Evidence-Based Review of Stroke Rehabilitation

Outcome Measures in Stroke Rehabilitation

Katherine Salter BA

Jeffrey Jutai PhD

Laura Zettler BHSc

Norine Foley MSc

Robert Teasell MD

From the Departments of Physical Medicine and Rehabilitation, St. Joseph Health Care, London, Parkwood Hospital, London and Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada

ST JOSEPH's
HEALTH CARE LONDON

Canadian Stroke Network
Réseau Canadien contre
Les accidents cérébrovasculaires

Western

9th Edition

Outcome Measures in Stroke Rehabilitation

Katherine Salter BA, Jeffrey Jutai PhD, Laura Zettler BHSc, Norine Foley MSc, Robert Teasell MD

Last updated July 2006

Table of Contents

21. Outcome Measures in Stroke Rehabilitation

21.1 Introduction

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

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 by Finch et al. (2002). This chapter will present only the information most relevant for stroke rehabilitation.

21.1.1 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 21.1 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 21.2 Classification of Outcome Measures*

Body structure (*impairments*)	Activities (*limitations to activity–disability*)	Participation (*barriers to participation--handicap*)
Beck Depression Inventory Canadian Neurological Scale Clock Drawing Test Fugl-Meyer Assessment of Motor Recovery after Stroke General Health Questionnaire (GHQ-28) Geriatric Depression Scale MMSE Modified Ashworth Scale MVPT National Institutes of Health Stroke Scale (NIHSS) Orpington Prognostic Scale	Action Research Arm Test Barthel Index Berg Balance Scale Chedoke McMaster Stroke Assessment Scale COVS Functional Ambulation Categories FIM Frenchay Activities Index Motor Assessment Scale Nine-hole Peg Test Rankin Handicap Scale TUG	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

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

21.1.2 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; van der Putten et al. 1999; Roberts & Counsell, 1998; Law, 2002). Finch et al. (2002) provide a good tutorial on issues for outcome measure selection.

The Health Technology Assessment (HTA) programme (Fitzpatrick et al. University of Southampton, UK, 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 21.3 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 21.3 Evaluation Criteria and Standards

Criterion	Definition	Standard
1. Appropriateness	The match of the instrument to the purpose/question under study. One must determine what information is required	Depends upon the specific purpose for which the measurement is intended.

Criterion	Definition	Standard
	and what use will be made of the information gathered (Wade 1992)	
2. 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]
3. Validity	Does the instrument measure what it purports to measure? Forms of validity include face, content, 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).	*Construct/convergent and concurrent correlations*: Excellent: ≥ 0.60, Adequate: $0.31 - 0.59$, Poor: ≤ 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)
4. 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	*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.

Criterion	Definition	Standard
	which no further improvement or deterioration can be noted.	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]
5. 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).
6. 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.
7. 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?	
8. 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 Andresen (2000).

Table 21.4 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 21.5, below).

Table 21.5 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 21.3.   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?

21.1.3 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 would seem 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).

21.1.4 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 (Segal et al. 1996; Sneeuw et al. 1997; Hachisuka 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, using a healthcare professional as a substitute for the family member or significant other as proxy does not solve this problem. 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 (Sneeuw et al. 1997; McGinnis et al. 1986). 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).

21.1.5 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. (1999), by 6 months post-stroke, physical recovery is complete, for the most part, with additional gains being a function of learning, practice and confidence. 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.

In this chapter, the main results of our evaluation are summarized. A table was prepared for each instrument detailing its reliability, validity, responsiveness and other properties, and citing the appropriate references from the published literature. To save space, the tables are not presented here. Please contact Katherine Salter (*Katherine.Salter@sjhc.london.on.ca*) to obtain this information.

21.2 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).

21.2.1 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. 1988; 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).

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 (Aben et al. 2002) especially among female patients.

The Beck Depression Inventory, while it is a self-report measure, has not been tested for use with proxy respondents, perhaps due to the highly subjective nature of the items included on the inventory. Aben et al. (2002) reported difficulty with scale completion within a stroke population.

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). The scale has not been tested for administration using proxy respondents.
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 21.6 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)

21.2.2 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. 1986, 1989). 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. 1986, 1989).

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 (Cote et al. 1986, Anamaet 2002).   It has been demonstrated that the CNS is a valid predictor of outcomes such as length of stay, death and dependency (see table).

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.
Fe*asibility* 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 21.7 – 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;

21.2.3 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; Suhr et al. 1998; McDowell & Newell 1996).

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 (Ruchinskas & Curyto 2003; McDowell & Newell 1996; 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 (Scanlan et al. 2002; Ruchinskas & Curyto 2003; McDowell & Newell 1996).

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 (Richardson & Glass 2002; Heinrik et al. 2004). 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).

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 (Ruchinskas & Curyto 2003; Lorentz et al. 2002), 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)

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 21.8 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;

21.2.4 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 (Fugl-Meyer, 1980; Duncan et al. 1994).

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.

Limitations
The assessment, administered in its entirety, is quite lengthy (Gladstone et al. 2002). 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).

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 (van der Lee et al. 2001). In these instances, the FMA may not be the most appropriate assessment tool.

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 (Hseuh et al. 2001 in 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).

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 (Fugl-Meyer et al. 1975; Fugl-Meyer, 1980; Duncan et al. 1994).

