subscale or dimension. Each of the 8 summed scores is linearly transformed onto a scale from 0 – 100 to provide a score for each scale. In addition, a physical component (PCS) and mental component score (MCS) can be derived from the scale items. Standardized population data for several countries are available for the SF-36 (McDowell & Newell 1996). The component scores have also been standardized with a mean of 50 and standard deviation of 10 (Finch et al. 2002).

The SF-36 questionnaire can be self-completed or administered in person or over the telephone by a trained interviewer. It is considered simple to administer and takes less than 10 minutes to complete (Andresen & Meyers 2000). Permission to use the instrument should be obtained from the Medical Outcomes Trust who oversee the standardized administration of the SF-36 and will provide updates on administration and scoring (McDowell & Newell 1996). Various computer applications are available to assist in scoring the SF-36 including free Excel templates that can be downloaded from the internet (Callahan et al. 2005).

Advantages

The SF-36 is simple to administer. Either form (self-completed or interview) of administration takes less than 10 minutes to complete (Hayes et al. 1995). As a self-completed, mailed questionnaire, it has been shown to have reasonably high response rates; 83% (Brazier et al. 1993; O'Mahony et al. 1998); 75% - 83% (Dorman et al. 1998); 85% (Dorman et al. 1999) 82% overall and 69% for those over age 85 (Walters et al. 2001).

Limitations

Higher rates of missing data have been reported among older patients when using a self-completed form of administration (Brazier et al. 1992, 1996; Hayes et al. 1995). O'Mahony et al. (1998) found item completion rates to range from 66% to 96%. At the scale level, complete data collection (amount required to compute a scale score) ranged from 67% (role limitations – emotional) to 97% (social functioning). Walters et al. (2001) reported scale completion rates among community dwelling older adults ranging from 86.4% to 97.7% with all eight scales being calculable for 72% of respondents. Dorman et al. (1999) reported proportion of missing data on the scale level ranging from 2% (social functioning) to 16% (role functioning – emotional). Given the lack of data completeness found, postal administration of the SF-36 may not be appropriate for use among older adults. However, low completion rates may not be limited to self-completion or postal administration. Andresen et al. (1999) administered the SF-36 to nursing home residents by face-to-face interview and reported that only 1 in 5 residents were able to complete it.

It has been suggested that data completeness may be indicative of respondent acceptance and understanding of the survey as relevant to them (Andresen et al. 1999; O'Mahony et al. 1998). Hayes et al. (1995) noted that the most common items missing on the self-completed questionnaire referred to work or to vigorous activity. Older respondents identified these questions as pertinent for much younger people and not relevant to their own situation. The authors suggested modifications to some of the questions, which may increase acceptability to older populations. In a qualitative assessment of the physical functioning and general health perceptions dimensions of the SF-36, Mallinson (2002) noted that the participants, who were all over the age of 65, tended to display signs of disengagement from the interview process and some participants expressed concern relating to the relevance of the questions. There was also considerable variation noted in subjective interpretation of items and most subjects used qualifying, contextual information to clarify their responses to the interviewer. As Mallinson (2002) pointed out, individual issues of subjective meaning and context are lost when the questionnaire is scored.

The SF-36 does not lend itself to the generation of an overall summary score. In scales using summed Likert scales, information contained within individual responses is lost in the total scale score (ie. any given total score can be achieved in a variety of ways from individual item responses) (Dorman et al. 1999). Hobart et al. (2002) examined the use of the 2-dimensional model, which consists of a mental health component (MCS) and physical health component (PCS) and found that these two scales could account for only 60% of the variance in SF-36 scores suggesting a significant loss of information when the 2-component model is used. In a recent factor analysis of the SF-36, Dallmeijer et al. (2006) reported that, while the 8-factor of structure of the SF-36 could be confirmed, use of the 2 summary scales in stroke populations should be reconsidered given that use of the 2 summary scales could account for only 56% of total variance and factor loadings deviated from the original factor structure. In addition, the general health, vitality and mental health subscales lacked unidimensionality when used to assess individuals with stroke (Dallmeijer et al. 2006).

The level of test re-test reliability reported in stroke populations indicate that the SF-36 may not be adequate for serial comparisons of individual patients, but rather should be used for large group comparisons only (Dorman et al. 1998). Weinberger et al. (1996) also questioned the usefulness of the SF-36 in serial evaluation of individuals given large reported absolute differences in SF-36 scores obtained via common modes of administration (face-to-face interview, self-administration and telephone interview) over short testing intervals.

Low rates of agreement were reported between proxy respondent and patient respondent ratings (Segal & Schall 1994) and test-retest reliability has also been shown to be negatively affected by the use of proxy respondents (Dorman et al. 1998). While the use of a proxy may be the only means by which to include data from more severely affected stroke survivors, the subjective nature of the SF-36 may make proxy use difficult or even inadvisable (Dorman et al. 1998).

Summary – Medical Outcomes Study Short Form 36

Interpretability: Use of scale scores and summary component scores represents a loss of information and decreases potential clinical interpretability. Standardized norms for several countries are available for the SF-36.

Acceptability: Completion times are approximately 10 minutes for either self-completed or interview administered questionnaires. Some items have been questioned for their relevance to elderly populations. The SF-36 has been studied for use by proxy, however, reliability of the test decreased when proxy respondents completed assessments.

Feasibility: The SF-36 questionnaire can be administered by self-completion questionnaire or by interview (either on the telephone or in-person). It has been used as a mail survey with reasonably high completion rates reported, however, data obtained are more complete when interview administration is used. Permission to use the instrument and additional information regarding its administration and scoring should be obtained from the Medical Outcomes Trust.

Table 20.4.5.1 SF-36 Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IC)	+++	+++	++	+++ *(Note: 1 study reported ES)*	+++ (total score – floor/ceiling) ++ (individual domains – floor) + / ++ (individual domains - ceiling)

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.6 Nottingham Health Profile (NHP)

The Nottingham Health Profile (NHP) was designed to be a brief, subjective measure of perceived health encompassing the social and personal effects of illness (Hunt et al. 1984, 1985, 1980, 1981). It was not intended to be a measure of health-related quality of life or as a means to identify specific health conditions (Bowling 1997; Hunt et al. 1984). Both the items and weights are intended to reflect the point of view of the lay person and were derived from statements regarding the effects of ill health collected from more than 700 patients with acute and chronic ailments (Hunt et al. 1981; McDowell & Newell 1996).

The NHP consists of 2 parts. Part I contains 38 items grouped into 6 dimensions or subsections of subjective health: physical mobility (8 items), pain (8 items), sleep (5 items), social isolation (5 items), emotional reactions (9 items) and energy level (3 items). Each item takes the form of a statement of a potential problem. Respondents answer yes or no to each statement according to whether or not they feel the item applies to them at the present time. Each statement carries with it a weight, based on perceived severity. Weights assigned to items in each dimension total 100. If a statement is affirmed, it is scored with its associated weight. All weighted responses within a section are summed to give a total score for that dimension out of 100. Higher scores correspond to poorer perceived health status. Results from the 6 dimensions should not be combined to provide a total score.

Part II contains 7 items representing areas or activities that may be influenced by the respondent's health: paid employment, jobs around the house, social life, personal relationships, sex life, hobbies & interests and holidays. Respondents provide yes or no answers as to whether each area is affected by the respondent's current state of health. Items in Part II are not weighted. A score out of 7 is obtained by adding together the number of positive responses. Administration of Part II is optional.

The NHP is a self-reported assessment that may be self-completed or administered by interview. It takes approximately 10 minutes to complete. A user's manual (Hunt et al. 1989) as well as reference scores for healthy people by age, group, sex and social class are available (Hunt et al. 1985).

Advantages

The NHP is a simple and concise measure. Reported completion times range from 5 to 15 minutes and, unless interview administration is necessary, administrative burden is minimal (Coons et al. 2000; de Haan et al. 1993; Tabali et al. 2012). As a postal questionnaire, reported response rates range from 68 – 93% (Brazier et al. 1992; Ebrahim et al. 1986; Hunt et al. 1985). Ebrahim et al. (1986) reported low rates of missing data (4 – 7%).

The NHP has been widely used and extensively studied. It was the first measure of perceived health developed for use in Europe.

Limitations

Overall, the NHP is a somewhat limited measure. It does not assess many areas of concern such as sensory deficits, incontinence, eating problems, stigma, memory, intellectual ability, or financial difficulty (Bowling 1997; Ebrahim et al. 1986). It is a negative measure of health assessing only the presence or absence of problems and does not address the presence of positive outcomes or feelings (Bowling 1997; Hunt et al. 1985). A score of zero is indicative only of an absence of the problems presented on the NHP and does not indicate a sense of well-being.

The statements comprising Part I reflect serious problems and this may limit the usefulness of the scale among less ill subjects. Given the prevalence of ceiling effects (scoring "0" – no problems), the NHP may not be suited for use in the general population or among individuals experiencing only minor illnesses or distress (Bowling 1997; de Haan et al. 1993; Stansfeld et al. 1997).

Although rates of completion may be high, in general, this may be affected somewhat by the presence of cognitive impairment. In a group of elderly nursing home residents (n=127, mean age = 83.6 years ±8.8), Tabali et al. (2012) reported significant differences in MMSE scores in individuals who completed the assessment compared with those who did not ($p<0.001$). Using ROC analysis, the authors determined that scale completion was most likely in residents with MMSE scores >16 (AUC = 0.80, sensitivity = 80%, specificity = 76%) (Tabali et al. 2012).

The use of the weights provided with the scale items has been criticized as being inappropriate and confounded (Anderson et al. 1993; Jenkinson 1991). In his 1991 study, Jenkinson (1991) gave values of 0 (no) and 1 (yes) to responses, summed the positive responses for each section and then expressed this summed total as a percentage. Scores derived by this simplified method were very highly correlated with results obtained using the traditional weighted system ($r=0.98$; $p<0.001$) suggesting that the use of weights may be unnecessary.

Part II is not well studied. Most evaluative research pertains to Part I. This may be due to its optional nature. The application of Part II may be more limited than Part I as many of the items would be inappropriate or irrelevant to a number of subject populations, such as the elderly, unemployed or disabled (Bowling 1997). It is has been reported that, subsequent to further developmental work, the authors no longer recommend the use of Part II (Bowling 1997; Coons et al. 2000).

Summary – Nottingham Health Profile

Interpretability: The NHP has been widely used in Europe and extensively studied. A complete user's manual is available (Hunt et al. 1989) as are population norms and scores for individual patient groups (Hunt et al. 1984).

Acceptability: The NHP is short & simple taking little time to complete. High response rates and low rates of missing data suggest that it is acceptable to respondents. It has been test for use with proxy respondents, however, reported reliability was low.

Feasibility: The test can be administered as either a self-report questionnaire or interview and has been used as a postal survey. The NHP is not suited for use in the general population or with mildly-impaired groups (Bowling 1997).

Table 20.4.6.1 NHP Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	+++ (TR) +++ (IC)	+++	+++	+	n/a	+ (ceiling) ++ (floor) (Cabral et al. 2012)

NOTE: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)

20.4.7 Reintegration to Normal Living Index (RNLI)

The Reintegration to Normal Living Index (Wood-Dauphinee & Williams 1987; Wood-Dauphinee et al. 1988) was developed as a short and simple way to assess, quantitatively, the degree to which individuals who had experienced traumatic or incapacitating illness achieve reintegration. Reintegration to normal living was defined by the scale authors as "the reorganization of physical, psychological and social characteristics of an individual into a harmonious whole so that one can resume well-adjusted living after an incapacitating illness or trauma" (Wood-Dauphinee & Williams 1987).

Based upon literature reviews and information gathered from consultations and testing with advisory panels consisting of healthcare professionals from a variety of disciplines, patients, relatives of patients and clergymen, 11 declarative statements were developed. Each of these statements are rated by the respondent on a 10 cm visual analogue scale (VAS) with the anchor statements of "Does not describe my situation" (1 or minimal reintegration) and "Fully describes my situation" (10 or maximum reintegration). Individual item scores are summed to provide a total score out of 110 points that is proportionally converted to create a score out of 100 (Wood-Dauphinee et al. 1988). Two subscales have been identified within the RNLI; Daily Functioning and Perceptions of Self. These may be calculated by combining the responses to the first 8 statements and the final 3 statements, respectively.

Three and 4-point categorical scoring systems were also developed (Wood-Dauphinee et al. 1988), however, the 10 cm VAS was selected over either of these. Despite this, the 3-point categorical system has been used in the evaluation of stroke patients (Mayo et al. 2002, 2000). In the 3-point system, an additional category is inserted between the two anchor points ("partially describes my situation") and the respondent selects the most applicable of the three categories. This option yields total scale scores from 0 – 22 (Mayo et al. 2002, 2000).

The RNL is short and simple. It requires no training to administer and is available free of charge. Patient and proxy formats are available as are English and French-Canadian versions.

Advantages

The RNLI is a simple, brief assessment tool. Versions are available for administration to either patient or appropriate proxy respondents in either French or English. The RNLI does not appear to be affected by either age or gender (Carter et al. 2000; Steiner et al. 1996).

The RNL focuses on the perception of the individual with regard to his or her own capabilities and personal autonomy rather than on the achievement of what is considered normal by society (Cardol et al. 1999). As such, it provides a patient-centred assessment of re-integration.

Limitations

Low correlations have been reported between responses given by healthcare professionals and patients. Given the subjective nature of the statements, the authors do not recommend that healthcare professionals be used as proxy respondents (Wood-Dauphinee et al. 1988).

While the use of subscales has the potential to provide more information than a single, summed score, the ideal composition of the subscales is uncertain. Using principal component analysis, the 2-factor structure of the index has been confirmed (Stark et al. 2005); however, the composition of the factors differed substantially from those identified by the authors of the RNLI. Stark et al. (2005) reported the presence of 2 factors; the first, labeled "social" consisted of 6 items (i.e. those concerned with personal relationships and family roles, socialization, coping with life events and social and recreational activities) while the second, labeled "physical" consisted of 5 items (i.e. those concerned with moving around in the home and community, taking trips, self-care and productivity). The authors suggested that this difference may be accounted for by the use of a different patient population than the one used in the initial validation study by Wood-Dauphinee et al. (1988)(Stark et al. 2005). Confirmation of the scale's factor structure has not been undertaken using a population of stroke patients.

While the RNLI has been used for the assessment of individuals who have experienced stroke, its reliability and validity have not been well-studied within this particular population. In addition, the use of a visual analogue scale in the assessment of stroke patients may not be appropriate. A study by Price et al. (1999) examined the use of visual analogue scales among stroke patients and found that, while the VAS was the most sensitive of the scales examine, it was associated with the poorest completion rates. Inability to complete the VAS correctly was associated with tactile inattention, hemineglect and cognitive and visuospatial impairments. A categorical rating system (in this case, consisting of none, mild, moderate, severe) was completed correctly more often than the VAS (Price et al. 1999). While a 3-point categorical system for the RNLI was developed and has been used in the stroke population, the reliability and validity of the 3-point response format has not been examined.

There are no generally accepted standards for interpretation presently available. A distribution of RNL scores was published in a study of patients (n=182) following subarachnoid haemorrhage (Carter et al. 2000). In that distribution, severe impairment included scores from 0 – 59, moderate impairment from 60 - 79, mild impairment from 80 – 99. A score of 100 was indicative of no impairment. However, this proposed distribution was obtained using a small sample of patients with subarachnoid haemorrhage. Further evaluation in a larger, less specialized population of stroke patients is required.

Summary – Reintegration to Normal Living Index

Interpretability: There are no generally accepted standards for interpretation. While a scoring distribution has been proposed for severe, moderate and mild impairment, the proposed distribution was based on a small subject sample. Further investigation using a large sample population is required.

Acceptability: Short and simple, administration of the RNLI represents minimal patient burden. It has been assessed for use with proxy respondents with moderate success when significant others are used.
Feasibility: The RNLI is available free of charge, although it is recommended that one contact the scale authors prior to use. No training is required to administer the RNLI and it has been assessed for use in longitudinal studies.

Table 20.4.7.1 RNLI Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (TR) +++ (IO)	+	++	+	++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.8 Stroke-Adapted Sickness Impact Profile (SA-SIP-30)

The Sickness Impact Profile (SIP) is a comprehensive, behaviourally-based measure of perceived health status originally intended for use in health surveys, program planning, policy formation & in monitoring patient progress in terms of sickness (Bergner et al. 1981, 1976). It has become one of the more commonly used generic instruments in the assessment of health-related quality of life.

The major drawback in the use of the SIP may be its length. It contains 136 items and may take more than 30 minutes to complete. As such, it represents considerable patient burden and may pose significant administrative difficulty for both clinical and research trial applications. A shorter version has been developed specifically for use in stroke outcome research in order to overcome problems of acceptability and feasibility associated with the longer SIP (van Straten et al. 1997).

The Stroke-Adapted Sickness Impact Profile (SA-SIP-30) was derived directly from the original scale. van Straten et al. (1997) followed a 3-stage process to eliminate items and subscales of least relevance to stroke survivors as well as those with the lowest levels of reliability (Golomb et al. 2001). The end result is a scale comprised of 30 items in 8 subscales (body care & movement, social interaction, mobility, communication, emotional behaviour, household management, alertness behaviour and ambulation). Scale items are weighted to reflect the relative importance of the item to health status. Weights used in the SA-SIP-30 are the same as those used in the parent version and were derived by health professionals, students and members of a group health plan (de Bruin et al. 1992).

Each item takes the form of a statement describing changes in behaviour that reflect the impact of illness on some aspect of daily life. Respondents are asked to mark items most descriptive of themselves on a given day. To score the SA-SIP-30, weights are applied to marked items, summed for each subscale and expressed as a percentage for each subscale. Higher scores are indicative of poorer health outcome (Cup et al. 2003; Finch et al. 2002; van Straten et al. 1997). Subscales can be combined to form 2 dimensions; physical (body care & movement, ambulation, household management and mobility) and psychosocial (alertness behaviour, communication, social interaction & emotional behaviour) (van Straten et al. 1997).

No special equipment or training is required though a user's manual and trainer's manual are available for the original SIP (McDowell & Newell 1996). Like the original SIP, the SA-SIP-30 may be self-administered or completed by interview.

Advantages

The SA-SIP-30 is a much shorter and simpler scale than the parent scale and is more suitable for use in stroke outcome research (Finch et al. 2002). Authors of the scale provide regression weights to allow for the calculation of estimated SIP scores from SA-SIP-30 scores (van Straten et al. 1997). In addition to maintaining much of the original subscale structure of the SIP, these weights help facilitate comparisons with studies using the original SIP-136. In addition, van Straten et al. (2000) have identified cutoff scores for representative of poor health. Patients with scores >33 were reported to be ADL disabled, unable to live independently, experienced some problems in self-care, mobility and in performing their main activity, and reported low values for health-related quality of life. Similar profiles were observed for physical dimension scores >40, but no cutoff values could be defined using the psychosocial dimension (van Straten et al. 2000).