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 21.9 – 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-S)	+ (FMA-S)

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

21.2.5 General Health Questionnaire – 28 (GHQ-28)

The General Health Questionnaire (GHQ) is a screening tool developed to detect possible cases of psychiatric disorders (McDowell and 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 and 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 and Hillier 1979). The GHQ is concerned with four aspects of distress: depression, anxiety, social impairment, and hypochondriasis (McDowell and 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 and Newell 1996). Thus, the GHQ was not designed to detect long-standing phenomena (chronic illnesses) that have become 'usual' to the individual (Richard 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 and 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 and Hillier 1979). The GHQ-28 has been recommended for detecting morbidity in posttraumatic clinical and research settings (Raphael, Lundin & Weisaeth, 1989 as cited in 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 xx. 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)

Table 21.10 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)	

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 and 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) and Lobo et al. (1988). 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. (1997), 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.

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 and 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 and 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. (1998) 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. 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 21.11 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;

21.2.6 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 are 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 (van Marwijk et al. 1995; Almeida & Almeida, 1999; MacNeill & Lichtenberg 2000; Rinaldi et al. 2003).

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 and 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 (McDowell & Newell 1996; Kafonck et al. 1989; 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 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 (van Marwijk et al. 1995; Almeida & Almeida 1999).

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 21.12 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;

21.2.7 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 & Chui (1987) increasing the content, number and difficulty of items included in the assessment. 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 cutoff 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, overall score (Wade 1992;Tombaugh & McIntyre 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).

MMSE scores have been shown to be affected by age, level of education and sociocultural background (Tombaugh & McIntyre, 1992; Bleeker et al. 1988; Lorentz et al. 2002). These variables may introduce bias leading to the misclassification of individuals. Though perhaps the prevalent view, such biases have not always been reported. For instance, Agrell & Dehlin (2000) found neither age nor education to influence scores. Lorentz et al. (2002) expressed concern that adjustments made for these biases may limit the general utility of the MMSE.

Perhaps the greatest limitation of the MMSE is its low reported levels of sensitivity particularly among individuals with mild cognitive impairment (Tombaugh & McIntyre, 1992; de Koning et al. 1998) and in patients with right-sided lesions within a general neurological patient population (Dick et al. 1984) and within a stroke population (Suhr & Grace 1999; Blake et al. 2002, Nys et al. 2005). 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 its low level of sensitivity derives from the emphasis placed on language items and a paucity of visual-spatial items (Grace et al. 1995; de Koning et al. 1998; Suhr & Grace, 1999; de Koning et al. 2000;). Blake et al. (2002) noted that while it was possible to use the MMSE as a screening assessment for general cognitive impairment (though with a low level of sensitivity – 62%), it was not possible to identify suitable cut-off points for use in assessment for the presence of either visual or verbal memory deficits.

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 (Lorentz et al. 2002) and would have little effect on the simplicity and accessibility of the test.

Summary – Mini Mental State Examination

Practicality

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 (Ruchinskas & Curyto 2003) *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 21.13 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)

21.2.8 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 & 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; Pandyan et al. 1999; Gregson et al. 2000). 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 21.14 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; Pandyan et al. 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; Pandyan et al. 2000) 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 confounded by contracture. Pandyan et al. (2003) suggest that even taken as a measure of resistance to passive movement only, 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, Pandyan et al. 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. The greater degree of discrimination introduced to the scale by Bohannon and Smith (1987) may be accompanied by a reduction in the scale's reliability (Damiano et al. 2002; Haas et al. 1996).

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; Gregson et al. 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 (Pandyan et al. 1999; Gregson et al. 2000). Assessments of ankle plantarflexors often demonstrate low levels of reliability (Pandyan et al. 1999, Gregson et al. 2000, Haas et al. 1996). 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).

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 MAS.

Table 21.15 Modified Ashworth 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;*

21.2.9 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 (Su et al. 2000; Mercier et al. 2001). 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 & Kwatny, 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 & Kwatney, 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 (Mercier et al. 1995) while maintaining high levels of reliability (Mercier et al. 1995). However, elimination of this variable may not always be desirable, as in a test of driving ability (Mazer et al. 1988).

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 21.16 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)*

21.2.10 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 used rapidly following acute admission (Anamaet 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 (Brott et al. 1989, Heinemann et al. 1997, Anamaet 2002, Schlegel 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 (Brott et al. 1989, Anamaet 2002).

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

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 et al. 1997, Dewey et al. 1999).

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. 1994).

Poor agreement for the item "limb ataxia" has been reported repeatedly (Goldstein et al. 1989; Schmulling et al. 1998; Dewey et al. 1999). Lyden et al. (1999) 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). Based on the results of their factor analysis, Lyden et al. (2001) proposed a scale revision that eliminated this item as well as several other that had demonstrated poor item loadings on identified factors (see mNIHSS in Table above).

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). It 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.

When used for retrospective evaluation, scoring is difficult. Lower reliability and item completion rates have been reported than for the CNS (Anamaet 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 21.17 NIHSS Evaluation Summary

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

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

23.2.11 Orpington Prognostic Scale (OPS)

The Orpington Prognostic Scale (OPS; Kalra and 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 21.12.