Limitations

In the process of creating the stroke-adapted scale, items less relevant to stroke were removed (ie. applying to fewer than 10% of stroke patients). However, no attempt was made to supplement the scale with items or domains of potential importance to stroke. The stroke-adapted version does not assess pain, recreation, energy, general health perceptions, overall quality of life or stroke symptoms (Golomb et al. 2001).

In examining the weights of removed items, van Straten et al. (1997) note that higher item weights tended to be associated with items that were removed and were descriptive of more severe health states. The new scale, therefore, may be less effective when used with patients who have suffered a severe stroke. Agreement between scores obtained with the SIP-136 and SA-SIP-30 was lower among more severely ill stroke patients than among healthier patients (van Straten et al. 1997).

Total scores of the SA-SIP-30 appear to be largely explained by its physical dimension (66% for the subscales of the physical dimension vs 25% for the subscales of the psychosocial dimension) (van Straten et al. 2000). As such, the SA-SIP-30 may represent a measure of physical disability rather than the more comprehensive constructs of health status or health-related quality of life.

Summary – Stroke Adapted Sickness Impact Profile

Interpretability: Maintenance of original structure and scoring procedures from the SIP in addition to the provision of constants with which to calculate estimated SIP scores from those obtained with the SA-SIP-30 have enhanced interpretability. Cut-off scores for poor health outcomes have been proposed (van Straten et al. 2000).
Acceptability: The SA-SIP-30 is shorter and simpler than the original, thereby reducing the associated patient burden. The original SIP has been tested for use with proxy respondents.
Feasibility: This shorter, simpler version of the SIP should represent less administrative burden and can be more easily included in both research and clinical settings.

Table 20.4.8.1 SA-SIP-30 Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	++ (IC)	++	++	+	++	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.9 Stroke Impact Scale (SIS)

The Stroke Impact Scale is a stroke-specific, comprehensive, health status measure. The scale was developed with input from both patients and caregivers (Duncan et al. 1999) and is intended to include domains from across the full impairment-participation continuum (Duncan et al. 2000).

Version 2.0 was comprised of 64 items in 8 domains (strength, hand function, ADL/IADL, mobility, communication, emotion, memory and thinking, participation) (Duncan et al. 1999). Based on the results of a Rasch analysis process, 5 items have been removed from version 2.0 to create the current version 3.0 (Duncan et al. 2003). The SIS is a patient-based, self-report scale in which each item is rated on a 5-point Likert scale in terms of the difficulty the patient has experienced in completing each item during the past week. A score of 1 represents an inability to complete the item and a score of 5 represents no difficulty experienced at all. Care should be taken when administering and scoring three of the questions in domain 3 (emotion) - 3f, 3h and 3i. These questions treat a score of 5 as the most negative score and a 1 as the most positive score. For final summary score purposes, these values need to be transformed (i.e. 5=1, 4=2, 3=3).

Using an algorithm equivalent to the one used in the SF-36, aggregate scores are generated for each domain. Domain scores range from 0 – 100. Factor analysis of the SIS 2.0 revealed that the 4 physical domains (strength, hand function, mobility and ADL/IADL) can be summed together to create a single, physical dimension score while all other domains should remain separate (Duncan et al. 1999). One item is included to assess the subject's overall perception of recovery. The item is presented in the form of a visual analog scale from 0 to 100 where 0 indicates "no recovery" and 100 indicate "full recovery".

The SIS was originally developed for administration by face-to-face interview. It is reported to take approximately 15 – 20 minutes to administer (Finch et al. 2002). A recent study by Jenkinson and colleagues (2013) validated the SIS in the UK setting and proposed a short from version with index score in order to create a less burdensome measure. The SIS (3.0), along with guides for administration and scoring the SIS are available via the internet at *www2.kumc.edu/coa*.

Advantages

The Stroke Impact Scale is intended to assess multiple domains of stroke recovery without administering multiple tests (Duncan et al. 2000). This may represent a decrease in patient burden and increased feasibility for researchers. German and Portuguese (Brazilian) versions of the SIS have been developed and evaluated (Carod-Artal et al. 2008; Petersen et al. 2001).

Published estimates of clinical importance differences by domain may improve interpretability of the results derived from repeat assessments (Lin et al. 2009).

Limitations

The emotion domain seems to be less psychometrically acceptable than the other 7 domains (Duncan et al. 1999) and even in version 3.0, the items are reported as being limited by their simplicity – that is, able to assess difficulties within only the severely affected stroke survivor (Duncan et al. 2003). Additional research on the psychometric acceptability of this scale is required.

As for other multi-dimensional assessments of health-related quality of life, agreements between patient and proxy raters were strongest in domains evaluating observable behaviours (Duncan et al. 2002). This was also reported by Carod-Artal et al. (2009) who demonstrated the poorest levels of patient/proxy agreements in the memory, communication, emotion and social participation domains.

Although the magnitude of bias reported was small in both studies, proxy raters tended to rate patients worse than the patients themselves (Carod-Artal et al. 2009; Duncan et al. 2002), particularly in the strength, ADL and composite physical domains (Carod-Artal et al. 2009).

Summary – Stroke Impact Scale

Interpretability: No standards or normative scores are available. The scale is new and has limited information available.
Acceptability: The patient-centered nature of the scale's development may enhance its relevance to patients and assessment across multiple levels may reduce patient burden. The scale has been evaluated successfully for use by proxy respondents.
Feasibility: Simple to administer and has been tested for use as a mailed questionnaire.

Table 20.4.9.1 SIS Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	++ (TR) +++ (IC)	+	+++	+	+	varied

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.4.10 Stroke Specific Quality of Life Scale (SSQOL)

The SSQOL is a patient-centered outcome measure intended to provide an assessment of health-related quality of life specific to stroke survivors. Scale domains and items were derived from a series of focused interviews with survivors of ischemic stroke (Kelly-Hayes 2000; Williams et al. 1999).

The SSQOL is a self-report scale containing 49 items in 12 domains: mobility, energy, upper extremity function, work/productivity, mood, self-care, social roles, family roles, vision, language, thinking, and personality. Each item is rated on a 5-point Likert scale on one of 3 keyed response sets (Williams et al. 1999). Higher scores indicate better function. The SSQOL yields both domain scores and an overall SSQOL summary score. The domain scores are unweighted averages of the associated items while the summary score is an unweighted average of all twelve domain scores (Williams et al. 1999).

Advantages

The method of development used assured content validity and a patient-based measure of meaning to stroke patients (Williams et al. 1999). Danish, German and Mandarin Chinese, Turkish, and Yoruba (South-Western Nigeria) versions of the scale have been developed (Ewert & Stucki 2007; Hsueh et al. 2011; Muus et al. 2011; Muus et al. 2009; Muus & Ringsberg 2005; Muus et al. 2007). Assessments of this tool in various languages and populations has furthered the evidence for the SSQOL. A 12-item Dutch-language short form has been developed and translated into Chinese (Chen et al. 2012; Post 2010). A Chinese version of the SSQOL has been developed and validated in patients with an aneurysmal subarachnoid hemorrhage (aSAH) (Wong et al. 2012). A short-form of this Chinese version, specifically for aSAH has also since been developed (Wong et al. 2013).

Limitations

The SSQOL is a relatively new scale which requires further third-party psychometric evaluation. It has not been tested in individuals with severe stroke.
The SSQOL does not appear to exhibit good sensitivity to change over time. Scale authors reported that one half of the SSQOL domains demonstrated less than moderate effect sizes and the amount of help

response set appeared to lack responsiveness (Williams et al. 1999). More recently, Lin et al. (2010) reported SRM values for SSQOL domains ranging from -0.03 (self-care) to 0.17 (language) based on assessments conducted before and after a 3-week therapeutic intervention targeting rehabilitation of the upper extremity post stroke. The SRM for the total SSQOL score was 0.14.

Several studies have examined the use of the SSQOL with proxy respondents. For observable, physical domains between-rater agreement has been reported to be moderate to excellent; however, in areas where responses may be based more on personal judgement or opinion than observation (e.g. psychological and social domains) the association between patient and proxy responses has been weaker (Muus et al. 2009; Williams et al. 2006; 2000). It is recommended that information obtained from proxy respondents be treated as supplementary rather than substantive and that use of proxy be restricted to individuals either living with or in daily contact with the patient (Lynn Snow et al. 2005; Muus et al. 2009).

Summary – Stroke-Specific Quality of Life Scale

Interpretability: There are no standardized or normative values available for comparison.
Acceptability: Its patient-centered development may increase its relevance to the patient's it is intended to assess.
Feasibility: No training necessary for administration. The SSQOL is a self-report questionnaire.

Table 20.4.10.1 SSQOL Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+	+++ (IC)	+	++	+	+	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

20.5 Conclusions and Recommendations

A careful review of the important measurement qualities obtainable from the published literature on stroke rehabilitation outcome measures produced the following main conclusions:

1. There appears to be adequate information available with which to evaluate the reliability and validity of commonly used measures.

2. Approaches taken to examine (and report) the measurement qualities of these instruments are inconsistent (especially with regard to validity).

3. Far less information is available on the responsiveness of measures, compared with reliability and validity (see Tables 20.42, 20.43 & 20.44 which present summaries of measures in each ICF category).

4. Of the three levels for classification from the ICF, the Participation category seems to be the most problematic with respect to: (a) lack of consensus on the range of domains required for measurement; (b) much greater emphasis on health-related quality of life, relative to subjective quality of life in general; (c) the inclusion of a mixture of measurements from all three ICF categories.

5. The literature offers very little specific guidance on how to ensure that the selection of an outcome measure is appropriate to a specific clinical purpose or research question. We found it impossible to

evaluate measures using this criterion. The relationship between the concepts of appropriateness and validity are not explained in a manner that would facilitate the selection of an outcome measure in stroke rehabilitation.

Clearly, there is no single form of rehabilitation that will be effective for all of the important features of a stroke-related condition, from the perspectives of all stakeholders. Therefore, one should be careful not to assume that strong evidence for intervention in a particular area necessarily implies that this intervention is likely to produce favourable outcomes in all domains that matter, for all those concerned. Based upon the conclusions from our review, we offer the following advice to the reader on how to enhance the clinical meaningfulness of the findings from the SREBR:

1. Wherever possible, try to interpret the strength of evidence for a particular form of stroke rehabilitation within the context of a theory, conceptual framework, or model for understanding the relationship between therapy and outcome. This will help you decide the forms, standards, and timeframes for reliability, validity, and responsiveness that are most appropriate to your clinical interests.

2. Consider what stakeholder values (e.g., patient, caregiver, practitioner), and balance of perspectives, are most important to you in interpreting the strength of evidence. You should be most concerned with interpreting the evidence from studies that have used reliable, valid, and responsive measures from these perspectives.

3. Examine carefully the nature and scope of outcome measurement used in reporting the strength of evidence for your area of interest in stroke rehabilitation. There is diversity in nature and scope of measures used within each of the 3 ICF categories, and a lack of consensus on what are the most important indicators of successful rehabilitation outcome in each domain.

20.5.1 Evaluation Summaries by ICF Category

Tables 20.5.1.1, 20.5.1.2 and 20.5.1.3 present a summary of the evaluation undertaken for measures in each ICF category.

Table 20.5.1.1 Evaluation Summary – Body Structure/Impairment Outcome Measures

Outcome Measure	Reliability		Validity		Responsiveness		
	Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
Beck Depression Inventory	+++	+++(TR) +++(IC)	+++	+++	+	+	n/a
Behavioral Inattention Test	+	+++(TR) +++ (IO) + (IC)	+++	+++	n/a	n/a	n/a
Canadian Neurological Scale	+	++(IO) +++(IC)	++	+++	+	+	n/a
Clock Drawing Test	++	+++(TR) ++ (IO)	+++	++	n/a	n/a	n/a
Frenchay Aphasia Screening Test	+	+++ (TR) +++ (IO)	+	+++	n/a	n/a	n/a
Fugl-Meyer Assessment	+++	+++(TR) +++(IO) ++ (IC-balance)	+++	+++ (problems balance & sensation	++	++ ++ (UE) +(sensation)	+(sensation)

				sections)			
General Health Questionnaire – 28	+	+++ (IC)	+++	+++	n/a	n/a	n/a
Geriatric Depression Scale	+++	+++(TR) +++(IC)	+++	+++	n/a	n/a	n/a
Hospital Anxiety and Depression Scale	+++	+++(TR) ++ (IO) ++ (IC)	+++	++	+	+	+++
Line Bisection Test	+	+++ (TR)	++	++	n/a	n/a	n/a
Mini Mental State Examination	+++	+++(TR) ++ (IO) ++ (IC)	+++	++	n/a	n/a	n/a
Modified Ashworth Scale	+++	++(TR) ++(IO)	+	++	+	++	n/a
Montreal Cognitive Assessment	+	+++(TR) ++ (IO)	++	+++	n/a	n/a	n/a
Motor-free Visual Perception Test	+	+++(TR) +++(IC)	++	++	n/a	n/a	n/a
National Institutes of Health Stroke Scale	++	++(TR) ++(IO) + (IC)	+++	+++	+	+	+
Orpington Prognostic Scale	+	+++(TR) +++(IO)	++	++	n/a	n/a	n/a
Stroke Rehabilitation Assessment of Movement	++	+++(TR) +++(IC)	++	+++	+	+++	+

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

Table 20.5.1.2 Evaluation Summary – Activity/Disability Outcome Measures

Outcome Measure	Reliability		Validity		Responsiveness		
	Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
Action Research Arm Test	++	+++(TR) +++(IO)	++	+++	++	+++	+
Barthel Index	+++	+++(TR) +++ (IO) +++ (IC)	+++	+++	+++	++	varied
Berg Balance Scale	++	+++(TR) +++(IO) +++(IC)	+++	+++	+++	+++	varied
Box and Block Test	++	+++(TR) +++(IO)	++	+++	++	++	n/a
Chedoke Arm and Hand Activity Inventory	+	+++(TR) ++ (IO)	+	+++	+	++	n/a
Chedoke-McMaster Stroke Assessment Scale	+	+++(TR) +++ (IO) +++ (IC)	+	+++	+	+++	n/a
Clinical Outcomes Variables Scale	+	+++(TR) +++ (IO) ++ (IC)	++	+++	++	++	++

Functional Ambulation Categories	+	+ (TR) +++(IO)	++	+++	+	+++	+
Functional Independence Measure	+++	+++(TR) +++ (IO) +++ (IC)	+++	++	+++	++	++
Frenchay Activities Index	+++	++ (TR) ++ (IO) +++ (IC)	+++	+++	+	++	+++
Modified Rankin Handicap Scale	++	++(TR) ++ (IO)	++	+++	+	++	+
Rivermead Motor Assessment	+	++ (TR) + (IO) +++ (IC)	++	++	+	++	Possible floor effect
Six-Minute Walk Test	++	+++(TR) +++(IO)	+++	+++	++	++	n/a
Motor Assessment Scale	++	+++(TR) +++ (IO)	+++	++	+	+	+
Nine Hole Peg Test	++	+++(TR) +++(IO)	++	+++	+	+	+
Rivermead Mobility Inventory	+++	+++(TR) +++ (IO) +++ (IC)	+++	+++	+++	+++	varied
Timed "Up & Go"	++	+++(TR) +++ (IO)	+++	+++	+	++	+ (floor, pts unable to complete)
Wolf Motor Function Test	+	+++ (TR) +++ (IO) ++ (IC)	+	+++	++	++	+

NOTE*: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

Table 20.5.1.3 Evaluation Summary – Participation/Handicap Outcome Measures

Outcome Measure	Reliability		Validity		Responsiveness		
	Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
Canadian Occupational Performance Measure	++	++ (TR)	++	++	+	+	n/a
EQ-5D	+	++ (TR) ++ (IO)	+++	++	+	++	varied
LIFE-H	++	++ (TR) ++(IO) ++ (IC)	+	++	n/a	n/a	n/a
London Handicap Scale	+	+++(TR)	+	+++	+	++	+
Medical Outcomes Study Short Form 36	+++	+++ (TR) +++ (IC)	+++	+++	++	+++ *(Note: 1 study reported ES)*	+++ (total score – floor/ceiling) ++ (individual domains – floor) + / ++ (individual domains - ceiling)
Nottingham Health Profile	+++	+++ (TR) +++ (IC)	+++	+++	+	n/a	+ (ceiling) ++ (floor)

Reintegration to Normal Living Index	+	+++(TR) +++ (IO)	+	++	+	++	n/a
Sickness Impact Profile (stroke-adapted version)	+	++ (IC)	++	++	+	++	n/a
Stroke Impact Scale	+	++ (TR) +++ (IC)	++	+++	++	+	varied
Stroke-Specific Quality of Life Scale	+	+++ (IC)	+	++	+	+	n/a

***NOTE**: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results)*

References

Aben, I., Verhey, F., Lousberg, R., Lodder, J., & Honig, A. (2002). Validity of the beck depression inventory, hospital anxiety and depression scale, SCL-90, and hamilton depression rating scale as screening instruments for depression in stroke patients. *Psychosomatics., 43*(5), 386-393.

Adams, S. A., Ashburn, A., Pickering, R. M., & Taylor, D. (1997a). The scalability of the Rivermead Motor Assessment in acute stroke patients. *Clinical Rehabilitation, 11*(1), 42-51.

Adams, S. A., Pickering, R. M., Ashburn, A., & Lincoln, N. B. (1997b). The scalability of the Rivermead Motor Assessment in nonacute stroke patients. *Clin Rehabil., 11*(1), 52-59.

Adunsky, A., Fleissig, Y., Levenkrohn, S., Arad, M., & Noy, S. (2002). Clock drawing task, mini-mental state examination and cognitive-functional independence measure: relation to functional outcome of stroke patients. *Arch Gerontol Geriatr., 35*(2), 153-160.

Agrell, B., & Dehlin, O. (2000). Mini mental state examination in geriatric stroke patients. Validity, differences between subgroups of patients, and relationships to somatic and mental variables. *Aging (Milano.). 12*(6), 439-444.

Ahmed, S., Mayo, N. E., Higgins, J., Salbach, N. M., Finch, L., & Wood-Dauphinee, S. L. (2003). The Stroke Rehabilitation Assessment of Movement (STREAM): a comparison with other measures used to evaluate effects of stroke and rehabilitation. *Phys Ther., 83*(7), 617-630.

Al-Khawaja, I., Wade, D. T., & Collin, C. F. (1996). Bedside screening for aphasia: a comparison of two methods. *J Neurol., 243*(2), 201-204.

Almeida, O. P., & Almeida, S. A. (1999). Short versions of the geriatric depression scale: a study of their validity for the diagnosis of a major depressive episode according to ICD-10 and DSM-IV. *Int J Geriatr Psychiatry., 14*(10), 858-865.

American Thoracic Society. (2002). ATS statement: guidelines for the six-minute walk test. *Am J Respir.Crit Care Med., 166*(1), 111-117.

Andersen, H. S., Sestoft, D., Lillebaek, T., Gabrielsen, G., & Hemmingsen, R. (2002). Validity of the General Health Questionnaire (GHQ-28) in a prison population: data from a randomized sample of prisoners on remand. *Int J Law Psychiatry., 25*(6), 573-580.

Anderson, R. T., Aaronson, N. K., & Wilkin, D. (1993). Critical review of the international assessments of health-related quality of life. *Qual.Life Res., 2*(6), 369-395.

Andersson, A. G., Kamwendo, K., Seiger, A., & Appelros, P. (2006). How to identify potential fallers in a stroke unit: validity indexes of 4 test methods. *J Rehabil Med., 38*(3), 186-191.

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil., 81*(12 Suppl 2), S15-S20.

Andresen, E. M., Gravitt, G. W., Aydelotte, M. E., & Podgorski, C. A. (1999). Limitations of the SF-36 in a sample of nursing home residents. *Age Ageing., 28*(6), 562-566.

Andresen, E. M., & Meyers, A. R. (2000). Health-related quality of life outcomes measures. *Arch Phys Med Rehabil., 81*(12 Suppl 2), S30-S45.

Anemaet, W. K. (2002). Using standardized measures to meet the challenge of stroke assessment. *Topics in geriatric rehabilitation, 18*(2), 47-62.

Ansari, N. N., Naghdi, S., Hasson, S., Mousakhani, A., Nouriyan, A., & Omidvar, Z. (2009). Inter-rater reliability of the Modified Modified Ashworth Scale as a clinical tool in measurements of post-stroke elbow flexor spasticity. *NeuroRehabilitation., 24*(3), 225-229.

Ansari, N. N., Naghdi, S., Moammeri, H., & Jalaie, S. (2006). Ashworth Scales are unreliable for the assessment of muscle spasticity. *Physiother Theory Pract., 22*(3), 119-125.

Ansari, N. N., Naghdi, S., Younesian, P., & Shayeghan, M. (2008). Inter- and intrarater reliability of the Modified Modified Ashworth Scale in patients with knee extensor poststroke spasticity. *Physiother Theory Pract., 24*(3), 205-213.

Appelros, P. (2007). Characteristics of the Frenchay Activities Index one year after a stroke: a population-based study. *Disabil Rehabil., 29*(10), 785-790.

Ashworth, B. (1964). Preliminary trial of carisoprodal in multiple sclerosis. *Practitioner, 192*, 540-542.

Aylard, P. R., Gooding, J. H., McKenna, P. J., & Snaith, R. P. (1987). A validation study of three anxiety and depression self-assessment scales. *J Psychosom.Res., 31*(2), 261-268.

Azouvi, P., Samuel, C., Louis-Dreyfus, A., Bernati, T., Bartolomeo, P., Beis, J. M., Chokron, S., Leclercq, M., Marchal, F., Martin, Y., De, M. G., Olivier, S., Perennou, D., Pradat-Diehl, P., Prairial, C., Rode, G., Sieroff, E., Wiart, L., & Rousseaux, M. (2002). Sensitivity of clinical and behavioural tests of spatial neglect after right hemisphere stroke. *J Neurol Neurosurg Psychiatry., 73*(2), 160-166.

Banks, J. L., & Marotta, C. A. (2007). Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. *Stroke., 38*(3), 1091-1096.

Barr, J. T., Schumacher, G. E., Freeman, S., LeMoine, M., Bakst, A. W., & Jones, P. W. (2000). American translation, modification, and validation of the St. George's Respiratory Questionnaire. *Clin Ther., 22*(9), 1121-1145.

Barreca, S., Gowland, C. K., Stratford, P., Huijbregts, M., Griffiths, J., Torresin, W., Dunkley, M., Miller, P., & Masters, L. (2004). Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Top Stroke Rehabil., 11*(4), 31-42.

Barreca, S. R., Stratford, P. W., Lambert, C. L., Masters, L. M., & Streiner, D. L. (2005). Test-retest reliability, validity, and sensitivity of the Chedoke arm and hand activity inventory: a new measure of upper-limb function for survivors of stroke. *Arch Phys Med Rehabil., 86*(8), 1616-1622.

Barreca, S. R., Stratford, P. W., Masters, L. M., Lambert, C. L., & Griffiths, J. (2006). Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. *Phys Ther., 86*(2), 245-253.

Barreca, S. R., Stratford, P. W., Masters, L. M., Lambert, C. L., Griffiths, J., & McBay, C. (2006). Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. *Physiotherapy Canada, 58*(2), 148-156.

Barton, G. R., Sach, T. H., Avery, A. J., Jenkinson, C., Doherty, M., Whynes, D. K., & Muir, K. R. (2008). A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. *Health Econ., 17*(7), 815-832.

Beck, A., Steer, R. A., & Brown, G. K. (1996). Beck Depression Inventory-II (BDI-II). The Psychological Corporation. *San Antonio, TX, 1996*.

Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice. A rapid technic. *Postgrad.Med., 52*(6), 81-85.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical psychology review, 8*(1), 77-100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Arch Gen.Psychiatry., 4:561-71.*, 561-571.

Beck, C. T., & Gable, R. K. (2000). Postpartum Depression Screening Scale: development and psychometric testing. *Nursing Research, 49*(5), 272-282.

Beninato, M., Gill-Body, K. M., Salles, S., Stark, P. C., Black-Schaffer, R. M., & Stein, J. (2006). Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Arch Phys Med Rehabil., 87*(1), 32-39.

Benjamin, S., Decalmer, P., & Haran, D. (1982). Community screening for mental illness: a validity study of the General Health Questionnaire. *Br J Psychiatry., 140:174-80.*, 174-180.

Berg, K. (1989). Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada, 41*(6), 304-311.

Berg, K., Wood-Dauphinee, S., & Williams, J. I. (1995). The Balance Scale: reliability assessment with elderly residents and patients with an acute stroke. *Scand J Rehabil Med., 27*(1), 27-36.

Berg, K. O., Wood-Dauphinee, S. L., Williams, J. I., & Maki, B. (1992). Measuring balance in the elderly: validation of an instrument. *Can J Public Health., 83 Suppl 2:S7-11.*, S7-11.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Med Care., 19*(8), 787-805.

Bergner, M., Bobbitt, R. A., Kressel, S., Pollard, W. E., Gilson, B. S., & Morris, J. R. (1976). The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. *Int J Health Serv., 6*(3), 393-415.

Bjelland, I., Dahl, A. A., Haug, T. T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression Scale. An updated literature review. *J Psychosom.Res., 52*(2), 69-77.

Blackburn, M., van, V. P., & Mockett, S. P. (2002). Reliability of measurements obtained with the modified Ashworth scale in the lower extremities of people with stroke. *Phys Ther., 82*(1), 25-34.

Blake, H., McKinney, M., Treece, K., Lee, E., & Lincoln, N. B. (2002). An evaluation of screening measures for cognitive impairment after stroke. *Age Ageing., 31*(6), 451-456.

Bleecker, M. L., Bolla-Wilson, K., Kawas, C., & Agnew, J. (1988). Age-specific norms for the Mini-Mental State Exam. *Neurology., 38*(10), 1565-1568.

Bodiam, C. (1999). The use of the Canadian Occupational Performance Measure for the assessment of outcome on a neurorehabilitation unit. *British Journal of Occupational Therapy, 62*, 123-126.

Bogard, K., Wolf, S., Zhang, Q., Thompson, P., Morris, D., & Nichols-Larsen, D. (2009). Can the Wolf Motor Function Test be streamlined? *Neurorehabilitation and Neural Repair, 23*(5), 422-428.

Bohannon, R. W., & Smith, M. B. (1987). Interrater reliability of a modified Ashworth scale of muscle spasticity. *Phys Ther., 67*(2), 206-207.

Bour, A., Rasquin, S., Boreas, A., Limburg, M., & Verhey, F. (2010). How predictive is the MMSE for cognitive performance after stroke? *J Neurol., 257*(4), 630-637.

Bouska, M. K., E. (1982). Manual for the application of the motor-free visual perception test to the adult population.

Bowling, A. (1997). *Measuring Heath: A review of quality of life measurement scales* (2 ed.). Philadelphia, PA: Open University Press.

Brazier, J., Jones, N., & Kind, P. (1993). Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Qual.Life Res., 2*(3), 169-180.

Brazier, J. E., Harper, R., Jones, N. M., O'Cathain, A., Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ., 305*(6846), 160-164.

Brazier, J. E., Walters, S. J., Nicholl, J. P., & Kohler, B. (1996). Using the SF-36 and Euroqol on an elderly population. *Qual.Life Res., 5*(2), 195-204.

Brenner, M. H., Curbow, B., & Legro, M. W. (1995). The proximal-distal continuum of multiple health outcome measures: the case of cataract surgery. *Med Care., 33*(4 Suppl), AS236-AS244.

Brock, K. A., Goldie, P. A., & Greenwood, K. M. (2002). Evaluating the effectiveness of stroke rehabilitation: choosing a discriminative measure. *Arch Phys Med Rehabil., 83*(1), 92-99.

Brooks, R. (1996). EuroQol: the current state of play. *Health Policy., 37*(1), 53-72.

Brott, T., Adams, H. P., Jr., Olinger, C. P., Marler, J. R., Barsan, W. G., Biller, J., Spilker, J., Holleran, R., Eberle, R., Hertzberg, V., & (1989). Measurements of acute cerebral infarction: a clinical examination scale. *Stroke., 20*(7), 864-870.

Brown, G. T., Rodger, S., & Davis, A. (2003). Motor-Free Visual Perception Test-Revised: an Overview and Critique. *The British Journal of Occupational Therapy, 66*(4), 159-167.

Buck, D., Jacoby, A., Massey, A., & Ford, G. (2000). Evaluation of measures used to assess quality of life after stroke. *Stroke., 31*(8), 2004-2010.

Burke, W. J., Houston, M. J., Boust, S. J., & Roccaforte, W. H. (1989). Use of the Geriatric Depression Scale in dementia of the Alzheimer type. *Journal of the American Geriatrics Society*.

Burvill, P. W., & Knuiman, M. W. (1983). Which version of the General Health Questionnaire should be used in community studies? *Aust.N Z.J Psychiatry., 17*(3), 237-242.

Bushnell, C. D., Johnston, D. C., & Goldstein, L. B. (2001). Retrospective assessment of initial stroke severity: comparison of the NIH Stroke Scale and the Canadian Neurological Scale. *Stroke., 32*(3), 656-660.

Butland, R. J., Pang, J., Gross, E. R., Woodcock, A. A., & Geddes, D. M. (1982). Two-, six-, and 12-minute walking tests in respiratory disease. *Br Med J (Clin Res Ed). 284*(6329), 1607-1608.

Cabral, D. L., Laurentino, G. E. C., Damascena, C. G., Faria, C. D. C. M., Melo, P. G., & Teixeira-Salmela, L. F. (2012). Comparisons of the Nottingham Health Profile and the SF-36 Health Survey for the assessment of quality of life in individuals with chronic stroke. *Revista Brasileira de Fisioterapia, 16*(4), July/August.

Callahan, C. D., Young, P. L., & Barisa, M. T. (2005). Using the SF-36 for longitudinal outcomes measurement in rehabilitation. *Rehabilitation Psychology, 50*(1), 65.

Can, S. S., Gencay-Can, A., & Gunendi, Z. (2012). Validity and reliability of the clock drawing test as a screening tool for cognitive impairment in patients with fibromyalgia. *Compr.Psychiatry., 53*(1), 81-86.

Cannon, B. J., Thaler, T., & Roos, S. (2002). Oral versus written administration of the Geriatric Depression Scale. *Aging & Mental Health, 6*(4), 418-422.

Cardol, M., Brandsma, J. W., de Groot, I. J., Van den Bos, G. A., de Haan, R. J., & de Jong, B. A. (1999). Handicap questionnaires: what do they assess? *Disabil Rehabil., 21*(3), 97-105.

Carey, J. R., Kimberley, T. J., Lewis, S. M., Auerbach, E. J., Dorsey, L., Rundquist, P., & Ugurbil, K. (2002). Analysis of fMRI and finger tracking training in subjects with chronic stroke. *Brain., 125*(Pt 4), 773-788.

Carod-Artal, F. J., Coral, L. F., Trizotto, D. S., & Moreira, C. M. (2008). The stroke impact scale 3.0: evaluation of acceptability, reliability, and validity of the Brazilian version. *Stroke., 39*(9), 2477-2484.

Carod-Artal, F. J., Ferreira, C. L., Stieven, T. D., & Menezes, M. C. (2009). Self- and proxy-report agreement on the Stroke Impact Scale. *Stroke., 40*(10), 3308-3314.

Carr, J. H., Shepherd, R. B., Nordholm, L., & Lynne, D. (1985). Investigation of a new motor assessment scale for stroke patients. *Phys Ther., 65*(2), 175-180.

Carroll, D. (1965). A Quantitative test of Upper Extremity Function. *J Chronic.Dis., 18,* 479-491.

Carswell, A., McColl, M. A., Baptiste, S., Law, M., Polatajko, H., & Pollock, N. (2004). The Canadian Occupational Performance Measure: a research and clinical literature review. *Can J Occup Ther., 71*(4), 210-222.

Carter, B. S., Buckley, D., Ferraro, R., Rordorf, G., & Ogilvy, C. S. (2000). Factors associated with reintegration to normal living after subarachnoid hemorrhage. *Neurosurgery., 46*(6), 1326-1333.

Cavanagh, S. J., Hogan, K., Gordon, V., & Fairfax, J. (2000). Stroke-specific FIM models in an urban population. *J Neurosci Nurs., 32*(1), 17-20.

Chae, J., Labatia, I., & Yang, G. (2003). Upper limb motor function in hemiparesis: concurrent validity of the Arm Motor Ability test. *Am J Phys Med Rehabil., 82*(1), 1-8.

Chan, C. C., & Lee, T. (1997). Validity of the Canadian occupational performance measure. *Occupational Therapy International, 4*(3), 231-249.

Chanubol, R., Wongphaet, P., Ot, N. C., Chira-Adisai, W., Kuptniratsaikul, P., & Jitpraphai, C. (2012). Correlation between the action research arm test and the box and block test of upper extremity function in stroke patients. *J Med Assoc Thai., 95*(4), 590-597.

Charoenpong, L. C., P. Limsriwilai, J. Chotikanuchit, S. Yamkaew, N. Lirathpong, N. Komoltri, C. Poungvarin, N. Nilanont, Y. (2013). Reliability and validity of the Canadian neurological scale, Thai version.

J Med Assoc Thai., 96(S2), S54-59.

Chen, H. F., Lin, K. C., Wu, C. Y., & Chen, C. L. (2012). Rasch validation and predictive validity of the action research arm test in patients receiving stroke rehabilitation. *Arch.Phys.Med Rehabil, 93*(6), 1039-1045.

Chen, H. F., Wu, C. Y., Lin, K. C., Chen, H. C., Chen, C. P., & Chen, C. K. (2012). Rasch validation of the streamlined Wolf Motor Function Test in people with chronic stroke and subacute stroke. *Phys Ther., 92*(8), 1017-1026.

Chen, H. F., Wu, C. Y., Lin, K. C., Li, M. W., & Yu, H. W. (2012). Validity, reliability and responsiveness of a short version of the Stroke-Specific Quality of Life Scale in patients receiving rehabilitation. *J Rehabil Med., 44*(8), 629-636.

Chen, Y. H., Rodger, S., & Polatjko, H. (2002). Experiences with the COPM and client-centred practice in adult neurorehabilitation in Taiwan. *Occup Ther Int., 9*(3), 167-184.

Chou, C. Y., Chien, C. W., Hsueh, I. P., Sheu, C. F., Wang, C. H., & Hsieh, C. L. (2006). Developing a short form of the Berg Balance Scale for people with stroke. *Phys Ther., 86*(2), 195-204.

Coast, J., Peters, T. J., Richards, S. H., & Gunnell, D. J. (1998). Use of the EuroQoL among elderly acute care patients. *Qual.Life Res., 7*(1), 1-10.

Cohen, M. E., & Marino, R. J. (2000). The tools of disability outcomes research functional status measures. *Arch Phys Med Rehabil., 81*(12 Suppl 2), S21-S29.

Colarusso, R. P., & Hammill, D. D. (1972). *Motor-free visual perception test*: Academic Therapy Pub.

Cole, B., & Basmajian, J. (1994). *Physical rehabilitation outcome measures*: Canadian Physiotherapy Association in cooperation with Health and Welfare Canada and Canada Communications Group, Publishing, Supply & Services Canada.

Collen, F. M., Wade, D. T., & Bradshaw, C. M. (1990). Mobility after stroke: reliability of measures of impairment and disability. *Int Disabil Stud., 12*(1), 6-9.

Collen, F. M., Wade, D. T., Robb, G. F., & Bradshaw, C. M. (1991). The Rivermead Mobility Index: a further development of the Rivermead Motor Assessment. *Int Disabil Stud., 13*(2), 50-54.

Collin, C., & Wade, D. T. (1990). Assessing motor impairment after stroke: a pilot reliability study. *Journal of Neurology, Neurosurgery & Psychiatry, 53*(7), 576-579.

Collin, C., Wade, D. T., Davies, S., & Horne, V. (1988). The Barthel ADL Index: a reliability study. *Disability & Rehabilitation, 10*(2), 61-63.

Coons, S. J., Rao, S., Keininger, D. L., & Hays, R. D. (2000). A comparative review of generic quality-of-life instruments. *Pharmacoeconomics., 17*(1), 13-35.

Cote, R., Battista, R. N., Wolfson, C., Boucher, J., Adam, J., & Hachinski, V. (1989). The Canadian Neurological Scale: validation and reliability assessment. *Neurology., 39*(5), 638-643.

Cote, R., Hachinski, V. C., Shurvell, B. L., Norris, J. W., & Wolfson, C. (1986). The Canadian Neurological Scale: a preliminary study in acute stroke. *Stroke., 17*(4), 731-737.

Cromwell, F. (1976). Occupational therapists manual for basic skill assessment: Primary prevocational evaluation (Fair Oaks Printing, Altadena, CA).

Crow, J. L., & BC, H.-v. d. W. (2008). Hierarchical properties of the motor function sections of the Fugl-Meyer assessment scale for people after stroke: a retrospective study. *Phys Ther., 88*(12), 1554-1567.

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA., 269*(18), 2386-2391.

Cunha, I. T., Lim, P. A., Henson, H., Monga, T., Qureshy, H., & Protas, E. J. (2002). Performance-based gait tests for acute stroke patients. *Am J Phys Med Rehabil., 81*(11), 848-856.

Cup, E. H., Scholte op Reimer, W. J., Thijssen, M. C., & van Kuyk-Minis, M. A. (2003). Reliability and validity of the Canadian Occupational Performance Measure in stroke patients. *Clin Rehabil., 17*(4), 402-409.

Curb, J. D., Ceria-Ulep, C. D., Rodriguez, B. L., Grove, J., Guralnik, J., Willcox, B. J., Donlon, T. A., Masaki, K. H., & Chen, R. (2006). Performance-based measures of physical function for high-function populations. *J Am Geriatr Soc., 54*(5), 737-742.