Table 21.18 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)	0.0
▪ MRC grade 4 (Diminished power)	0.4
▪ MRC grade 3 (Movement against gravity)	0.8
▪ MRC Grade 1 – 2 (Movement with gravity eliminated or trace)	1.2
▪ MRC Grade 0 (No movement)	1.6
B. Proprioception (eyes closed)	
(Locates affected thumb)	
▪ Accurately	0.0
▪ Slight difficulty	0.4
▪ Finds thumb via arm	0.8
▪ Unable to find thumb	1.2
C. Balance	
▪ Walks 10 feet without help	0.0
▪ Maintains standing position	0.4
▪ Maintains sitting position	0.8
▪ No sitting balance	1.2
D. Cognition	
Based on administration of Hodkinson's Mental Test	
▪ Mental test score 10	0.0
▪ Mental test score 8-9	0.4
▪ Mental test score 5-7	0.8
▪ Mental test score 0-4	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	

Ref: 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 and 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 and 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 and meal preparation 6 months post stroke (Lai et al. 1998). 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. 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. (1994) reported that assessment 2 weeks after the stroke event is optimal with regard to predictive ability. However, several studies have demonstrated significant predictive ability of OPS scores obtained within 14 days of the stroke event (Lai et al. 1998; 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).

Kalra et al. (1994) reported that the predictive values for dependence and discharge destination was not as strong in the middle group of patients (OPS 3 – 5, 2 weeks post stroke) as for patients with mild or severe deficits. 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

Practicality

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 21.19 Orpington Prognostic Scale 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;*

21.3 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 category is rarely perfect, measures appearing in this section focus primarily on the identification or assessment of limitations in activity.

21.3.1 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 (1981) 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). 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 (deWeerdt & Harrison 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. The test covers most aspects of arm function, including proximal control and dexterity. No formal training is required to administer the test. Since the scoring of the ARA test is based on a hierarchical Guttman scale, the testing can be completely quickly on higher functioning patients.

Limitations
The scale was originally developed using a small sample of patients (n=20), with differing diagnosis, including non-stroke. Although the scale has good concurrent validity, other forms of validity have not been evaluated within the stroke population.

Significant floor and ceiling effects have been identified. In patients with severe impairments or near normal function, the scale may not be able to assess change in performance (Van der Lee et al. 2002).

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 only guidelines for administration appear in the original publication and contain limited detail.

Table 21.20 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*

21.3.2 Barthel Index (BI)

The Barthel Index of Activities of Daily Living (BI) has been in use since 1955 (Mahoney & Barthel 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 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 and has been shown to be equally reliable when administered by skilled and unskilled individuals (Collin & Wade 1988).

Advantages
The clearest advantage of the Barthel 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.

Limitations
Perhaps the most common criticism of the Barthel Index is its relative insensitivity and lack of comprehensiveness particularly as is reflected in large reported ceiling and floor effects. 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 & 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. 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 Barthel Index 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 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 (Roberts & Counsell 1998; Sulter et al. 1999; Duncan et al. 2000).

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 21.21 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)*

21.3.3 Berg Balance Scale (BBS)

The Berg Balance Scale provides a quantitative assessment of balance in older adults (Berg et al. 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 requires a ruler, a stopwatch, chair, step or stool, room to turn 360° and 10 – 15 minutes and is administered via direct observation of task completion (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 is represents 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. 1992a; Zwick et al. 2000).

Advantages. The Berg Balance Scale measures a number of different aspects of balance, both static and dynamic, and does so with relatively little equipment or space required (Whitney et al. 1998; Nakamura, 1998; Zwick et al. 2000). No specialized training is required to be able to administer the BBS (Nakamura et al. 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. It should also be noted that the extremely high values reported for internal consistency reliability might be indicative of item redundancy.

As the BBS takes somewhat longer than other balance measures to administer (Whitney et al. 1998, Chou et al. 2006) 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 increased floor effect may, in part, be attributed to the removal of the simplest item on the scale (unsupported sitting). Further evaluation of this alternate version of the BBS is required.

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.

Limitations. The BBS takes somewhat longer than other balance measures (Whitney et al. 1998) and may not be suitable for the evaluation of active, elderly persons, as the items included are not sufficiently challenging for this group (Berg et al. 1989; Nakamura et al. 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 Rasch analysis of the Berg Balance Scale 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 (Kornetti et al. 2004). 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).

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 to indicate 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 21.22 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)

21.3.4 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 et al. 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 et al. 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 (2001) concluded that, by comparison, the CMSA is comprehensive and has been well studied for reliability and validity.

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 21.23 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)

21.3.5 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 (Seaby & Torrance 1989; Hajek et al. 1997; Hajek et al. 1997; Eng et al. 2002). 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 (Seaby & Torrance 1989; Finch et al. 2002). 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 (Barclay-Goddard 2000; Low Choy et al 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 and is user-friendly in that it was designed to be performed as part of a routine physiotherapy assessment (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 (Huijbregts, 1996).

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 21.24 – COVS 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;*

21.3.6 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 (Holden et al. 1984, 1986, Hesse et al. 1994), multiple sclerosis (Holden et al. 1984, 1986), stroke (Collen et al. 1990, Stevenson 1999, Simondson et al. 2003, Brock et al. 2002, Lord et al. 2004, Cunha et al. 2002, da Cunha et al. 2002) 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. 1984, 1986; 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. (1984, 1986).

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 lacks responsiveness, especially if using it to distinguish between groups at lower levels of functioning (Collen et al. 1990). Given the large ceiling effects

associated with its use, diagnostic or prognostic conclusions should not be based on the FAC.

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 21.25 FAC 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;*

21.3.7 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.

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 enhance the interpretability of FIM change scores; however, it should be noted that the external criterion around which these figures were developed were retrospective physician ratings of change (Beninato et al. 2006). Assessments of change provided by the patient, caregiver or family assessments were not included. 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 demonstrate 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 21.26 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)

21.3.7.1 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 & Svensson 2000; Hsueh IP 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 & 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 (van der Putten et al. 1999; Hobart et al. 2001; Wallace et al. 2002; Hsueh 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 (van der Putten et al. 1999; Hsueh et al. 2002) 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.

21.3.7.2 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 (CIHI 1999). 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.

21.3.8 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 an objective measurement 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. Administered in an interview format (with or without the patient's family), the FAI takes approximately 5 minutes to complete (Segal & Schall 1994).

Advantages
The brevity and simplicity of the FAI make it easy to use in a clinical setting (Wade 1992). 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

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. 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 seemed to be a balance of gender dominance.

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 (Wade et al. 1985; Piercy et al. 2000; Green et al. 2001). 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 and de Witte 2003).

While the FAI has been assessed for use by proxy with good results for the total score, there is less agreement between proxy and patient assessments at the item level (Wyller et al. 1996; Tooth et al. 2003). 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 recent 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 (Tooth et al. 2003). 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 21.27 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)*

21.3.9 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" (Bamford et al. 1989). The conventional method of administration for the Rankin Scale is via a guided interview process.

Table 21.28 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 measure with well-studied reliability and is a time efficient tool by which to categorize level of functional outcome. As such, it is feasible for use large centers or in large trials (Wade 1992; deHaan et al. 1995). DeHaan et al. (1995) 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.

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). Recently, a structured interview format for the administration of the Modified Rankin Scale has become available. Use of the structured interview has been associated with significant improvements in interobserver reliability (Wilson et al. 2002, Wilson et al. 2005). In addition, a recent guided interview and accompanying questionnaire in Japanese has been published (Shinohara et al. 2006).

Although the scale might be suitable for dichotomized groupings, there is no standardized or consistent point at which this is done (Sulter et al. 1999, New & Buchbinder 2006) 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 et al. (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 favourable vs. poor outcomes.

Table 21.29 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)*

21.3.10 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, Poole and Whitney 1988, Malouin et al. 1994, Sabari et al. 2005). Patients perform each task 3 times and the best performance 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. (1985) as are the criteria for grading each item and a list of general rules and equipment for the administration of the MAS. While Carr et al. (1985) suggest that administration of the MAS requires approximately 15 minutes, subsequent studies report administration times ranging from 15 to 60 minutes (Poole and Whitney 1988, Malouin et al. 1994).

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 such items as a stopwatch, 8 jellybeans, a rubber ball, a stool, comb, spoon, pen, teacups, water and a table. Carr et al. (1985) recommend a short instruction and practice period, including practice assessment on at least 6 patients, prior to using the test in a formal setting.

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 and Whitney 1988). This item is often omitted from the scale and reports using the MAS or about the MAS may not include it (Malouin et al. 1994, Loewen & Anderson 1990).

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. Given these results, use of the upper limb items as a separate scale to evaluate upper extremity function (UL/UE-MAS) should be approached with caution, despite reports of acceptable reliability and validity (Lannin 2004, Hsueh and Hsieh 2002).

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. (1985). A period of instruction and practice assessment is recommended prior to formal use in a clinical or research setting. While the list of equipment required for administration is relatively long, items are commonly available.

Table 21.30 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;*

21.3.11 Nine-hole Peg Test (NHPT)

Kellor et al. developed the Nine-hole Peg Test (NHPT) in 1971 as a test of manual dexterity to be used primarily within the practice of occupational therapy. The NHPT is classified as an activity measure as it assesses fine motor coordination and the ability to manipulate small objects with the hands and fingers. This tool is extensively used in the physical disability and rehabilitation setting as a test of upper-extremity motor function in patients suffering from conditions including, but not limited to, arthritis (Backman et al. 1992), Parkinson's disease, multiple sclerosis, stroke (Wade 1994), and tetraplegia (van Tuijl et al. 2002). It has also been suggested as a useful measure of disability resulting from sensory loss and ataxia (Wade 1994)

The NHPT is a relatively simple and quick test. While seated at a table, the subject picks up nine wooden dowels and then places them into nine corresponding holes on a board (Wade 1994). Each hand is tested separately. Several versions of the NHPT are available commercially (Davis et al. 1999, Grice et al. 2003), but they are all relatively similar. Wade (1989) distinguishes between two main variations of the test: some prefer to time how long it takes to place the nine pegs (Heller et al. 1987), whereas others observe the time it takes to place *and* remove all nine pegs (Kellor et al. 1971). The results can then be conveyed in three different ways: 1) the time it took the subject to complete the test; 2) the number of pegs placed in 50 seconds if the subject was unable to place all nine pegs within that time or 3) the time it took to place *each* peg (Wade et al. 1989). Wade (1994, p.171) suggests the latter as the most appropriate method to express the results of the NHPT.

As mentioned above, commercial versions of the NHPT can be purchased or it is relatively easy to construct the apparatus. Wade (1994) provides an excellent description of the tool and Mathoiwetz et al. (1985) gives a description and diagrammatic representation of the one used in their study. Backman et al. (1992) reported that it takes less than five minutes to administer and score the test.