Cuspineda, E., Machado, C., Aubert, E., Galan, L., Llopis, F., & Avila, Y. (2003). Predicting outcome in acute stroke: a comparison between QEEG and the Canadian Neurological Scale. *Clin Electroencephalogr., 34*(1), 1-4.

da Cunha, I. T. J., Lim, P. A., Qureshy, H., Henson, H., Monga, T., & Protas, E. J. (2002). Gait outcomes after acute stroke rehabilitation with supported treadmill ambulation training: a randomized controlled pilot study. *Arch Phys Med Rehabil., 83*(9), 1258-1265.

Daley, K., Mayo, N., & Wood-Dauphinee, S. (1999). Reliability of scores on the Stroke Rehabilitation Assessment of Movement (STREAM) measure. *Phys Ther., 79*(1), 8-19.

Dalgas, U., Severinsen, K., & Overgaard, K. (2012). Relations between 6 minute walking distance and 10 meter walking speed in patients with multiple sclerosis and stroke. *Arch Phys Med Rehabil., 93*(7), 1167-1172.

Dallmeijer, A. J., Dekker, J., Knol, D. L., Kalmijn, S., Schepers, V. P., De, G., V, Lindeman, E., Beelen, A., & Lankhorst, G. J. (2006). Dimensional structure of the SF-36 in neurological patients. *J Clin Epidemiol., 59*(5), 541-543.

Damiano, D. L., Quinlivan, J. M., Owen, B. F., Payne, P., Nelson, K. C., & Abel, M. F. (2002). What does the Ashworth scale really measure and are instrumented measures more valid and precise? *Dev.Med Child Neurol., 44*(2), 112-118.

Dancer, S., Brown, A. J., & Yanase, L. R. (2009). National Institutes of Health Stroke Scale reliable and valid in plain English. *J Neurosci Nurs., 41*(1), 2-5.

Danielsson, A., Willen, C., & Sunnerhagen, K. S. (2011). Is walking endurance associated with activity and participation late after stroke? *Disabil Rehabil., 33*(21-22), 2053-2057.

Davis, J., Kayser, J., Matlin, P., Mower, S., & Tadano, P. (1999). Clinical Analysis. Nine-hole peg tests: are they all the same? *OT Practice, 4*, 59-61.

de Bruin, A. F., de Witte, L. P., Stevens, F., & Diederiks, J. P. (1992). Sickness Impact Profile: the state of the art of a generic functional status measure. *Soc Sci Med., 35*(8), 1003-1014.

de Haan, R., Aaronson, N., Limburg, M., Hewer, R. L., & van, C. H. (1993). Measuring quality of life in stroke. *Stroke., 24*(2), 320-327.

de Haan, R., Horn, J., Limburg, M., van der Meulen, J., & Bossuyt, P. (1993). A comparison of five stroke scales with measures of disability, handicap, and quality of life. *Stroke., 24*(8), 1178-1181.

de Koning, I., van, K. F., & Koudstaal, P. J. (1998). Value of screening instruments in the diagnosis of post-stroke dementia. *Haemostasis., 28*(3-4), 158-166.

De Weerdt, W. J. G. (1985). Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnstrom-Fugl-Meyer test and the Action Research Arm test. *Physiotherapy Canada, 37*(2), 65-70.

Dedding, C., Cardol, M., Eyssen, I. C., Dekker, J., & Beelen, A. (2004). Validity of the Canadian Occupational Performance Measure: a client-centred outcome measurement. *Clin Rehabil., 18*(6), 660-667.

Demaerschalk, B. V., S. et al. (2012). Reliability of Real-Time Video Smartphone for Assessing National Institutes of Health Stroke Scale Scores in Acute Stroke Patients. *Stroke, 43*, 3271-3277.

Desrosiers, J., Bravo, G., Hebert, R., Dutil, E., & Mercier, L. (1994). Validation of the Box and Block Test as a measure of dexterity of elderly people: reliability, validity, and norms studies. *Arch Phys Med Rehabil., 75*(7), 751-755.

Desrosiers, J., Noreau, L., Robichaud, L., Fougeyrollas, P., Rochette, A., & Viscogliosi, C. (2004). Validity of the Assessment of Life Habits in older adults. *J Rehabil Med., 36*(4), 177-182.

Desrosiers, J., Noreau, L., Rochette, A., Bravo, G., & Boutin, C. (2002). Predictors of handicap situations following post-stroke rehabilitation. *Disabil Rehabil., 24*(15), 774-785.

Dewey, H. M., Donnan, G. A., Freeman, E. J., Sharples, C. M., Macdonell, R. A., McNeil, J. J., & Thrift, A. G. (1999). Interrater reliability of the National Institutes of Health Stroke Scale: rating by neurologists and nurses in a community-based stroke incidence study. *Cerebrovasc Dis., 9*(6), 323-327.

Dick, J. P., Guiloff, R. J., Stewart, A., Blackstock, J., Bielawska, C., Paul, E. A., & Marsden, C. D. (1984). Mini-mental state examination in neurological patients. *J Neurol Neurosurg Psychiatry., 47*(5), 496-499.

Dijkers, M. P., Whiteneck, G., & El-Jaroudi, R. (2000). Measures of social outcomes in disability research. *Arch Phys Med Rehabil., 81*(12 Suppl 2), S63-S80.

Dong, Y., Sharma, V. K., Chan, B. P., Venketasubramanian, N., Teoh, H. L., Seet, R. C., Tanicala, S., Chan, Y. H., & Chen, C. (2010). The Montreal Cognitive Assessment (MoCA) is superior to the Mini-Mental State Examination (MMSE) for the detection of vascular cognitive impairment after acute stroke. *J Neurol Sci., 299*(1-2), 15-18.

Dorman, P., Slattery, J., Farrell, B., Dennis, M., & Sandercock, P. (1998). Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. United Kingdom Collaborators in the International Stroke Trial. *Stroke., 29*(1), 63-68.

Dorman, P. J., Dennis, M., & Sandercock, P. (1999). How do scores on the EuroQol relate to scores on the SF-36 after stroke? *Stroke., 30*(10), 2146-2151.

Dorman, P. J., Slattery, J., Farrell, B., Dennis, M. S., & Sandercock, P. A. (1997). A randomised comparison of the EuroQol and Short Form-36 after stroke. United Kingdom collaborators in the International Stroke Trial. *BMJ., 315*(7106), 461.

Dorman, P. J., Waddell, F., Slattery, J., Dennis, M., & Sandercock, P. (1997). Is the EuroQol a valid measure of health-related quality of life after stroke? *Stroke., 28*(10), 1876-1882.

Dubuc, N., Haley, S., Ni, P., Kooyoomjian, J., & Jette, A. (2004). Function and disability in late life: comparison of the Late-Life Function and Disability Instrument to the Short-Form-36 and the London Handicap Scale. *Disabil Rehabil., 26*(6), 362-370.

Duffy, L., Gajree, S., Langhorne, P., Stott, D. J., & Quinn, T. J. (2013). Reliability (inter-rater agreement) of the Barthel Index for assessment of stroke survivors: systematic review and meta-analysis. *Stroke, 44*(2), 462-468.

Duncan, P. W., Bode, R. K., Min, L. S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Arch Phys Med Rehabil., 84*(7), 950-963.

Duncan, P. W., Goldstein, L. B., Horner, R. D., Landsman, P. B., Samsa, G. P., & Matchar, D. B. (1994). Similar motor recovery of upper and lower extremities after stroke. *Stroke., 25*(6), 1181-1188.

Duncan, P. W., Jorgensen, H. S., & Wade, D. T. (2000). Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. *Stroke., 31*(6), 1429-1438.

Duncan, P. W., Lai, S. M., Bode, R. K., Perera, S., & DeRosa, J. (2003). Stroke Impact Scale-16: A brief assessment of physical function. *Neurology., 60*(2), 291-296.

Duncan, P. W., Lai, S. M., Tyler, D., Perera, S., Reker, D. M., & Studenski, S. (2002). Evaluation of proxy responses to the Stroke Impact Scale. *Stroke., 33*(11), 2593-2599.

Duncan, P. W., Reker, D. M., Horner, R. D., Samsa, G. P., Hoenig, H., LaClair, B. J., & Dudley, T. K. (2002). Performance of a mail-administered version of a stroke-specific outcome measure, the Stroke Impact Scale. *Clin Rehabil., 16*(5), 493-505.

Duncan, P. W., Samsa, G. P., Weinberger, M., Goldstein, L. B., Bonito, A., Witter, D. M., Enarson, C., & Matchar, D. (1997). Health status of individuals with mild stroke. *Stroke., 28*(4), 740-745.

Duncan, P. W., Wallace, D., Lai, S. M., Johnson, D., Embretson, S., & Laster, L. J. (1999). The stroke impact scale version 2.0. Evaluation of reliability, validity, and sensitivity to change. *Stroke., 30*(10), 2131-2140.

Ebrahim, S., Barer, D., & Nouri, F. (1986). Use of the Nottingham Health Profile with patients after a stroke. *J Epidemiol.Community Health., 40*(2), 166-169.

Edwards, D. F., Lang, C. E., Wagner, J. M., Birkenmeier, R., & Dromerick, A. W. (2012). An evaluation of the Wolf Motor Function Test in motor trials early after stroke. *Arch.Phys.Med Rehabil, 93*(4), 660-668.

Ehreke, L., Luck, T., Luppa, M., Kanig, H. H., Villringer, A., & Riedel-Heller, S. G. (2011). Clock Drawing Test- screening utility for mild cognitive impairment according to different scoring systems: results of the Leipzig Longitudinal Study of the Aged (LEILA 75+). *International Psychogeriatrics, 23*(10), 1592.

Enderby, P., & Crow, E. (1996). Frenchay Aphasia Screening Test: validity and comparability. *Disabil Rehabil., 18*(5), 238-240.

Enderby, P., Wood, V. A., Wade, D. T., & Hewer, R. L. (1987a). Aphasia after stroke: a detailed study of recovery in the first 3 months. *Int Rehabil Med., 8*(4), 162-165.

Enderby, P. M., Wood, V. A., Wade, D. T., & Hewer, R. L. (1987b). The Frenchay Aphasia Screening Test: a short, simple test for aphasia appropriate for non-specialists. *Int Rehabil Med., 8*(4), 166-170.

Endres, M., Nyary, I., Banhidi, M., & Deak, G. (1990). Stroke rehabilitation: a method and evaluation. *Int J Rehabil Res., 13*(3), 225-236.

Eng, J. J., Dawson, A. S., & Chu, K. S. (2004). Submaximal exercise in persons with stroke: test-retest reliability and concurrent validity with maximal oxygen consumption. *Arch Phys Med Rehabil., 85*(1), 113-118.

Eng, J. J., Rowe, S. J., & McLaren, L. M. (2002). Mobility status during inpatient rehabilitation: a comparison of patients with stroke and traumatic brain injury. *Arch Phys Med Rehabil., 83*(4), 483-490.

Enright, P. L., & Sherrill, D. L. (1998). Reference equations for the six-minute walk in healthy adults. *Am J Respir.Crit Care Med., 158*(5 Pt 1), 1384-1387.

Essink-Bot, M. L., Krabbe, P. F., Bonsel, G. J., & Aaronson, N. K. (1997). An empirical comparison of four generic health status measures: the Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Medical Care, 35*(5), 522-537.

Ewert, T., & Stucki, G. (2007). Validity of the SS-QOL in Germany and in survivors of hemorrhagic or ischemic stroke. *Neurorehabil Neural Repair., 21*(2), 161-168.

Eyssen, I. C., Beelen, A., Dedding, C., Cardol, M., & Dekker, J. (2005). The reproducibility of the Canadian Occupational Performance Measure. *Clin Rehabil., 19*(8), 888-894.

Faria, C. D., Teixeira-Salmela, L. F., Neto, M. G., & Rodrigues-de-Paula, F. (2012). Performance-based tests in subjects with stroke: outcome scores, reliability and measurement errors. *Clin.Rehabil, 26*(5), 460-469.

Ferber, S., & Karnath, H. O. (2001). How to assess spatial neglect--line bisection or cancellation tasks? *J Clin Exp.Neuropsychol., 23*(5), 599-607.

Finch, E., Brooks, D., P.W., S., & N.E., M. (2002). *Physical Rehabilitations Outcome Measures. A guide to enhanced clinical decision-making* (2 ed.). Toronto: Canadian Physiotherapy Association.

Fiorelli, M., Alperovitch, A., Argentino, C., Sacchetti, M. L., Toni, D., Sette, G., Cavalletti, C., Gori, M. C., & Fieschi, C. (1995). Prediction of long-term outcome in the early hours following acute ischemic stroke. Italian Acute Stroke Study Group. *Arch Neurol., 52*(3), 250-255.

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess., 2*(14), i-74.

Flansbjer, U. B., Holmback, A. M., Downham, D., Patten, C., & Lexell, J. (2005). Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med., 37*(2), 75-82.

Flint, A. J., & Rifat, S. L. (2002). Factor structure of the hospital anxiety and depression scale in older patients with major depression. *Int J Geriatr Psychiatry., 17*(2), 117-123.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr.Res., 12*(3), 189-198.

Fordell, H., Bodin, K., Bucht, G., & Malm, J. (2011). A virtual reality test battery for assessment and screening of spatial neglect. *Acta Neurol Scand., 123*(3), 167-174.

Forlander, D. A., & Bohannon, R. W. (1999). Rivermead Mobility Index: a brief review of research to date. *Clin Rehabil., 13*(2), 97-100.

Fougeyrollas, P., Noreau, L., Bergeron, H., Cloutier, R., Dion, S. A., & St-Michel, G. (1998). Social consequences of long term impairments and disabilities: conceptual approach and assessment of handicap. *Int J Rehabil Res., 21*(2), 127-141.

Franchignoni, F., Tesio, L., Benevolo, E., & Ottonello, M. (2003). Psychometric properties of the Rivermead Mobility Index in Italian stroke rehabilitation inpatients. *Clin Rehabil., 17*(3), 273-282.

Fritz, S. L., Blanton, S., Uswatte, G., Taub, E., & Wolf, S. L. (2009). Minimal detectable change scores for the Wolf Motor Function Test. *Neurorehabilitation and Neural Repair, 23*(7), 662-667.

Fu, T. S., Wu, C. Y., Lin, K. C., Hsieh, C. J., Liu, J. S., Wang, T. N., & Ou-Yang, P. (2012). Psychometric comparison of the shortened Fugl-Meyer Assessment and the streamlined Wolf Motor Function Test in stroke rehabilitation. *Clin.Rehabil, 26*(11), 1043-1047.

Fugl-Meyer, A. R. (1980). Post-stroke hemiplegia assessment of physical properties. *Scand J Rehabil Med Suppl., 7:85-93.*, 85-93.

Fugl-Meyer, A. R., Jaasko, L., Leyman, I., Olsson, S., & Steglind, S. (1975). The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J Rehabil Med., 7*(1), 13-31.

Fulk, G. D., Echternach, J. L., Nof, L., & O'Sullivan, S. (2008). Clinometric properties of the six-minute walk test in individuals undergoing rehabilitation poststroke. *Physiother Theory Pract., 24*(3), 195-204.

Furukawa, T. A., Goldberg, D. P., Rabe-Hesketh, S., & Ustun, T. B. (2001). Stratum-specific likelihood ratios of two versions of the general health questionnaire. *Psychol Med., 31*(3), 519-529.

Gagnon, C., Mathieu, J., & Noreau, L. (2006). Measurement of participation in myotonic dystrophy: reliability of the LIFE-H. *Neuromuscul.Disord., 16*(4), 262-268.

Ghotbi, N., Nakhostin, A. N., Naghdi, S., & Hasson, S. (2011). Measurement of lower-limb muscle spasticity: intrarater reliability of Modified Modified Ashworth Scale. *J Rehabil Res Dev., 48*(1), 83-88.

Gibson, L., MacLennan, W. J., Gray, C., & Pentland, B. (1991). Evaluation of a comprehensive assessment battery for stroke patients. *Int J Rehabil Res., 14*(2), 93-100.

Gladstone, D. J., Danells, C. J., & Black, S. E. (2002). The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabilitation and Neural Repair, 16*(3), 232-240.

Godefroy, O., Fickl, A., Roussel, M., Auribault, C., Bugnicourt, J. M., Lamy, C., Canaple, S., & Petitnicolas, G. (2011). Is the Montreal Cognitive Assessment superior to the Mini-Mental State Examination to detect poststroke cognitive impairment? A study with neuropsychological evaluation. *Stroke., 42*(6), 1712-1716.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol Med., 27*(1), 191-197.

Goldberg, D. P., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychol Med., 9*(1), 139-145.

Goldberg, D. P., Oldehinkel, T., & Ormel, J. (1998). Why GHQ threshold varies from one place to another. *Psychol Med., 28*(4), 915-920.

Goldstein, L. B., Bertels, C., & Davis, J. N. (1989). Interrater reliability of the NIH stroke scale. *Arch Neurol., 46*(6), 660-662.

Goldstein, L. B., & Samsa, G. P. (1997). Reliability of the National Institutes of Health Stroke Scale. Extension to non-neurologists in the context of a clinical trial. *Stroke., 28*(2), 307-310.

Golomb, B. A., Vickrey, B. G., & Hays, R. D. (2001). A review of health-related quality-of-life measures in stroke. *Pharmacoeconomics., 19*(2), 155-185.

Goodchild, M. E., & Duncan-Jones, P. (1985). Chronicity and the General Health Questionnaire. *Br J Psychiatry., 146:55-61.*, 55-61.

Gosman-Hedstrom, G. S., E. (2000). Parallel reliability of the Functional Independence Measure and the Barthel ADL Index. *Disabil Rehabil, 22*, 702-715.

Gowland, C. (1995). *Chedoke-McMaster Stroke Assessment: development, validation and administration manual*: Chedoke-McMaster Hospitals and McMaster University.

Gowland, C., Stratford, P., Ward, M., Moreland, J., Torresin, W., Van, H. S., Sanford, J., Barreca, S., Vanspall, B., & Plews, N. (1993). Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. *Stroke., 24*(1), 58-63.

Grace, J., Nadler, J. D., White, D. A., Guilmette, T. J., Giuliano, A. J., Monsch, A. U., & Snow, M. G. (1995). Folstein vs modified Mini-Mental State Examination in geriatric stroke. Stability, validity, and screening utility. *Arch.Neurol., 52*(5), 477-484.

Granger, C. V., Cotter, A. C., Hamilton, B. B., & Fiedler, R. C. (1993). Functional assessment scales: a study of persons after stroke. *Arch.Phys.Med Rehabil., 74*(2), 133-138.

Granger, C. V., Sherwood, C. C., & Greer, D. S. (1977). Functional status measures in a comprehensive stroke care program. *Archives of physical medicine and rehabilitation, 58*(12), 555.

Green, J., & Young, J. (2001). A test-retest reliability study of the Barthel Index, the Rivermead Mobility Index, the Nottingham Extended Activities of Daily Living Scale and the Frenchay Activities Index in stroke patients. *Disability & Rehabilitation, 23*(15), 670-676.

Gregson, J. M., Leathley, M., Moore, A. P., Sharma, A. K., Smith, T. L., & Watkins, C. L. (1999). Reliability of the Tone Assessment Scale and the modified Ashworth scale as clinical tools for assessing poststroke spasticity. *Arch.Phys.Med Rehabil., 80*(9), 1013-1016.

Gregson, J. M., Leathley, M. J., Moore, A. P., Smith, T. L., Sharma, A. K., & Watkins, C. L. (2000). Reliability of measurements of muscle tone and muscle power in stroke patients. *Age Ageing., 29*(3), 223-228.

Group, E. (1990). EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. *Health Policy., 16*(3), 199-208.

Guyatt, G. H., Pugsley, S. O., Sullivan, M. J., Thompson, P. J., Berman, L., Jones, N. L., Fallen, E. L., & Taylor, D. W. (1984). Effect of encouragement on walking test performance. *Thorax., 39*(11), 818-822.

Guyatt, G. H., Sullivan, M. J., Thompson, P. J., Fallen, E. L., Pugsley, S. O., Taylor, D. W., & Berman, L. B. (1985). The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Can.Med Assoc.J., 132*(8), 919-923.

Guyatt, G. H., Townsend, M., Keller, J., Singer, J., & Nogradi, S. (1991). Measuring functional status in chronic lung disease: conclusions from a randomized control trial. *Respir.Med., 85 Suppl B:17-21; discussion 33-7.*, 17-20.

Haas, B. M., Bergstrom, E., Jamous, A., & Bennie, A. (1996). The inter rater reliability of the original and of the modified Ashworth scale for the assessment of spasticity in patients with spinal cord injury. *Spinal Cord., 34*(9), 560-564.

Hachisuka, K., Ogata, H., Ohkuma, H., Tanaka, S., & Dozono, K. (1997). Test-retest and inter-method reliability of the self-rating Barthel Index. *Clin Rehabil., 11*(1), 28-35.

Hajek, V. E., Gagnon, S., & Ruderman, J. E. (1997). Cognitive and functional assessments of stroke patients: an analysis of their relation. *Arch.Phys.Med Rehabil., 78*(12), 1331-1337.

Halligan, P. W., Cockburn, J., & Wilson, B. A. (1991). The behavioural assessment of visual neglect. *Neuropsychological rehabilitation, 1*(1), 5-32.

Han, C. W., Yajima, Y., Nakajima, K., Lee, E. J., Meguro, M., & Kohzuki, M. (2006). Construct validity of the Frenchay Activities Index for community-dwelling elderly in Japan. *Tohoku J Exp.Med., 210*(2), 99-107.

Harvey, R. F., & Jellinek, H. M. (1981). Functional performance assessment: a program approach. *Archives of physical medicine and rehabilitation, 62*(9), 456-460.

Harwood, R. H., & Ebrahim, S. (2000a). A comparison of the responsiveness of the Nottingham extended activities of daily living scale, London handicap scale and SF-36. *Disabil.Rehabil., %20;22*(17), 786-793.

Harwood, R.H., & Ebrahim, S. (2000b). Measuring the outcomes of day hospital attendance: a comparison of the Barthel Index and London Handicap Scale. *Clin Rehabil., 14*(5), 527-531.

Harwood, R. H., Gompertz, P., & Ebrahim, S. (1994a). Handicap one year after a stroke: validity of a new scale. *J Neurol Neurosurg.Psychiatry., 57*(7), 825-829.

Harwood, R. H., Rogers, A., Dickinson, E., & Ebrahim, S. (1994b). Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. *Qual.Health Care., 3*(1), 11-16.

Hayes, V., Morris, J., Wolfe, C., & Morgan, M. (1995). The SF-36 health survey questionnaire: is it suitable for use with older adults? *Age Ageing., 24*(2), 120-125.

Healey, A. K., Kneebone, I. I., Carroll, M., & Anderson, S. J. (2008). A preliminary investigation of the reliability and validity of the Brief Assessment Schedule Depression Cards and the Beck Depression Inventory-Fast Screen to screen for depression in older stroke survivors. *Int.J Geriatr.Psychiatry., 23*(5), 531-536.

Heinemann, A. W., Harvey, R. L., McGuire, J. R., Ingberman, D., Lovell, L., Semik, P., & Roth, E. J. (1997). Measurement properties of the NIH Stroke Scale during acute rehabilitation. *Stroke., 28*(6), 1174-1180.

Heinik, J., Solomesh, I., & Berkman, P. (2004). Correlation between the CAMCOG, the MMSE, and three clock drawing tests in a specialized outpatient psychogeriatric service. *Arch.Gerontol.Geriatr., 38*(1), 77-84.

Heller, A., Wade, D. T., Wood, V. A., Sunderland, A., Hewer, R. L., & Ward, E. (1987). Arm function after stroke: measurement and recovery over the first three months. *J Neurol Neurosurg.Psychiatry., 50*(6), 714-719.

Helvik, A. S., Engedal, K., Skancke, R. H., & Selbaek, G. (2011). A psychometric evaluation of the Hospital Anxiety and Depression Scale for the medically hospitalized elderly. *Nord.J Psychiatry., 65*(5), 338-344.

Herrmann, C. (1997). International experiences with the Hospital Anxiety and Depression Scale--a review of validation data and clinical results. *J Psychosom.Res., 42*(1), 17-41.

Hershkovitz, A., & Brill, S. (2006). Get up and go--home. *Aging Clin Exp.Res., 18*(4), 301-306.

Hesse, S., Bertelt, C., Schaffrin, A., Malezic, M., & Mauritz, K. H. (1994). Restoration of gait in nonambulatory hemiparetic patients by treadmill training with partial body-weight support. *Arch.Phys.Med Rehabil., 75*(10), 1087-1093.

Hiengkaew, V., Jitaree, K., & Chaiyawat, P. (2012). Minimal detectable changes of the Berg Balance Scale, Fugl-Meyer Assessment Scale, Timed "Up & Go" Test, gait speeds, and 2-minute walk test in individuals with chronic stroke with different degrees of ankle plantarflexor tone. *Arch.Phys.Med Rehabil., 93*(7), 1201-1208.

Higgins, J., Mayo, N. E., Desrosiers, J., Salbach, N. M., & Ahmed, S. (2005). Upper-limb function and recovery in the acute phase poststroke. *J Rehabil Res.Dev., 42*(1), 65-76.

Hobart, J. C., Lamping, D. L., Freeman, J. A., Langdon, D. W., McLellan, D. L., Greenwood, R. J., & Thompson, A. J. (2001). Evidence-based measurement: which disability scale for neurologic rehabilitation? *Neurology., 57*(4), 639-644.

Hobart, J. C., & Thompson, A. J. (2001). The five item Barthel index. *J Neurol Neurosurg.Psychiatry., 71*(2), 225-230.

Hobart, J. C., Williams, L. S., Moran, K., & Thompson, A. J. (2002). Quality of life measurement after stroke: uses and abuses of the SF-36. *Stroke., 33*(5), 1348-1356.

Hoffmann, T., Worrall, L., Eames, S., & Ryan, A. (2010). Measuring outcomes in people who have had a stroke and their carers: can the telephone be used? *Top.Stroke Rehabil., 17*(2), 119-127.

Holbrook, M., & Skilbeck, C. E. (1983). An activities index for use with stroke patients. *Age Ageing., 12*(2), 166-170.

Holden, M. K., Gill, K. M., & Magliozzi, M. R. (1986). Gait assessment for neurologically impaired patients Standards for outcome assessment. *Physical therapy, 66*(10), 1530-1539.

Holden, M. K., Gill, K. M., Magliozzi, M. R., Nathan, J., & Piehl-Baker, L. (1984). Clinical gait assessment in the neurologically impaired. Reliability and meaningfulness. *Phys.Ther., 64*(1), 35-40.

Hou, W. H., Shih, C. L., Chou, Y. T., Sheu, C. F., Lin, J. H., Wu, H. C., Hsueh, I. P., & Hsieh, C. L. (2012). Development of a computerized adaptive testing system of the Fugl-Meyer motor scale in stroke patients. *Arch Phys Med Rehabil., 93*(6), 1014-1020

House, A., Dennis, M., Mogridge, L., Warlow, C., Hawton, K., & Jones, L. (1991). Mood disorders in the year after first stroke. *Br.J Psychiatry., 158:83-92.*, 83-92.

Hsieh, C. L., Hsueh, I. P., & Mao, H. F. (2000). Validity and responsiveness of the rivermead mobility index in stroke patients. *Scand.J Rehabil Med., 32*(3), 140-142.

Hsieh, Y. W., Hsueh, I. P., Chou, Y. T., Sheu, C. F., Hsieh, C. L., & Kwakkel, G. (2007). Development and validation of a short form of the Fugl-Meyer motor scale in patients with stroke. *Stroke., 38*(11), 3052-3054.

Hsieh, Y. W., Wang, C. H., Sheu, C. F., Hsueh, I. P., & Hsieh, C. L. (2008). Estimating the minimal clinically important difference of the Stroke Rehabilitation Assessment of Movement measure. *Neurorehabil.Neural Repair., 22*(6), 723-727.

Hsieh, Y. W., Wang, C. H., Wu, S. C., Chen, P. C., Sheu, C. F., & Hsieh, C. L. (2007). Establishing the minimal clinically important difference of the Barthel Index in stroke patients. *Neurorehabil.Neural Repair., 21*(3), 233-238.

Hsieh, Y. W., Wu, C. Y., Lin, K. C., Chang, Y. F., Chen, C. L., & Liu, J. S. (2009). Responsiveness and validity of three outcome measures of motor function after stroke rehabilitation. *Stroke., 40*(4), 1386-1391.

Hsueh, I. P., & Hsieh, C. L. (2002). Responsiveness of two upper extremity function instruments for stroke inpatients receiving rehabilitation. *Clin Rehabil., 16*(6), 617-624.

Hsueh, I. P., Hsu, M. J., Sheu, C. F., Lee, S., Hsieh, C. L., & Lin, J. H. (2008). Psychometric comparisons of 2 versions of the Fugl-Meyer Motor Scale and 2 versions of the Stroke Rehabilitation Assessment of Movement. *Neurorehabil.Neural Repair., 22*(6), 737-744.

Hsueh, I. P., Jeng, J. S., Lee, Y., Sheu, C. F., & Hsieh, C. L. (2011). Construct validity of the stroke-specific quality of life questionnaire in ischemic stroke patients. *Arch.Phys.Med Rehabil., 92*(7), 1113-1118.

Hsueh, I. P., Lee, M. M., & Hsieh, C. L. (2001). Psychometric characteristics of the Barthel activities of daily living index in stroke patients. *J Formos.Med Assoc., 100*(8), 526-532.

Hsueh, I. P., Lee, M. M., & Hsieh, C. L. (2002). The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? *Clin Rehabil., 16*(4), 382-388.

Hsueh, I. P., Lin, J. H., Jeng, J. S., & Hsieh, C. L. (2002). Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. *J Neurol Neurosurg.Psychiatry., 73*(2), 188-190.

Hsueh, I. P., Wang, C. H., Sheu, C. F., & Hsieh, C. L. (2003). Comparison of psychometric properties of three mobility measures for patients with stroke. *Stroke., 34*(7), 1741-1745.

Hsueh, I. P., Wang, W. C., Wang, C. H., Sheu, C. F., Lo, S. K., Lin, J. H., & Hsieh, C. L. (2006). A simplified stroke rehabilitation assessment of movement instrument. *Phys.Ther., 86*(7), 936-943.

Huijbregts, M. (1996). The physiotherapy clinical outcome variables (COVS) reliability testing videotape. *Physiotherapy Canada, 48*, 285.

Hung, C. I., Liu, C. Y., Wang, S. J., Yao, Y. C., & Yang, C. H. (2012). The cut-off points of the Depression and Somatic Symptoms Scale and the Hospital Anxiety and Depression Scale in detecting non-full remission and a current major depressive episode. *Int.J Psychiatry Clin Pract., 16*(1), 33-40.

Hunt, S. M., McEwen, J., & McKenna, S. P. (1984). Perceived health: age and sex comparisons in a community. *J Epidemiol.Community Health., 38*(2), 156-160.

Hunt, S. M., McEwen, J., & McKenna, S. P. (1985). Measuring health status: a new tool for clinicians and epidemiologists. *J R.Coll.Gen.Pract., 35*(273), 185-188.

Hunt, S. M., McKenna, S. P., & McEwen, J. (1989). The Nottingham health profile user's manual. *Manchester: Galen Research and Consultancy*.

Hunt, S. M., McKenna, S. P., McEwen, J., Backett, E. M., Williams, J., & Papp, E. (1980). A quantitative approach to perceived health status: a validation study. *J Epidemiol.Community Health., 34*(4), 281-286.

Hunt, S. M., McKenna, S. P., McEwen, J., Williams, J., & Papp, E. (1981). The Nottingham Health Profile: subjective health status and medical consultations. *Soc.Sci.Med A., 15*(3 Pt 1), 221-229.

ICIDH. (1990). International Classification of Impairments, Disabilities and Handicaps (ICIDH).

Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., & Busschbach, J. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res, 22*(7), 1717-1727.

Janssen, P. M., Visser, N. A., Dorhout Mees, S. M., Klijn, C. J., Algra, A., & Rinkel, G. J. (2010). Comparison of telephone and face-to-face assessment of the modified Rankin Scale. *Cerebrovasc.Dis., 29*(2), 137-139.

Jehkonen, M., Ahonen, J. P., Dastidar, P., Koivisto, A. M., Laippala, P., Vilkki, J., & Molnar, G. (2000). Visual neglect as a predictor of functional outcome one year after stroke. *Acta Neurol Scand., 101*(3), 195-201.

Jenkinson, C. (1991). Why are we weighting? A critical examination of the use of item weights in a health status measure. *Soc.Sci.Med., 32*(12), 1413-1416.

Jenkinson, C., Fitzpatrick, R., Crocker, H., & Peters, M. (2013). The stroke impact scale: Validation in a UK setting and development of a SIS short form and SIS index. *Stroke, 44*(9), September.

Jenkinson, C., Mant, J., Carter, J., Wade, D., & Winner, S. (2000). The London handicap scale: a re-evaluation of its validity using standard scoring and simple summation. *J Neurol Neurosurg.Psychiatry., 68*(3), 365-367.

Johnston, M., Pollard, B., & Hennessey, P. (2000). Construct validation of the hospital anxiety and depression scale with clinical populations. *J Psychosom.Res., 48*(6), 579-584.

Jorgensen, H. S., Nakayama, H., Raaschou, H. O., Vive-Larsen, J., Stoier, M., & Olsen, T. S. (1995). Outcome and time course of recovery in stroke. Part II: Time course of recovery. The Copenhagen Stroke Study. *Arch.Phys.Med Rehabil., 76*(5), 406-412.

Josephson, S. A., Hills, N. K., & Johnston, S. C. (2006). NIH Stroke Scale reliability in ratings from a large sample of clinicians. *Cerebrovasc.Dis., 22*(5-6), 389-395.

Juneja, G., Czyrny, J. J., & Linn, R. T. (1998). Admission balance and outcomes of patients admitted for acute inpatient rehabilitation. *Am.J Phys.Med Rehabil., 77*(5), 388-393.

Jutai, J. W., & Teasell, R. W. (2003). The necessity and limitations of evidence-based practice in stroke rehabilitation. *Top.Stroke Rehabil., 10*(1), 71-78.

Kafonek, S., Ettinger, W. H., Roca, R., Kittner, S., Taylor, N., etal. (1989). Instruments for screening for depression and dementia in a long term care. *Journal of the American Geriatrics Society, 37*, 29-34.

Kalra, L., & Crome, P. (1993). The role of prognostic scores in targeting stroke rehabilitation in elderly patients. *J Am.Geriatr.Soc., 41*(4), 396-400.

Kalra, L., Dale, P., & Crome, P. (1994). Evaluation of a clinical score for prognostic stratification of elderly stroke patients. *Age Ageing., 23*(6), 492-498.

Kaya, T., Karatepe, A. G., Gunaydin, R., Koc, A., & Altundal, E. U. (2011). Inter-rater reliability of the Modified Ashworth Scale and modified Modified Ashworth Scale in assessing poststroke elbow flexor spasticity. *Int.J Rehabil Res., 34*(1), 59-64.

Kellor, M., Frost, J., Silberberg, N., Iversen, I., & Cummings, R. (1971). Hand strength and dexterity. *Am.J Occup.Ther., 25*(2), 77-83.

Kelly-Hayes, M. (2000). Stroke outcomes measurement: new developments. *Journal of Rehabilitation Outcomes Measurement, 4*(4), 57-63.

Kerr, D. M., Fulton, R. L., & Lees, K. R. (2012). Seven-day NIHSS is a sensitive outcome measure for exploratory clinical trials in acute stroke: evidence from the Virtual International Stroke Trials Archive. *Stroke., 43*(5), 1401-1403.

Khan, A., Chien, C. W., & Brauer, S. G. (2013). Rasch-based scoring offered more precision in differentiating patient groups in measuring upper limb function. *J Clin.Epidemiol., 66*(6), 681-687.

Kidd, D., Stewart, G., Baldry, J., Johnson, J., Rossiter, D., Petruckevitch, A., & Thompson, A. J. (1995). The Functional Independence Measure: a comparative validity and reliability study. *Disabil.Rehabil., 17*(1), 10-14.

Kilic, C., Rezaki, M., Rezaki, B., Kaplan, I., Ozgen, G., Sagduyu, A., & Ozturk, M. O. (1997). General Health Questionnaire (GHQ12 & GHQ28): psychometric properties and factor structure of the scales in a Turkish primary care sample. *Soc.Psychiatry Psychiatr.Epidemiol., 32*(6), 327-331.

Kimberley, T. L., SM. Auerbach, EJ. Dorsey, LL., & Lojovich, J. C., JR. (2004). Electrical stimulation driving functional improvements and cortical changes in subjects with stroke. *Experimental Brain Research, 154*, 450-460.

Koeter, M. W. (1992). Validity of the GHQ and SCL anxiety and depression scales: a comparative study. *J Affect.Disord., 24*(4), 271-279.

Koh, C. L., Hsueh, I. P., Wang, W. C., Sheu, C. F., Yu, T. Y., Wang, C. H., & Hsieh, C. L. (2006). Validation of the action research arm test using item response theory in patients after stroke. *J Rehabil Med., 38*(6), 375-380.

Kornetti, D. L., Fritz, S. L., Chiu, Y. P., Light, K. E., & Velozo, C. A. (2004). Rating scale analysis of the Berg Balance Scale. *Arch.Phys.Med Rehabil., 85*(7), 1128-1135.