Advantages
The NHPT is a simple, brief, and portable test, which makes it feasible and acceptable for both client and clinician (Wade 1994). By nature of being a timed test, the NHPT should be a sensitive measure (Wade 1989), and studies have shown this sensitivity, especially at the upper range of function and in the later stages of recovery (Heller et al. 1987, Sunderland et al. 1989).

Limitations

According to Wade (1994), because the main focus of the NHPT is manual dexterity, it lacks ability to detect loss of proximal strength. Also, neglect and other various cognitive problems may have an effect on test performance. The NHPT has been found to have very large floor effects, which makes it unable to distinguish between moderate and severe cases of disability (Wade 1994). Mathiowetz et al. (1985) urges that the NHPT is not a measure for normal subjects or for determining the effects of treatment on finger dexterity.

Performance on the NHPT may be influenced by age. Grice et al. (2003) found a strong correlation between NHPT scores and age (males, $r = 0.908, 0.918$; females, $r = 0.890, 0.896$). Although Heller et al. (1987) found a significant correlation between pegs/second and age ($r = -0.42$, $p < 0.01$), age only explained 17% of the variance, and the correlations of age with grip strength were non-significant.

While normative values for adults have been published, the generalizability of these values is questionable since few people 75 years of age or older participated in the normative studies (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 (1994) 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. Normative values for completion of the NHPT have been published. However, they should be used with caution when evaluating the performance of elderly individuals.
Acceptability. The test is quick and easy to administer and evaluate. It may not prove to be an easy task for the patient, depending on his/her level of functional disability however, it will only require 50 seconds of their time, if the method described by Heller et al. (1987) is used.
Feasibility. The apparatus is a small, portable, and inexpensive. To use it requires very little time from the perspective of both patient and clinician.

Table 21.31 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;

21.3.12 Timed "Up & Go" Test (TUG)

An objective measure of basic mobility and balance manoeuvres; 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 timed up & go 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. 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. 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).

No normative data is available for the TUG, so its primary use has been assessment of change within the individual (Thompson & Medley 1995). Thompson & 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 ($p<0.0001$).

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 21.32 TUG 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)

21.4 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.

21.4.1 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 1980). 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. 1994a). 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, Harwood et al. 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 and Ebrahim 2000a, 2000b). 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 1994a, 1994b).

The LHS is designed as a self-report questionnaire, although it may be completed by a carer or appropriate informant (Harwood et al. 1994a). 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 et al. 2003). The LHS has been translated into several other languages including Dutch (Perenboom et al. 2003) and Hong Kong Chinese (Lo et al. 2001).

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 and 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 and 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. 1994a). As such, the authors recommend that the scale be used for group comparisons (eg. 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. (2004) reported a large ceiling effect when the scale was used to assess handicap in a group of healthy, community dwelling adults.

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 21.33 LHS Evaluation Summary

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

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

21.4.2 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 (Ware & Sherbourne 1992; McHorney et al. 1993).

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 (Andreson & 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 1995). As a self-completed, mailed questionnaire, it has been shown to have reasonably high response rates (83% – Brazier et al. 1992, O'Mahoney et al. 1998; 75% - 83% Dorman et al. 1998; 85% - Dorman et al. 1999; 82% overall & 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; Hayes et al. 1995; Brazier et al. 1996). O'Mahoney 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 (O;Mahoney et al. 1998; Andresen et al. 1999). 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 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). These two scales can account for only 60% of the variance in SF-36 scores suggesting a significant loss of information when the 2-component model is used.

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 21.34 SF-36 Evaluation Summary

Reliability		Validity		Responsiveness		
Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
+++	++ (TR) ++ (IC)	+++	+++	++	+++ *(Note: 1 study reported ES)*	+

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

21.4.3 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. 1980; Hunt et al. 1981; Hunt et al. 1984, Hunt et al. 1985). It was not intended to be a measure of health-related quality of life or as a means to identify specific health conditions (Hunt et al. 1984; Bowling, 1997). 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 (de Haan et al. 1993; Coons et al. 2000). As a postal questionnaire, reported response rates range from 68 – 93% (Hunt et al. 1985; Brazier et al. 1992; Ebrahim et al. 1986). 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 (Hunt et al. 1985; Bowling, 1997). 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 (de Haan et al. 1993; Bowling, 1997; Stansfeld et al. 1997; Coons et al. 2000).

The use of the weights provided with the scale items has been criticized as being inappropriate and confounded (Jenkinson, 1991; Anderson et al. 1993). In his 1991 study, Jenkinson 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 21.35 Evaluation Summary for Nottingham Health Profile

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

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

21.4.4 Reintegration to Normal Living Index (RNLI)

The Reintegration to Normal Living Index (Wood-Dauphinee and 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 and incapacitating illness or trauma*" (Wood-Dauphinee and 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. 2000, Mayo et al. 2002). 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. 2000, Mayo et al. 2002).

The RNLI 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 (Steiner et al. 1996, Carter et al. 2000).

The RNLI 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 RNLI 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 21.36 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;

21.4.5 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. 1976,1981). 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 (van Straten et al. 1997; 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 (van Straten et al. 1997; Finch et al. 2002; Cup et al. 2003). 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 (van Straten et al. 1997) provide regression weights to allow for the calculation of estimated SIP scores from SA-SIP-30 scores. 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). However, the scoring shift toward healthier outcomes has not been entirely supported in subsequent study. Van de Port et al. (2004) demonstrated a shift toward healthier outcomes only on the physical dimension of the SA-SIP-30 and even then, the trend was less notable than on the SIP68.