Kurtais, Y., Kucukdeveci, A., Elhan, A., Yilmaz, A., Kalli, T., Tur, B. S., & Tennant, A. (2009). Psychometric properties of the Rivermead Motor Assessment: its utility in stroke. *Journal of Rehabilitation Medicine, 41*(13), 1055-1061.

Kutlay, S., Kucukdeveci, A. A., Yanik, B., Elhan, A., Oztuna, D., & Tennant, A. (2011). The interval scaling properties of the London Handicap Scale: an example from the adaptation of the scale for use in Turkey. *Clin Rehabil., 25*(3), 248-255.

Kwakkel, G., Veerbeek, J. M., Harmeling-van der Wel, B., van Wegen, E., & Kollen, B. J. (2011). Diagnostic Accuracy of the Barthel Index for Measuring Activities of Daily Living Outcome After Ischemic Hemispheric Stroke Does Early Poststroke Timing of Assessment Matter? *Stroke, 42*(2), 342-346.

Kwon, S., Hartzema, A. G., Duncan, P. W., & Min-Lai, S. (2004). Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. *Stroke., 35*(4), 918-923.

La, P. F., Caselli, S., Susassi, S., Cavallini, P., Tennant, A., & Franceschini, M. (2012). Is the Berg Balance Scale an internally valid and reliable measure of balance across different etiologies in neurorehabilitation? A revisited Rasch analysis study. *Arch.Phys.Med Rehabil, 93*(7), 1209-1216.

Lai, S. M., & Duncan, P. W. (2001). Stroke recovery profile and the Modified Rankin assessment. *Neuroepidemiology., 20*(1), 26-30.

Lai, S. M., Duncan, P. W., & Keighley, J. (1998). Prediction of functional outcome after stroke: comparison of the Orpington Prognostic Scale and the NIH Stroke Scale. *Stroke., 29*(9), 1838-1842.

Lam, S. P., Tsui, E., Chan, K. S., Lam, C. L., & So, H. P. (2006). The validity and reliability of the functional impairment checklist (FIC) in the evaluation of functional consequences of severe acute respiratory distress syndrome (SARS). *Qual.Life Res., 15*(2), 217-231.

Lang, C. E., Edwards, D. F., Birkenmeier, R. L., & Dromerick, A. W. (2008). Estimating minimal clinically important differences of upper-extremity measures early after stroke. *Arch.Phys.Med Rehabil., 89*(9), 1693-1700.

Lannin, N. (2004). Reliability, validity and factor structure of the upper limb subscale of the Motor Assessment Scale (UL-MAS) in adults following stroke. *Disabil.Rehabil., 26*(2), 109-116.

Law, M., Baptiste, S., McColl, M., Opzoomer, A., Polatajko, H., & Pollock, N. (1990). The Canadian occupational performance measure: an outcome measure for occupational therapy. *Can.J Occup.Ther., 57*(2), 82-87.

Law, M., & MacDermid, J. (2002). Introduction to evidence-based practice. *Evidence-based rehabilitation: A guide to practice*, 3-12.

Law, M., Polatajko, H., Pollock, N., McColl, M. A., Carswell, A., & Baptiste, S. (1994). Pilot testing of the Canadian Occupational Performance Measure: clinical and measurement issues. *Can.J Occup.Ther., 61*(4), 191-197.

Lee, K. S., Kim, E. A., Hong, C. H., Lee, D. W., Oh, B. H., & Cheong, H. K. (2008). Clock drawing test in mild cognitive impairment: quantitative analysis of four scoring methods and qualitative analysis. *Dement.Geriatr.Cogn Disord., 26*(6), 483-489.

Lepage, C., Noreau, L., Bernard, P. M., & Fougeyrollas, P. (1998). Profile of handicap situations in children with cerebral palsy. *Scand J Rehabil Med., 30*(4), 263-272.

Li, K. Y., Lin, K. C., Wang, T. N., Wu, C. Y., Huang, Y. H., & Ouyang, P. (2012). Ability of three motor measures to predict functional outcomes reported by stroke patients after rehabilitation. *NeuroRehabilitation., 30*(4), 267-275.

Liaw, L. J., Hsieh, C. L., Lo, S. K., Chen, H. M., Lee, S., & Lin, J. H. (2008). The relative and absolute reliability of two balance performance measures in chronic stroke patients. *Disabil Rehabil., 30*(9), 656-661.

Lin, J. H., Hsu, M. J., Sheu, C. F., Wu, T. S., Lin, R. T., Chen, C. H., & Hsieh, C. L. (2009). Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. *Phys Ther., 89*(8), 840-850.

Lin, J. H., Hsueh, I. P., Sheu, C. F., & Hsieh, C. L. (2004). Psychometric properties of the sensory scale of the Fugl-Meyer Assessment in stroke patients. *Clin Rehabil., 18*(4), 391-397.

Lin, K. C., Chen, H. F., Wu, C. Y., Yu, T. Y., & Ouyang, P. (2012). Multidimensional Rasch validation of the Frenchay Activities Index in stroke patients receiving rehabilitation. *J Rehabil Med, 44*(1), 58-64.

Lin, K. C., Fu, T., Wu, C. Y., Hsieh, Y. W., Chen, C. L., & Lee, P. C. (2010). Psychometric comparisons of the Stroke Impact Scale 3.0 and Stroke-Specific Quality of Life Scale. *Qual.Life Res., 19*(3), 435-443.

Lin, K. C., Hsieh, Y. W., Wu, C. Y., Chen, C. L., Jang, Y., & Liu, J. S. (2009). Minimal detectable change and clinically important difference of the Wolf Motor Function Test in stroke patients. *Neurorehabil Neural Repair., 23*(5), 429-434.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Arch Phys Med Rehabil., 75*(2), 127-132.

Lincoln, N., & Leadbitter, D. (1979). Assessment of motor function in stroke patients. *Physiotherapy., 65*(2), 48-51.

Lincoln, N. B., Nicholl, C. R., Flannaghan, T., Leonard, M., & Van der Gucht, E. (2003). The validity of questionnaire measures for assessing depression after stroke. *Clin Rehabil., 17*(8), 840-846.

Liu, J., Drutz, C., Kumar, R., McVicar, L., Weinberger, R., Brooks, D., & Salbach, N. M. (2008). Use of the six-minute walk test poststroke: is there a practice effect? *Arch Phys Med Rehabil., 89*(9), 1686-1692.

Lo, R., Harwood, R., Woo, J., Yeung, F., & Ebrahim, S. (2001). Cross-cultural validation of the London Handicap Scale in Hong Kong Chinese. *Clin Rehabil., 15*(2), 177-185.

Lo, R. S., Kwok, T. C., Cheng, J. O., Yang, H., Yuan, H. J., Harwood, R., & Woo, J. (2007). Cross-cultural validation of the London Handicap Scale and comparison of handicap perception between Chinese and UK populations. *Age Ageing., 36*(5), 544-548.

Lobo, A., Perez-Echeverria, M. J., Jimenez-Aznarez, A., & Sancho, M. A. (1988). Emotional disturbances in endocrine patients. Validity of the scaled version of the General Health Questionnaire (GHQ-28). *Br J Psychiatry., 152:807-12.*, 807-812.

Loewen, S. C., & Anderson, B. A. (1990). Predictors of stroke outcome using objective measurement scales. *Stroke., 21*(1), 78-81.

Lopes, M. A. L., Ferreira, H. P., Carvalho, J. C., Cardoso, L. A., & Andra C. (2007). Screening tests are not enough to detect hemineglect. *Arquivos de neuro-psiquiatria, 65*(4B), 1192-1195.

Lord, S. E., McPherson, K., McNaughton, H. K., Rochester, L., & Weatherall, M. (2004). Community ambulation after stroke: how important and obtainable is it and what measures appear predictive? *Archives of physical medicine and rehabilitation, 85*(2), 234-239.

Lorentz, W., Scanlan, J., & Borson, S. (2002). Brief screening tests for dementia. *Journal of Psychiatry, 47*, 723-733.

Lourenco, R. A., Ribeiro-Filho, S. T., Moreira, I. F., Paradela, E. M., & Miranda, A. S. (2008). The Clock Drawing Test: performance among elderly with low educational level. *Rev Bras.Psiquiatr., 30*(4), 309-315.

Low Choy, N., Kuys, S., Richards, M., & Isles, R. (2002). Measurement of functional ability following traumatic brain injury using the Clinical Outcomes Variable Scale: a reliability study. *Aust.J Physiother., 48*(1), 35-39.

Lu, W. S., Chen, C. C., Huang, S. L., & Hsieh, C. L. (2012). Smallest real difference of 2 instrumental activities of daily living measures in patients with chronic stroke. *Arch.Phys.Med Rehabil, 93*(6), 1097-1100.

Luis, C. A., Keegan, A. P., & Mullan, M. (2009). Cross validation of the Montreal Cognitive Assessment in community dwelling older adults residing in the Southeastern US. *International Journal of Geriatric Psychiatry, 24*(2), 197-201.

Lundgren, N. A., & Tennant, A. (2011). Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. *J Rehabil Med., 43*(10), 884-891.

Lyden, P., Lu, M., Jackson, C., Marler, J., Kothari, R., Brott, T., & Zivin, J. (1999). Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. NINDS tPA Stroke Trial Investigators. *Stroke., 30*(11), 2347-2354.

Lyden, P., Raman, R., Liu, L., Emr, M., Warren, M., & Marler, J. (2009). National Institutes of Health Stroke Scale certification is reliable across multiple venues. *Stroke., 40*(7), 2507-2511.

Lyden, P. L., M. Levine, SR. Brott, TG. Broderick, J. (2001). A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. *Stroke, 32*, 1310-1317.

Lykouras, L., Adrachta, D., Kalfakis, N., Oulis, P., Voulgari, A., Christodoulou, G. N., Papageorgiou, C., & Stefanis, C. (1996). GHQ-28 as an aid to detect mental disorders in neurological inpatients. *Acta Psychiatr.Scand., 93*(3), 212-216.

Lyle, R. C. (1981). A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res., 4*(4), 483-492.

Lynn Snow, A., Cook, K. F., Lin, P. S., Morgan, R. O., & Magaziner, J. (2005). Proxies and other external raters: methodological considerations. *Health Serv.Res., 40*(5 Pt 2), 1676-1693.

MacKenzie, G., Gould, L., Ireland, S., LeBlanc, K., & Sahlas, D. (2011). Detecting cognitive impairment in clients with mild stroke or transient ischemic attack attending a stroke prevention clinic. *Can.J Neurosci Nurs., 33*(1), 47-50.

MacNeill, S. E., & Lichtenberg, P. A. (2000). The MacNeill-Lichtenberg Decision Tree: a unique method of triaging mental health problems in older medical rehabilitation patients. *Arch Phys Med Rehabil., 81*(5), 618-622.

Mahoney, F. I. (1965). Functional evaluation: the Barthel index. *Maryland state medical journal, 14*, 61-65.

Mallinson, S. (2002). Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Soc Sci Med., 54*(1), 11-20.

Malouin, F., Pichard, L., Bonneau, C., Durand, A., & Corriveau, D. (1994). Evaluating motor recovery early after stroke: comparison of the Fugl-Meyer Assessment and the Motor Assessment Scale. *Arch Phys Med Rehabil., 75*(11), 1206-1212.

Mao, H. F., Hsueh, I. P., Tang, P. F., Sheu, C. F., & Hsieh, C. L. (2002). Analysis and comparison of the psychometric properties of three balance measures for stroke patients. *Stroke., 33*(4), 1022-1027.

Marinus, J., Leentjens, A. F., Visser, M., Stiggelbout, A. M., & van Hilten, J. J. (2002). Evaluation of the hospital anxiety and depression scale in patients with Parkinson's disease. *Clin Neuropharmacol., 25*(6), 318-324.

Mathias, S., Nayak, U. S., & Isaacs, B. (1986). Balance in elderly patients: the "get-up and go" test. *Arch Phys Med Rehabil., 67*(6), 387-389.

Mathiowetz, V., Volland, G., Kashman, N., & Weber, K. (1985). Adult norms for the Box and Block Test of manual dexterity. *Am J Occup Ther., 39*(6), 386-391.

Mayo, E. (1999). Disablement following stroke. *Disability & Rehabilitation, 21*(5-6), 258-268.

Mayo, N. E., Wood-Dauphinee, S., Cote, R., Durcan, L., & Carlton, J. (2002). Activity, participation, and quality of life 6 months poststroke. *Arch Phys Med Rehabil., 83*(8), 1035-1042.

Mayo, N. E., Wood-Dauphinee, S., Cote, R., Gayton, D., Carlton, J., Buttery, J., & Tamblyn, R. (2000). There's no place like home : an evaluation of early supported discharge for stroke. *Stroke., 31*(5), 1016-1023.

Mazer, B. L., Korner-Bitensky, N. A., & Sofer, S. (1998). Predicting ability to drive after stroke. *Arch Phys Med Rehabil., 79*(7), 743-750.

McColl, M. A., Paterson, M., Davies, D., Doubt, L., & Law, M. (2000). Validity and community utility of the Canadian Occupational Performance Measure. *Can.J Occup Ther., 67*(1), 22-30.

McDowell, I., & Newell, C. (1996). Measuring health: A guide to rating scales and questionnaires: OUP.

McEwan, S. (1995). *Performance-based correlates of health-related quality of life in community-dwelling persons with stroke.*, McGill University, Quebec, Canada.

McGavin, C. R., Gupta, S. P., & McHardy, G. J. (1976). Twelve-minute walking test for assessing disability in chronic bronchitis. *Br Med J., 1*(6013), 822-823.

McGinnis, G. E., Seward, M. L., DeJong, G., & Osberg, J. S. (1986). Program evaluation of physical medicine and rehabilitation departments using self-report Barthel. *Arch Phys Med Rehabil., 67*(2), 123-125.

McGivney, S. A., Mulvihill, M., & Taylor, B. (1994). Validating the GDS depression screen in the nursing home. *J Am Geriatr Soc., 42*(5), 490-492.

McHorney, C. A., Ware, J. E., Jr., & Raczek, A. E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care., 31*(3), 247-263.

Mehrholz, J., Wagner, K., Rutte, K., Meissner, D., & Pohl, M. (2007). Predictive validity and responsiveness of the functional ambulation category in hemiparetic patients after stroke. *Arch Phys Med Rehabil., 88*(10), 1314-1319.

Meldrum, D., Pittock, S. J., Hardiman, O., Ni, D. C., & O'Regan, M. (2004). Recovery of the upper limb post ischaemic stroke and the predictive value of the Orpington Prognostic Score. *Clin Rehabil., 18*(6), 694-702.

Menon, A., & Korner-Bitensky, N. (2004). Evaluating unilateral spatial neglect post stroke: working your way through the maze of assessment choices. *Top Stroke Rehabil., 11*(3), 41-66.

Mercier, L., Desrosiers, J., HÃ©bert, R. Ã., Rochette, A., & Dubois, M. F. (2001). Normative data for the Motor-Free Visual Perception Test-Vertical. *Physical & Occupational Therapy in Geriatrics, 19*(2), 39-50.

Meyer, B. C., Hemmen, T. M., Jackson, C. M., & Lyden, P. D. (2002). Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. *Stroke., 33*(5), 1261-1266.

Miller, K. J., Slade, A. L., Pallant, J. F., & Galea, M. P. (2010). Evaluation of the psychometric properties of the upper limb subscales of the Motor Assessment Scale using a Rasch analysis model. *J Rehabil Med., 42*(4), 315-322.

Millis, S. R., Straube, D., Iramaneerat, C., Smith, E. V., Jr., & Lyden, P. (2007). Measurement properties of the National Institutes of Health Stroke Scale for people with right- and left-hemisphere lesions: further analysis of the clomethiazole for acute stroke study-ischemic (class-I) trial. *Arch Phys Med Rehabil., 88*(3), 302-308.

Miyamoto, S., Nagaya, N., Satoh, T., Kyotani, S., Sakamaki, F., Fujita, M., Nakanishi, N., & Miyatake, K. (2000). Clinical correlates and prognostic significance of six-minute walk test in patients with primary pulmonary hypertension. Comparison with cardiopulmonary exercise testing. *Am J Respir.Crit Care Med., 161*(2 Pt 1), 487-492.

Morris, D. M., Uswatte, G., Crago, J. E., Cook, E. W., III, & Taub, E. (2001). The reliability of the wolf motor function test for assessing upper extremity function after stroke. *Arch Phys Med Rehabil., 82*(6), 750-755.

Morris, S., Morris, M. E., & Iansek, R. (2001). Reliability of measurements obtained with the Timed "Up & Go" test in people with Parkinson disease. *Phys Ther., 81*(2), 810-818.

Muir, K. W., Weir, C. J., Murray, G. D., Povey, C., & Lees, K. R. (1996). Comparison of neurological scales and scoring systems for acute stroke prognosis. *Stroke., 27*(10), 1817-1820.

Muus, I., Christensen, D., Petzold, M., Harder, I., Johnsen, S. P., Kirkevold, M., & Ringsberg, K. C. (2011). Responsiveness and sensitivity of the Stroke Specific Quality of Life Scale Danish version. *Disabil Rehabil., 33*(25-26), 2425-2433.

Muus, I., Petzold, M., & Ringsberg, K. C. (2009). Health-related quality of life after stroke: reliability of proxy responses. *Clin Nurs Res., 18*(2), 103-118.

Muus, I., & Ringsberg, K. C. (2005). Stroke Specific Quality of Life Scale: Danish adaptation and a pilot study for testing psychometric properties. *Scand J Caring Sci., 19*(2), 140-147.

Muus, I., Williams, L. S., & Ringsberg, K. C. (2007). Validation of the Stroke Specific Quality of Life Scale (SS-QOL): test of reliability and validity of the Danish version (SS-QOL-DK). *Clin Rehabil., 21*(7), 620-627.

Naghdi, S., Ansari, N. N., Mansouri, K., Olyaei, G. R., Asgari, A., & Kazemnejad, A. (2008). The correlation between Modified Ashworth Scale scores and the new index of alpha motoneurones excitability in post-stroke patients. *Electromyogr.Clin Neurophysiol., 48*(2), 109-115.

Naghdi, S., Ebrahimi, I., Asgari, A., Olyaei, G. R., Kazemnejad, A., Mansouri, K., & Ansari, N. N. (2007). A preliminary study into the criterion validity of the Modified Modified Ashworth Scale using the new measure of the alpha motoneuron excitability in spastic hemiplegia. *Electromyogr.Clin Neurophysiol., 47*(3), 187-192.

Nakamura, D. H., M. & Wilson, A. (1998). Measures of balance and fear of falling in the elderly: A review. *Physical and Occupational Therapy in Geriatrics, 15*, 17-32.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc., 53*(4), 695-699.

New, P. W., & Buchbinder, R. (2006). Critical appraisal and review of the Rankin scale and its derivatives. *Neuroepidemiology., 26*(1), 4-15.

Ng, S. S. (2011). Contribution of subjective balance confidence on functional mobility in subjects with chronic stroke. *Disabil Rehabil., 33*(23-24), 2291-2298.