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 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. The SA-SIP-30 has demonstrated moderate responsiveness in a longitudinal study.

Table 21.37 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)*

21.4.6 Stroke Impact Scale (SIS)

The Stroke Impact Scale (SIS) 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-pt 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. 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 indicates "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). 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.

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 the majority of information currently available originates from the scales' authors.

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 21.38 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)

21.4.7 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 (Williams et al. 1999a, Kelly-Hayes 2000).

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. 1999a). 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. 1999b).

Advantages

The method of development used assured content validity and a patient-based measure of meaning to stroke patients (Williams et al. 1999). A Danish version of the scale is currently being developed and evaluated (Muus et al. 2005).

Limitations

The SSQOL is a new scale and not well studied. It has not been tested among severe stroke populations.

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 – though in the studies cited here, it was administered by interview.

Table 21.39 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)

21.5 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 21.23, 21.24 & 21.25 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.

21.5.1 Evaluation Summaries by ICF Category

Tables 21.40, 21.41 and 21.42 present a summary of the evaluation undertaken for measures in each ICF category. *Please note, for each table: +++=Excellent; ++=Adequate; +=Poor; n/a = insufficient information; TR=Test re-test; IC= internal consistency; IO = Interobserver; varied (re. floor/ceiling effects; mixed results.*

Table 21.40 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
Canadian Neurological Scale	+	++(IO) +++(IC)	++	+++	+	+	n/a
Clock Drawing Test	++	+++(TR) ++ (IO)	+++	++	n/a	n/a	n/a
Fugl-Meyer Assessment	+++	+++(TR) +++(IO) ++ (IC-balance)	+++	+++ (problems with balance & sensation)	++	++ (sensation)	+(sensation)
General Health Questionnaire - 28	+	+++ (IC)	+++	+++	n/a	n/a	n/a
Geriatric Depression Scale	+++	+++(TR) +++(IC)	+++	+++	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	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)	+++	+++	+	+	+
Orpington Prognostic Scale	+	+++(TR) +++(IO)	++	++	N/a	n/a	n/a

Table 21.41 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
Chedoke-McMaster Stroke Assessment Scale	+	+++(TR) +++ (IO) +++ (IC)	+	+++	+	+++	n/a
Clinical Outcomes Variables Scale	+	+++(TR) +++ (IO) ++ (IC)	++	+++	+	+	n/a
Functional Ambulation Categories	+	+ (TR) +++(IO)	+	++	+	n/a	+
Functional Independence Measure	+++	+++(TR) +++ (IO) +++ (IC)	+++	++	+++	++	++

Outcome Measure	Reliability		Validity		Responsiveness		
	Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
Frenchay Activities Index	+++	++ (TR) ++ (IO) +++ (IC)	+++	+++	+	+	+++
Modified Rankin Handicap Scale	++	+++(TR) ++ (IO)	++	+++	+	++	+
Motor Assessment Scale	++	+++(TR) +++ (IO)	++	++	+	+	+
Nine Hole Peg Test	++	++(TR) +++(IO)	++	+++	+	+	+
Timed "Up & Go"	++	+++(TR) +++ (IO)	+++	+++	+	+	n/a

Table 21.42 Evaluation Summary – Participation/Handicap Outcome Measures

Outcome Measure	Reliability		Validity		Responsiveness		
	Rigor	Results	Rigor	Results	Rigor	Results	Floor/ceiling
London Handicap Scale	+	+++(TR)	+	+++	+	++	+
Medical Outcomes Study Short Form 36	+++	++ (TR) ++ (IC)	+++	+++	++	+++	+
Nottingham Health Profile	+++	++ (TR) ++(IC)	+++	++	+	n/a	+ (ceiling)
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

Bibliography

Aben I, Verhey F, Lousberg R, Lodder J, Honig A. 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 2002;43:386-393.

Adams HP, Jr., Davis PH, Leira EC, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). Neurology 1999;53:126-131.

Adunsky A, Fleissig Y, Levenkrohn A, Arad M, Noy S. Clock drawing task, mini mental state examination and cognitive-functional independence measure: relation to functional outcome of stroke patients. Archives of Gerontology and Geriatrics 2002;35:153-160.

Agrell B, Dehlin O. Comparison of six depression rating scales in geriatric stroke patients. Stroke 1989;20:1190-1194.

Agrell B, Dehlin O. Mini Mental State Examination in geriatric stroke patients. Validity, differences between subgroups of patients, and relationships to somatic and mental variables. Aging Clin Exp Res 2000;12:439-444.

Allison SC, Abraham LD, Peterson CL. Reliability of the Modified Ashworth Scale in the assessment of plantarflexor muscle spasticity in patients with traumatic brain injury. Journal of International Rehabilitation Research 1996;19:67-78.

Almeida OP, Almeida SA. 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. International Journal of Geriatric Psychiatry 1999;14:858-865.

Anemaet WK. Using standardized measures to meet the challenge of stroke assessment. Top Geriatr Rehabil 2002;18:47-62.

Andersen HS, Sestoft D, Lillebaek T, Gabrielsen G, Hemmingsen R. 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 2002;25:573-580.

Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health-related quality of life. Quality of Life Research 1993; 2: 369-395.

Anderson C, Laubscher S, Burns R. Validation of the Short Form 36 (SF-36) health survey questionnaire among stroke patients. Stroke 1996;27:1812-1816.

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

Andresen EM. Criteria for assessing the tools of disability outcomes research. Archives of Physical Medicine and Rehabilitation 2000;81:S15-S20.

Andresen EM, Meyers AR. Health-related quality of life outcomes measures. Archives of Physical Medicine and Rehabilitation 2000;81:S30-S45.

Andresen EM, Gravitt GW, Aydelotte ME, Podgorski CA. Limitations of the SF-36 in a sample ot nursing home residents. Age and Ageing 1999;28:562-566.

Ashworth B. Preliminary trial of carisoprodal in multiple sclerosis. Practitioner 1964;192:540-542.

Au-Yeung SSY, Ng JTW, Lo SK. Does balance or motor impairment of limbs discriminate the ambulatory status of stroke survivors? Am J Phys Med Rehabil 2003;82:279-283.

Aylard PR, Gooding JH, McKenna PJ, Snaith RP. A validation study of three anxiety and depression self-assessment scales. J Psychosom Res 1987;31:261-268.

Backman C, Cork S, Gibson D, Parsons J. Assessment of hand function: the relationship between pegboard dexterity and applied dexterity. Canadian Journal of Occupational Therapy1992 Oct; 59:208-213.

Bamford JM, Sandercock P, Warlow CP, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients (letter). Stroke 1989;20:828.

Banks MH. Validation of the General Health Questionnaire in a young community sample. Psychol Med 1983;13:349-353.

Barclay-Goddard R. Physical function outcome measurement in acute neurology. Physiotherapy Canada 2000;52:138-145.

Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J. Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. Phys Ther 2006;86:245-253.

Beck AT, Ware CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Archives of General Psychiatry 1961;4:561-571.

Beck AT, Beck RW. Screening depressed patients in family practice: A rapid technique. Postgraduate Medicine 1972;52:85.

Beck AT, Steer RA, Garbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical Psychology Review 1988;8:77-100.

Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. Arch Phys Med Rehabil 2006;87:32-39.

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

Benjamin S, Lennon S, Gardner G. The validity of the General Health Questionnaire for first-stage screening for mental illness in pain clinic patients. Pain 1991;47:197-202.

Berg KO, Wood-Dauphinee S, Williams JL, Maki B. Measuring balance in the elderly: preliminary development of an instrument. Physiotherapy Canada 1989;41:304-311.

Berg KO, Wood-Dauphinee S, Williams JL, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;83:S7-S11.

Berg KO, Maki B, Williams JL, Holliday PJ, Wood-Dauphinee S. Clinical and laboratory measures of postural balance in an elderly population. Arch Phys Med Rehabil 1992;73:1073-1080.

Berg KO, Wood-Dauphinee S, Williams JL. The Balance Scale: Reliability assessment with elderly residents and patients with acute stroke. Scan J Rehab Med 1995;27:27-36.

Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. International Journal of Health Services 1976;6:393.

Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Medical Care 1981;19:787-805.

Blackburn M, van Vliet P MS. Reliability of measurements obtained with the Modified Ashworth Scale in the lower extremities of people with stroke. Physical Therapy 2002;82:25-34.

Blake H, McKinney M, Treece K, Lee E, Lincoln NB. An evaluation of screening measures for cognitive impairment after stroke. Age and Ageing 2002;31:451-456.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1986;Feb 8 :307-310.

Bleecker ML, Bolla-Wilson K, Kawas C, Agnew J. Age-specific norms for the Mini Mental State Examination. Neurology 1988;38:1565-1568.

Bluvol A, Ford-Gilboe M. Hope, health work and quality of life in families of stroke survivors. J Adv Nurs 2004;48:322-332.

Bogle-Thorbahn DB, Newton RA. Use of the Berg Balance Test to predict falls in elderly persons. Phys Ther 1996;76:576-583.

Bohannon RW, Smith MB. Inter rater reliability of a modified Ashworth Scale of muscle spasticity. Physical Therapy 1987;67:206-207.

Bouska MJ, Kwatney E. Manual for the application of the motor-free visual perception test to the adult population. Philadelphia, PA: Temple University Rehabilitation Research and Training Centre, 1982.

Bowling A. Measuring Health: A review of quality of life measurement scales. 2nd ed. Philadelphia, PA: Open University Press: 1997.

Brashear A, Zafone R, Corcoran M, et al. Inter and intrarater reliability of the disability assessment scale in patients with upper limb post-stroke spasticity. Archives of Physical Medicine and Rehabilitation 2002;83:1349-1354.

Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992;305:160-164.

Brazier JE, Jones N, Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. Quality of Life Research 1993;2:169-180.

Brazier JE, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and the Euroqol on an elderly population. Quality of Life Research 1996;5:195-204.

Brenner MH, Curbow B, Legro MW. The proximal-distal continuum of multiple health outcome measures: the case of cataract surgery. Medical Care 1995;33:AS236-244.

Bridges KW, Goldberg DP. The validation of the GHQ-28 and the use of the MMSE in neurological in-patients. Br J Psychiatry 1986;148:548-553.

Brock JA, Goldie PA, Greenwood KM. Evaluating the effectiveness of stroke rehabilitation: choosing a discriminative measure. Arch Phys Med Rehabil 2002;83:92-99.