Ng, S. S., & Hui-Chan, C. W. (2005). The timed up & go test: its reliability and association with lower-limb impairments and locomotor capacities in people with chronic stroke. *Arch Phys Med Rehabil., 86*(8), 1641-1647.

Nijland, R., van, W. E., Verbunt, J., van, W. R., van, K. J., & Kwakkel, G. (2010). A comparison of two validated tests for upper limb function after stroke: The Wolf Motor Function Test and the Action Research Arm Test. *J Rehabil Med., 42*(7), 694-696.

Nokleby, K., Boland, E., Bergersen, H., Schanke, A. K., Farner, L., Wagle, J., & Wyller, T. B. (2008). Screening for cognitive deficits after stroke: a comparison of three screening tools. *Clin Rehabil., 22*(12), 1095-1104.

Nordin, E., Lindelof, N., Rosendahl, E., Jensen, J., & Lundin-Olsson, L. (2008). Prognostic validity of the Timed Up-and-Go test, a modified Get-Up-and-Go test, staff's global judgement and fall history in evaluating fall risk in residential care facilities. *Age Ageing., 37*(4), 442-448.

Nordin, E., Rosendahl, E., & Lundin-Olsson, L. (2006). Timed "Up & Go" test: reliability in older people dependent in activities of daily living--focus on cognitive state. *Phys Ther., 86*(5), 646-655.

Noreau, L., & Fougeyrollas, P. (2000). Long-term consequences of spinal cord injury on social participation: the occurrence of handicap situations. *Disabil Rehabil., 22*(4), 170-180.

Nys, G. M., van Zandvoort, M. J., de Kort, P. L., Jansen, B. P., Kappelle, L. J., & de Haan, E. H. (2005). Restrictions of the Mini-Mental State Examination in acute stroke. *Arch Clin Neuropsychol., 20*(5), 623-629.

O'Mahony, P. G., Rodgers, H., Thomson, R. G., Dobson, R., & James, O. F. (1998). Is the SF-36 suitable for assessing health status of older stroke patients? *Age Ageing., 27*(1), 19-22.

O'Neill, P. A., Cheadle, B., Wyatt, R., McGuffog, J., & Fullerton, K. J. (1990). The value of the Frenchay Aphasia Screening Test in screening for dysphasia: better than the clinician? *Clinical Rehabilitation, 4*(2), 123-128.

O'Rourke, S., MacHale, S., Signorini, D., & Dennis, M. (1998). Detecting psychiatric morbidity after stroke: comparison of the GHQ and the HAD Scale. *Stroke., 29*(5), 980-985.

Oemar, M., & Janssen, B. (2013). EQ-5D-5L User Guide: Basic information on how to use the EQ-5D-5L instrument.

Ottenbacher, K. J., Hsu, Y., Granger, C. V., & Fiedler, R. C. (1996). The reliability of the functional independence measure: a quantitative review. *Arch Phys Med Rehabil., 77*(12), 1226-1232.

Oxford, G. K., Vogel, K. A., Le, V., Mitchell, A., Muniz, S., & Vollmer, M. A. (2003). Adult norms for a commercially available Nine Hole Peg Test for finger dexterity. *Am J Occup Ther., 57*(5), 570-573.

Page, S. J., Fulk, G. D., & Boyne, P. (2012). Clinically important differences for the upper-extremity Fugl-Meyer Scale in people with minimal to moderate impairment due to chronic stroke. *Phys Ther., 92*(6), 791-798.

Pais-Ribeiro, J., Silva, I., Ferreira, T., Martins, A., Meneses, R., & Baltar, M. (2007). Validation study of a Portuguese version of the Hospital Anxiety and Depression Scale. *Psychol Health Med., 12*(2), 225-235.

Pandyan, A. D., Johnson, G. R., Price, C. I., Curless, R. H., Barnes, M. P., & Rodgers, H. (1999). A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. *Clin Rehabil., 13*(5), 373-383.

Pandyan, A. D., Price, C. I., Barnes, M. P., & Johnson, G. R. (2003). A biomechanical investigation into the validity of the modified Ashworth Scale as a measure of elbow spasticity. *Clin Rehabil., 17*(3), 290-293.

Pandyan, A. D., Price, C. I., Rodgers, H., Barnes, M. P., & Johnson, G. R. (2001). Biomechanical examination of a commonly used measure of spasticity. *Clin Biomech.(Bristol., Avon.). 16*(10), 859-865.

Patrick, E., & Ada, L. (2006). The Tardieu Scale differentiates contracture from spasticity whereas the Ashworth Scale is confounded by it. *Clin Rehabil., 20*(2), 173-182.

Pedersen, P. M., Jorgensen, H. S., Nakayama, H., Raaschou, H. O., & Olsen, T. S. (1997). Comprehensive assessment of activities of daily living in stroke. The Copenhagen Stroke Study. *Arch Phys Med Rehabil., 78*(2), 161-165.

Pendlebury, S. T., Cuthbertson, F. C., Welch, S. J., Mehta, Z., & Rothwell, P. M. (2010). Underestimation of cognitive impairment by Mini-Mental State Examination versus the Montreal Cognitive Assessment in patients with transient ischemic attack and stroke: a population-based study. *Stroke., 41*(6), 1290-1293.

Pendlebury, S. T., Welch, S. J., Cuthbertson, F. C., Mariz, J., Mehta, Z., & Rothwell, P. M. (2013). Telephone assessment of cognition after transient ischemic attack and stroke: modified telephone interview of cognitive status and telephone Montreal Cognitive Assessment versus face-to-face Montreal Cognitive Assessment and neuropsychological battery. *Stroke., 44*(1), 227-229.

Perenboom, R. J., & Chorus, A. M. (2003). Measuring participation according to the International Classification of Functioning, Disability and Health (ICF). *Disabil Rehabil., 25*(11-12), 577-587.

Perera, S., Mody, S. H., Woodman, R. C., & Studenski, S. A. (2006). Meaningful change and responsiveness in common physical performance measures in older adults. *J Am Geriatr Soc., 54*(5), 743-749.

Petersen, C., Morfeld, M., & Bullinger, M. (2001). [Testing and validation of the German version of the Stroke Impact Scale]. *Fortschr.Neurol Psychiatr., 69*(6), 284-290.

Pickard, A. S., Johnson, J. A., Feeny, D. H., Shuaib, A., Carriere, K. C., & Nasser, A. M. (2004). Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index. *Stroke., 35*(2), 607-612.

Piercy, M., Carter, J., Mant, J., & Wade, D. T. (2000). Inter-rater reliability of the Frenchay activities index in patients with stroke and their careers. *Clin Rehabil., 14*(4), 433-440.

Pittock, S. J., Meldrum, D., Ni, D. C., Hardiman, O., & Moroney, J. T. (2003). The Orpington Prognostic Scale within the first 48 hours of admission as a predictor of outcome in ischemic stroke. *J Stroke Cerebrovasc Dis., 12*(4), 175-181.

Platz, T., Pinkowski, C., van, W. F., Kim, I. H., di, B. P., & Johnson, G. (2005). Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. *Clin Rehabil., 19*(4), 404-411.

Poole, J. B., PA. Torres, TA. (2005). Measuring Dexterity in Children Using the Nine-hole Peg Test. *J HAND THER, 18*, 348-351.

Poole, J. L., & Whitney, S. L. (1988). Motor assessment scale for stroke patients: concurrent validity and interrater reliability. *Arch Phys Med Rehabil., 69*(3 Pt 1), 195-197.

Poole, J. L., & Whitney, S. L. (2001). Assessments of motor function post stroke: A review. *Physical & Occupational Therapy in Geriatrics, 19*(2), 1-22.

Post, M. B., H. van Zandvoort, MM. Passier, PECA. Rinkel, GLE. Visser-Meily, JMA. (2010). Development and validation of a short version of the Stroke-Specific Quality of Life Scale. *Journal of Neurology, Neurosurgery and Psychiatry, 82*(3).

Post, M. W., & de Witte, L. P. (2003). Good inter-rater reliability of the Frenchay Activities Index in stroke patients. *Clin Rehabil., 17*(5), 548-552.

Post, P. N., Stiggelbout, A. M., & Wakker, P. P. (2001). The utility of health states after stroke: a systematic review of the literature. *Stroke., 32*(6), 1425-1429.

Prescott, R. J., Garraway, W. M., & Akhtar, A. J. (1982). Predicting functional outcome following acute stroke using a standard clinical examination. *Stroke., 13*(5), 641-647.

Price, C. I., Curless, R. H., & Rodgers, H. (1999). Can stroke patients use visual analogue scales? *Stroke., 30*(7), 1357-1361.

Quinn, T. J., Dawson, J., Walters, M. R., & Lees, K. R. (2008). Variability in modified Rankin scoring across a large cohort of international observers. *Stroke., 39*(11), 2975-2979.

Quinn, T. J., Dawson, J., Walters, M. R., & Lees, K. R. (2009). Exploring the reliability of the modified rankin scale. *Stroke., 40*(3), 762-766.

Quinn, T. J., Langhorne, P., & Stott, D. J. (2011). Barthel index for stroke trials: development, properties, and application. *Stroke., 42*(4), 1146-1151.

Quinn, T. J., Lees, K. R., Hardemark, H. G., Dawson, J., & Walters, M. R. (2007). Initial experience of a digital training resource for modified Rankin scale assessment in clinical trials. *Stroke., 38*(8), 2257-2261.

R., B.-G. (2002). Physical function outcome measurement in acute neurology. *Physiotherapy Canada, 52*, 138-145.

Rabins, P. V., & Brooks, B. R. (1981). Emotional disturbance in multiple sclerosis patients: validity of the General Health Questionnaire (GHQ). *Psychol Med., 11*(2), 425-427.

Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scottish medical journal, 2*(5), 200.

Rehab Measures. (2010). Rehabilitation Measures Database.

Richard, C., Lussier, M. T., Gagnon, R., & Lamarche, L. (2004). GHQ-28 and cGHQ-28: implications of two scoring methods for the GHQ in a primary care setting. *Soc Psychiatry Psychiatr.Epidemiol., 39*(3), 235-243.

Richards, L., Stoker-Yates, J., Pohl, P., Wallace, D., & Duncan, P. (2001). Reliability and validity of two tests of upper extremity motor function post-stroke. *Occupational Therapy Journal of Research, 21*(3), 201-219.

Richardson, H. E., & Glass, J. N. (2002). A comparison of scoring protocols on the Clock Drawing Test in relation to ease of use, diagnostic group, and correlations with Mini-Mental State Examination. *J Am Geriatr Soc., 50*(1), 169-173.

Rinaldi, P., Mecocci, P., Benedetti, C., Ercolani, S., Bregnocchi, M., Menculini, G., Catani, M., Senin, U., & Cherubini, A. (2003). Validation of the five-item geriatric depression scale in elderly subjects in three different settings. *J Am Geriatr Soc., 51*(5), 694-698.

Ripat, J., Etcheverry, E., Cooper, J., & Tate, R. B. (2001). A comparison of the Canadian Occupational Performance Measure and the Health Assessment Questionnaire. *Can.J Occup Ther., 68*(4), 247-253.

Roberts, L., & Counsell, C. (1998). Assessment of clinical outcomes in acute stroke trials. *Stroke., 29*(5), 986-991.

Roberts, S. B., Bonnici, D. M., Mackinnon, A. J., & Worcester, M. C. (2001). Psychometric evaluation of the Hospital Anxiety and Depression Scale (HADS) among female cardiac patients. *Br J Health Psychol., 6*(Part 4), 373-383.

Rockwood, K., Awalt, E., Carver, D., & MacKnight, C. (2000). Feasibility and measurement properties of the functional reach and the timed up and go tests in the Canadian study of health and aging. *J Gerontol A Biol.Sci Med Sci., 55*(2), M70-M73.

Roorda, L. D., Green, J. R., Houwink, A., Bagley, P. J., Smith, J., Molenaar, I. W., & Geurts, A. C. (2012). Item hierarchy-based analysis of the Rivermead Mobility Index resulted in improved interpretation and enabled faster scoring in patients undergoing rehabilitation after stroke. *Arch.Phys.Med Rehabil, 93*(6), 1091-1096.

Roorda, L. D., Green, J. R., Houwink, A., Bagley, P. J., Smith, J., Molenaar, I. W., & Geurts, A. C. (2012). The Rivermead Mobility Index allows valid comparisons between subgroups of patients undergoing rehabilitation after stroke who differ with respect to age, sex, or side of lesion. *Arch.Phys.Med Rehabil, 93*(6), 1086-1090.

Royall, D. R., Cordes, J. A., & Polk, M. (1998). CLOX: an executive clock drawing task. *Journal of Neurology, Neurosurgery & Psychiatry, 64*(5), 588-594.

Ruchinskas, R. A., & Curyto, K. J. (2003). Cognitive screening in geriatric rehabilitation. *Rehabilitation Psychology, 48*(1), 14.

Sabari, J. S., Lim, A. L., Velozo, C. A., Lehman, L., Kieran, O., & Lai, J. S. (2005). Assessing arm and hand function after stroke: a validity test of the hierarchical scoring system used in the motor assessment scale for stroke. *Arch Phys Med Rehabil., 86*(8), 1609-1615.

Sackley, C. M., & Lincoln, N. B. (1990). The verbal administration of the gross function scale of the Rivermead Motor Assessment. *Clinical Rehabilitation, 4*(4), 301-303.

Sarker, S. J., Rudd, A. G., Douiri, A., & Wolfe, C. D. (2012). Comparison of 2 extended activities of daily living scales with the Barthel Index and predictors of their outcomes: cohort study within the South London Stroke Register (SLSR). *Stroke, 43*(5), 1362-1369.

Saver, J. L., Filip, B., Hamilton, S., Yanes, A., Craig, S., Cho, M., Conwit, R., & Starkman, S. (2010). Improving the reliability of stroke disability grading in clinical trials and clinical practice: the Rankin Focused Assessment (RFA). *Stroke., 41*(5), 992-995.

Scanlan, J. M., Brush, M., Quijano, C., & Borson, S. (2002). Comparing clock tests for dementia screening: naive judgments vs formal systems--what is optimal? *Int J Geriatr Psychiatry., 17*(1), 14-20.

Schenkenberg, T., Bradford, D. C., & Ajax, E. T. (1980). Line bisection and unilateral visual neglect in patients with neurologic impairment. *Neurology., 30*(5), 509-517.

Schindl, M. R., Forstner, C., Kern, H., & Hesse, S. (2000). Treadmill training with partial body weight support in nonambulatory patients with cerebral palsy. *Arch Phys Med Rehabil., 81*(3), 301-306.

Schlegel, D. J., Tanne, D., Demchuk, A. M., Levine, S. R., & Kasner, S. E. (2004). Prediction of hospital disposition after thrombolysis for acute ischemic stroke using the National Institutes of Health Stroke Scale. *Arch Neurol., 61*(7), 1061-1064.

Schmulling, S., Grond, M., Rudolf, J., & Kiencke, P. (1998). Training as a prerequisite for reliable use of NIH Stroke Scale. *Stroke., 29*(6), 1258-1259.

Schuster, C., Hahn, S., & Ettlin, T. (2010). Objectively-assessed outcome measures: a translation and cross-cultural adaptation procedure applied to the Chedoke McMaster Arm and Hand Activity Inventory (CAHAI). *BMC Med Res Methodol., 10:106.*

Seaby, L., & Torrance, G. (1989). Reliability of a physiotherapy functional assessment used in a rehabilitation setting. *Physiother Can, 41*, 264-271.

Segal, M. E., Gillard, M., & Schall, R. (1996). Telephone and in-person proxy agreement between stroke patients and caregivers for the functional independence measure. *Am J Phys Med Rehabil., 75*(3), 208-212.

Segal, M. E., & Schall, R. R. (1994). Determining functional/health status and its relation to disability in stroke survivors. *Stroke., 25*(12), 2391-2397.

Sheikh, J. I., Yesavage, J. A., & Brink, T. L. (1986). Clinical Gerontology: A Guide to Assessment and Intervention. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version: New York: The Haworth Press1986.

Shinohara, Y., Minematsu, K., Amano, T., & Ohashi, Y. (2006). Modified Rankin scale with expanded guidance scheme and interview questionnaire: interrater agreement and reproducibility of assessment. *Cerebrovasc Dis., 21*(4), 271-278.

Shoemaker, M. J., Mullins-MacRitchie, M., Bennett, J., Vryhof, K., & Boettcher, I. (2006). Predicting response to rehabilitation in elderly patients with stroke using the Orpington Prognostic Scale and selected clinical variables. *J Geriatr Phys Ther., 29*(2), 69-73.

Siegert, R. J., Walkey, F. H., & Turner-Stokes, L. (2009). An examination of the factor structure of the Beck Depression Inventory-II in a neurorehabilitation inpatient sample. *J Int Neuropsychol.Soc., 15*(1), 142-147.

Siggeirsdottir, K., Jonsson, B. Y., Jonsson, H., Jr., & Iwarsson, S. (2002). The timed 'Up & Go' is dependent on chair type. *Clin Rehabil., 16*(6), 609-616.

Simondson, J. A., Goldie, P., & Greenwood, K. M. (2003). The Mobility Scale for Acute Stroke Patients: concurrent validity. *Clin Rehabil., 17*(5), 558-564.

Smith, T., Gildeh, N., & Holmes, C. (2007). The Montreal Cognitive Assessment: validity and utility in a memory clinic setting. *Can.J Psychiatry., 52*(5), 329-332.

Smith, Y. A., & Hong , E. U. N. S. (2000). Normative and validation studies of the Nine-hole Peg Test with children. *Perceptual and Motor Skills, 90*(3), 823-843.

Sneeuw, K. C., Aaronson, N. K., de Haan, R. J., & Limburg, M. (1997). Assessing quality of life after stroke. The value and limitations of proxy ratings. *Stroke., 28*(8), 1541-1549.

Solway, S., Brooks, D., Lacasse, Y., & Thomas, S. (2001). A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. *Chest., 119*(1), 256-270.

Spreen, O., & Risser, A. H. (2002). *Assessment of aphasia*: Oxford University Press, USA.

Stansfeld, S. A., Roberts, R., & Foot, S. P. (1997). Assessing the validity of the SF-36 General Health Survey. *Qual.Life Res., 6*(3), 217-224.

Stark, S. L., Edwards, D. F., Hollingsworth, H., & Gray, D. B. (2005). Validation of the Reintegration to Normal Living Index in a population of community-dwelling people with mobility limitations. *Arch Phys Med Rehabil., 86*(2), 344-345.

Steele, B. (1996). Timed walking tests of exercise capacity in chronic cardiopulmonary illness. *J Cardiopulm Rehabil., 16*(1), 25-33.

Steffen, T. M., Hacker, T. A., & Mollinger, L. (2002). Age- and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds. *Phys Ther., 82*(2), 128-137.

Steiner, A., Raube, K., Stuck, A. E., Aronow, H. U., Draper, D., Rubenstein, L. Z., & Beck, J. C. (1996). Measuring psychosocial aspects of well-being in older community residents: performance of four short scales. *Gerontologist., 36*(1), 54-62.

Stevens, D., Elpern, E., Sharma, K., Szidon, P., Ankin, M., & Kesten, S. (1999). Comparison of hallway and treadmill six-minute walk tests. *Am J Respir.Crit Care Med., 160*(5 Pt 1), 1540-1543.

Stevenson, T. J. (1999). Using impairment inventory scores to determine ambulation status in individuals with stroke. *Physiotherapy Canada, 51*, 168-174.

Stevenson, T. J. (2001). Detecting change in patients with stroke using the Berg Balance Scale. *Aust.J Physiother., 47*(1), 29-38.

Stiles, P. G., & McGarrahan, J. F. (1998). The Geriatric Depression Scale: A comprehensive review. *Journal of Clinical Geropsychology*.

Stone, S. P., Ali, B., Auberleek, I., Thompsell, A., & Young, A. (1994). The Barthel index in clinical practice: use on a rehabilitation ward for elderly people. *J R Coll Physicians Lond., 28*(5), 419-423.

Straube, D., Moore, J., Leech, K., & Hornby, T. G. (2013). Item analysis of the berg balance scale in individuals with subacute and chronic stroke. *Top.Stroke Rehabil, 20*(3), 241-249.

Stroke Unit Trialist's Collaboration. (2007). Organised inpatient (stroke unit) care for stroke. *The Cochrane Database of Systematic Reviews, Oct 17;(4).*

Studenski, S. A., Wallace, D., Duncan, P. W., Rymer, M., & Lai, S. M. (2001). Predicting stroke recovery: three- and six-month rates of patient-centered functional outcomes based on the orpington prognostic scale. *J Am Geriatr Soc., 49*(3), 308-312.

Su, C. Y., Chang, J. J., Chen, H. M., Su, C. J., Chien, T. H., & Huang, M. H. (2000). Perceptual differences between stroke patients with cerebral infarction and intracerebral hemorrhage. *Arch Phys Med Rehabil., 81*(6), 706-714.

Suhr, J., Grace, J., Allen, J., Nadler, J., & McKenna, M. (1998). Quantitative and qualitative performance of stroke versus normal elderly on six clock drawing systems. *Arch Clin Neuropsychol., 13*(6), 495-502.

Suhr, J. A., & Grace, J. (1999). Brief cognitive screening of right hemisphere stroke: relation to functional outcome. *Arch Phys Med Rehabil., 80*(7), 773-776.

Sulter, G., Steen, C., & De, K. J. (1999). Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke., 30*(8), 1538-1541.

Sveen, U., Bautz-Holter, E., Sodring, K. M., Wyller, T. B., & Laake, K. (1999). Association between impairments, self-care ability and social activities 1 year after stroke. *Disabil Rehabil., 21*(8), 372-377.

Svensson, E., & Hager-Ross, C. (2006). Hand function in Charcot Marie Tooth: test retest reliability of some measurements. *Clin Rehabil., 20*(10), 896-908.

Tabali, M., Jeschke, E., Dassen, T., Ostermann, T., & Heinze, C. (2012). The Nottingham Health Profile: a feasible questionnaire for nursing home residents? *Int Psychogeriatr., 24*(3), 416-424.

Tang, A., Sibley, K. M., Bayley, M. T., McIlroy, W. E., & Brooks, D. (2006). Do functional walk tests reflect cardiorespiratory fitness in sub-acute stroke? *J Neuroeng.Rehabil., 3:23.*

Tang, W. K., Mok, V., Chan, S. S., Chiu, H. F., Wong, K. S., Kwok, T. C., Lam, W. W., & Ungvari, G. S. (2005). Screening of dementia in stroke patients with lacunar infarcts: comparison of the mattis dementia rating scale and the mini-mental state examination. *J Geriatr Psychiatry Neurol., 18*(1), 3-7.

Teng, E. L., & Chui, H. C. (1987). The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry., 48*(8), 314-318.

Thompson, M., & Medley, A. (1995). Performance of community dwelling elderly on the Timed Up and Go Test. *Physical & Occupational Therapy in Geriatrics, 13*(3), 17-30.

Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*.

Tooth, L. R., McKenna, K. T., Smith, M., & O'Rourke, P. (2003). Further evidence for the agreement between patients with stroke and their proxies on the Frenchay Activities Index. *Clin Rehabil., 17*(6), 656-665.

Troosters, T., Gosselink, R., & Decramer, M. (1999). Six minute walking distance in healthy elderly subjects. *Eur Respir.J., 14*(2), 270-274.

Turner-Stokes, L., & Hassan, N. (2002). Depression after stroke: a review of the evidence base to inform the development of an integrated care pathway. Part 1: Diagnosis, frequency and impact. *Clinical Rehabilitation, 16*(3), 231-247.

Uyttenboogaart, M., Stewart, R. E., Vroomen, P. C., De, K. J., & Luijckx, G. J. (2005). Optimizing cutoff scores for the Barthel index and the modified Rankin scale for defining outcome in acute stroke trials. *Stroke., 36*(9), 1984-1987.

Valach, L., Signer, S., Hartmeier, A., Hofer, K., & Steck, G. C. (2003). Chedoke-McMaster stroke assessment and modified Barthel Index self-assessment in patients with vascular brain damage. *Int.J.Rehabil.Res., 26*(2), 93-99.

van Bloemendaal, M., Kokkeler, A. M., & van de Port, I. G. (2012). The shuttle walk test: a new approach to functional walking capacity measurements for patients after stroke? *Arch.Phys.Med.Rehabil., 93*(1), 163-166.

van der Lee, J. H., Beckerman, H., Lankhorst, G. J., & Bouter, L. M. (2001). The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. *J.Rehabil.Med., 33*(3), 110-113.

van der Lee, J. H., Roorda, L. D., Beckerman, H., Lankhorst, G. J., & Bouter, L. M. (2002). Improving the Action Research Arm test: a unidimensional hierarchical scale. *Clin Rehabil., 16*(6), 646-653.

van der Putten, J. J., Hobart, J. C., Freeman, J. A., & Thompson, A. J. (1999). Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. *Journal of Neurology, Neurosurgery & Psychiatry, 66*(4), 480-484.

van Marwijk, H. W., Wallace, P., de Bock, G. H., Hermans, J., Kaptein, A. A., & Mulder, J. D. (1995). Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the geriatric depression scale. *Br.J.Gen.Pract., 45*(393), 195-199.

van Straten, A., de Haan, R. J., Limburg, M., Schuling, J., Bossuyt, P. M., & Van den Bos, G. A. (1997). A stroke-adapted 30-item version of the Sickness Impact Profile to assess quality of life (SA-SIP30). *Stroke., 28*(11), 2155-2161.

van Straten, A., de Haan, R. J., Limburg, M., & Van den Bos, G. A. (2000). Clinical meaning of the Stroke-Adapted Sickness Impact Profile-30 and the Sickness Impact Profile-136. *Stroke., 31*(11), 2610-2615.

van Swieten, J. C., Koudstaal, P. J., Visser, M. C., Schouten, H. J., & van, G. J. (1988). Interobserver agreement for the assessment of handicap in stroke patients. *Stroke., 19*(5), 604-607.

van Wijck, F. M., Pandyan, A. D., Johnson, G. R., & Barnes, M. P. (2001). Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. *Neurorehabil.Neural Repair., 15*(1), 23-30.

Velozo, C. A., & Woodbury, M. L. (2011). Translating measurement findings into rehabilitation practice: an example using Fugl-Meyer Assessment-Upper Extremity with patients following stroke. *J.Rehabil.Res.Dev., 48*(10), 1211-1222.

Visser, M. C., Koudstaal, P. J., Erdman, R. A., Deckers, J. W., Passchier, J., van, G. J., & Grobbee, D. E. (1995). Measuring quality of life in patients with myocardial infarction or stroke: a feasibility study of four questionnaires in The Netherlands. *J.Epidemiol.Community Health., 49*(5), 513-517.

Wade, D. T. (1992). Measurement in neurological rehabilitation. *Current Opinion in Neurology, 5*(5), 682-686.

Wade, D. T., & Collin, C. (1988). The Barthel ADL Index: a standard measure of physical disability? *Int.Disabil.Stud., 10*(2), 64-67.

Wade, D. T., Legh-Smith, J., & Langton, H. R. (1985). Social activities after stroke: measurement and natural history using the Frenchay Activities Index. *Int.Rehabil.Med., 7*(4), 176-181.

Wallace, D., Duncan, P. W., & Lai, S. M. (2002). Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. *J.Clin Epidemiol., 55*(9), 922-928.

Walters, S. J., Munro, J. F., & Brazier, J. E. (2001). Using the SF-36 with older adults: a cross-sectional community-based survey. *Age Ageing., 30*(4), 337-343.

Wang, C. H., Hsieh, C. L., Dai, M. H., Chen, C. H., & Lai, Y. F. (2002). Inter-rater reliability and validity of the stroke rehabilitation assessment of movement (stream) instrument. *J.Rehabil.Med., 34*(1), 20-24.

Wang, C. H., Hsueh, I. P., Sheu, C. F., Yao, G., & Hsieh, C. L. (2004). Psychometric properties of 2 simplified 3-level balance scales used for patients with stroke. *Phys.Ther., 84*(5), 430-438.

Ward, I., Pivko, S., Brooks, G., & Parkin, K. (2011). Validity of the stroke rehabilitation assessment of movement scale in acute rehabilitation: a comparison with the functional independence measure and stroke impact scale-16. *PM.R., 3*(11), 1013-1020.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med.Care., 30*(6), 473-483.

Waters, D. (1995). Recovering from a depressive episode using the Canadian Occupational Performance Measure. *Canadian Journal of Occupational Therapy, 62*(5), 278-282.

Wee, J. Y., Bagg, S. D., & Palepu, A. (1999). The Berg balance scale as a predictor of length of stay and discharge destination in an acute stroke rehabilitation setting. *Arch.Phys.Med.Rehabil., 80*(4), 448-452.

Wee, J. Y., Wong, H., & Palepu, A. (2003). Validation of the Berg Balance Scale as a predictor of length of stay and discharge destination in stroke rehabilitation. *Arch.Phys.Med.Rehabil., 84*(5), 731-735.

Wei, X. J., Tong, K. Y., & Hu, X. L. (2011). The responsiveness and correlation between Fugl-Meyer Assessment, Motor Status Scale, and the Action Research Arm Test in chronic stroke with upper-extremity rehabilitation robotic training. *Int.J.Rehabil.Res., 34*(4), 349-356.

Weinberger, M., Oddone, E. Z., Samsa, G. P., & Landsman, P. B. (1996). Are health-related quality-of-life measures affected by the mode of administration? *J.Clin Epidemiol., 49*(2), 135-140.

Wendel, K. A., StÇ¾hl, A., & Iwarsson, S. (2013). Inter-rater agreement of a modified and extended Swedish version of the Frenchay Activities Index (FAI). *European Journal of Ageing, 10*(3), 247-255.

Werneke, U., Goldberg, D. P., Yalcin, I., & Ustun, B. T. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychol.Med., 30*(4), 823-829.

Westergren, A., & Hagell, P. (2006). Initial validation of the Swedish version of the London Handicap Scale. *Qual.Life Res., 15*(7), 1251-1256.

Wevers, L. E., Kwakkel, G., & van de Port, I. G. (2011). Is outdoor use of the six-minute walk test with a global positioning system in stroke patients' own neighbourhoods reproducible and valid? *J.Rehabil.Med., 43*(11), 1027-1031.

Whitall, J., Savin, D. N., Jr., Harris-Love, M., & Waller, S. M. (2006). Psychometric properties of a modified Wolf Motor Function test for people with mild and moderate upper-extremity hemiparesis. *Arch.Phys.Med.Rehabil., 87*(5), 656-660.

Whitney, S. L., Poole, J. L., & Cass, S. P. (1998). A review of balance instruments for older adults. *Am.J.Occup.Ther., 52*(8), 666-671.

WHO, W. H. O. (2001). International classification of functioning disability and health (ICF).

WHO, W. H. O. (2002). Towards a common language for functioning, disability and health: ICF. *Geneva: WHO*, 8-9.

Williams, C. L., Rittman, M. R., Boylstein, C., Faircloth, C., & Haijing, Q. (2005). Qualitative and quantitative measurement of depression in veterans recovering from stroke. *J.Rehabil.Res.Dev., 42*(3), 277-290.

Williams, L. S., Bakas, T., Brizendine, E., Plue, L., Tu, W., Hendrie, H., & Kroenke, K. (2006). How valid are family proxy assessments of stroke patients' health-related quality of life? *Stroke., 37*(8), 2081-2085.

Williams, L. S., Redmon, G., Martinez, B., & Weinberger, M. (2000). Proxy ratings of stroke specific quality of life (SSQOL) scores. *Stroke, 31*, 301.

Williams, L. S., Weinberger, M., Harris, L. E., & Biller, J. (1999). Measuring quality of life in a way that is meaningful to stroke patients. *Neurology., 53*(8), 1839-1843.

Williams, L. S., Weinberger, M., Harris, L. E., Clark, D. O., & Biller, J. (1999). Development of a stroke-specific quality of life scale. *Stroke., 30*(7), 1362-1369.

Willmott, S. A., Boardman, J. A., Henshaw, C. A., & Jones, P. W. (2004). Understanding General Health Questionnaire (GHQ-28) score and its threshold. *Soc.Psychiatry Psychiatr.Epidemiol., 39*(8), 613-617.

Wilson, B., Cockburn, J., & Halligan, P. (1987). Development of a behavioral test of visuospatial neglect. *Arch.Phys.Med.Rehabil., 68*(2), 98-102.

Wilson, J. T., Hareendran, A., Grant, M., Baird, T., Schulz, U. G., Muir, K. W., & Bone, I. (2002). Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. *Stroke., 33*(9), 2243-2246.

Wilson, J. T., Hareendran, A., Hendry, A., Potter, J., Bone, I., & Muir, K. W. (2005). Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke., 36*(4), 777-781.

Wolf, S. L., Catlin, P. A., Ellis, M., Archer, A. L., Morgan, B., & Piacentino, A. (2001). Assessing Wolf motor function test as outcome measure for research in patients after stroke. *Stroke., 32*(7), 1635-1639.

Wolf, S. L., Lecraw, D. E., Barton, L. A., & Jann, B. B. (1989). Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Exp.Neurol., 104*(2), 125-132.

Wolf, S. L., McJunkin, J. P., Swanson, M. L., & Weiss, P. S. (2006). Pilot normative database for the Wolf Motor Function Test. *Arch.Phys.Med.Rehabil., 87*(3), 443-445.

Wolf, S. L., Thompson, P. A., Morris, D. M., Rose, D. K., Winstein, C. J., Taub, E., Giuliani, C., & Pearson, S. L. (2005). The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. *Neurorehabil.Neural Repair., 19*(3), 194-205.

Wolfe, C. D., Taub, N. A., Woodrow, E. J., & Burney, P. G. (1991). Assessment of scales of disability and handicap for stroke patients. *Stroke., 22*(10), 1242-1244.

Wong, G. K., Lam, S. W., Ngai, K., Wong, A., Poon, W. S., & Mok, V. (2012). Validation of the Stroke-specific Quality of Life for patients after aneurysmal subarachnoid hemorrhage and proposed summary subscores. *J Neurol Sci, 320*(1-2), 97-101.

Wong, G. K., Lam, S. W., Ngai, K., Wong, A., Poon, W. S., & Mok, V. (2013). Development of a short form of Stroke-Specific Quality of Life Scale for patients after aneurysmal subarachnoid hemorrhage. *J Neurol Sci*.

Woo, D., Broderick, J. P., Kothari, R. U., Lu, M., Brott, T., Lyden, P. D., Marler, J. R., & Grotta, J. C. (1999). Does the National Institutes of Health Stroke Scale favor left hemisphere strokes? NINDS t-PA Stroke Study Group. *Stroke., 30*(11), 2355-2359.

Wood-Dauphinee, S., & Williams, J. I. (1987). Reintegration to Normal Living as a proxy to quality of life. *J.Chronic.Dis., 40*(6), 491-502.

Wood-Dauphinee, S. L., Opzoomer, M. A., Williams, J. I., Marchand, B., & Spitzer, W. O. (1988). Assessment of global function: The Reintegration to Normal Living Index. *Arch.Phys.Med.Rehabil., 69*(8), 583-590.

Woodbury, M. L., Velozo, C. A., Richards, L. G., Duncan, P. W., Studenski, S., & Lai, S. M. (2007). Dimensionality and construct validity of the Fugl-Meyer Assessment of the upper extremity. *Arch.Phys.Med.Rehabil., 88*(6), 715-723.

Woodbury, M. L., Velozo, C. A., Richards, L. G., Duncan, P. W., Studenski, S., & Lai, S. M. (2008). Longitudinal stability of the Fugl-Meyer Assessment of the upper extremity. *Arch.Phys.Med.Rehabil., 89*(8), 1563-1569.

Wressle, E., Marcusson, J., & Henriksson, C. (2002). Clinical utility of the Canadian Occupational Performance Measure--Swedish version. *Can.J.Occup.Ther., 69*(1), 40-48.

Wright, C. J., Swinton, L. C., Green, T. L., & Hill, M. D. (2004). Predicting final disposition after stroke using the Orpington Prognostic Score. *Can.J.Neurol.Sci., 31*(4), 494-498.

Wu, C. Y., Chuang, L. L., Lin, K. C., & Horng, Y. S. (2011). Responsiveness and validity of two outcome measures of instrumental activities of daily living in stroke survivors receiving rehabilitative therapies. *Clin Rehabil., 25*(2), 175-183.

Wyller, T. B., Sveen, U., & Bautz-Holter, E. (1996). The Frenchay Activities Index in stroke patients: agreement between scores by patients and by relatives. *Disabil.Rehabil., 18*(9), 454-459.

Yim, S. C., JR. Lee, IY. (2003). Normative data and developmental characteristics of hand function for elementary school children in Suwon area of Korea: grip, pinch and dexterity study. *J Korean Med Sci, 18*, 552-558.

Yozbatiran, N., Der-Yeghiaian, L., & Cramer, S. C. (2008). A standardized approach to performing the action research arm test. *Neurorehabil.Neural Repair., 22*(1), 78-90.

Zandieh, A., Kahaki, Z. Z., Sadeghian, H., Pourashraf, M., Parviz, S., Ghaffarpour, M., & Ghabaee, M. (2012). The underlying factor structure of National Institutes of Health Stroke scale: an exploratory factor analysis. *Int.J.Neurosci., 122*(3), 140-144.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatr.Scand., 67*(6), 361-370.

Zwick, D., Rochelle, A., Choksi, A., & Domowicz, J. (2000). Evaluation and treatment of balance in the elderly: A review of the efficacy of the Berg Balance Test and Tai Chi Quan. *NeuroRehabilitation., 15*(1), 49-56.