Brott T, Adams HP, Jr., Olinger CP, et al. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 1989;20:864-870.

Brown GT, Rodger S, Davis A. Motor-free visual perception test -- revised: an overview and critique. British Journal of Occupational Therapy 2003;66:159-167.

Buck D, Jacoby A, Massey A, Ford G. Evaluation of Measures used to assess quality of life after stroke. Stroke 2000;31:2004-2010.

Bureau-Chalot F, Novella JL, Jolly D, Ankri J, et al. Feasibility, acceptability and internal consistency reliability of the Nottingham Health Profile in dementia patients. Gerontology 2002; 48(4):220–225.

Burke WJ, Roccaforte WH, Wengel SP, Conley DM, Potter JF. The reliability and validity of the geriatric depression rating scale administered by telephone. Journal of the American Geriatrics Society 1995;43:674-679.

Burvill PW, Knuiman MW. Which version of the General Health Questionnaire should be used in community studies? Aust N Z J Psychiatry 1983;17:237-242.

Bushnell CD, Johnston DC, Goldstein LB. Retrospective assessment of initial stroke severity: comparison of the NIH Stroke Scale and the Canadian Neurological Scale. Stroke 2001;32:656-660.

Callahan CD, Young PL, Barisa MT. Using the SF-36 for longitudinal outcomes assessment in rehabilitation. Rehabiliation Psychology 2005;50:65-70.

Cardol M, Brandsma JW, de Groot IJM, van den Bos GAM, de Haan RJ, de Jong BA. Handicap questionnaires: what do they assess? Disability and Rehabilitation 1999;21:97-105.

Carr JH, Shepherd RB, Nordholm L, Lynne D. Investigation of a new motor assessment scale for stroke patients. Phys Ther 1985;65:175-180.

Carroll D. A quantitative test of upper extremity function. J Chronic Dis 1965;18:479-491.

Carter BS, Buckley D, Ferraro R, Rordorf G, Ogilvy CS. Factors associated with reintegration to normal living after subarachnoid hemorrhage. Neurosurgery 2000;46:1326-1333.

Carter J, Mant F, Wade DT, Winner S. Comparison of the postal version of the Frenchay Activities Index with interviewer-administered version for use in people with stroke. Clin Rehabil 1997;11:131-138.

Cate Y, Richards L. Relationship between performance on tests of basic visual functions and visual-perceptual processing in persons after brain injury. American Journal of Occupational Therapy 2000;54:326-334.

Cavanagh SJ, Hogan K, Gordon V, Fairfax J. Stroke-specific FIM models in an urban population. Journal of Neurological Nursing 2000;32:17-21.

Chae J, Labatia I, Yang G. Upper limb motor function in hemiparesis: concurrent validity of the Arm Motor Ability Test. Americal Jouranl of Physical Medicine and Rehabilitation 2003;82:1-8.

Chong DK. Measurement of instrumental activities of daily living in stroke. Stroke 1995;26:1119-1122.

Chou CY, Chien CW, Hsueh IP, Sheu CF, Wang CH, Hsieh CL. Developing a short form of the Berg Balance Scale for people with stroke. Phys Ther 2006;86:195-204.

Cifu DX, Keyser-Marcus L, Lopez E, et al. Acute predictors of successful return to work 1 year after traumatic brain injury: a multicenter analysis. Arch Phys Med Rehabil 1997;78:125-131.

Clarke PJ, Black SE, Badley EM, Lawrence JM, Williams JI. Handicap in stroke survivors. Disabil Rehabil 1999;21:116-123.

Cohen MI, Marino RJ. The tools of disability outcomes research functional status measures. Archives of Physical Medicine and Rehabilitation 2000;81:S21-S29.

Colarusso RP, Hammill DD. Motor-free Visual Perception Test. Novato, CA: Academic Therapy Publications, 1972.

Colarusso RP, Hammill DD. Motor-free visual perception test--revised. Novato CA: Academic Therapy Publications, 1996.

Cole B, Finch E, Gowland C, Mayo N. Physical rehabilitation outcome measures. Toronto, Ontario: Canadian Physical Therapy Association, 1994.

Collen FM, Wade DT, Bradshaw CM. Mobility after stroke: reliability of measures of impairment and disability. Int Disabil Stud 1990;12:6-9.

Collin C, Davies S, Horne V, Wade DT. Reliability of the Barthel ADL Index. Int J Rehabil Res 1987;10:356-357.

Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. Int Disabil Stud 1988;10:61-63.

Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000; 17(1): 13-35.

Corrigan JD, Smith-Knapp K, Granger C. Validity of the Functional Independence Measure for persons with traumatic brain injury. Arch Phys Med Rehabil 1997;78:828-834.

Cote R, Hachinski VC, Shurvell BL, Norris JW, Wolfson C. The Canadian Neurological Scale: a preliminary study in acute stroke. Stroke 1986;17:731-737.

Cote R, Battista RN, Wolfson C, Boucher J, Adam J, Hachinski V. The Canadian Neurological Scale: validation and reliability assessment. Neurology 1989;39:638-643.

Croarkin E, Danoff J, Barnes C. Evidence-based rating of upper-extremity motor function tests used for people following a stroke. Phys Ther 2004;84:62-74.

Cunha IT, Lim PA, Henson H, Monga T, Qureshy H, Protas EJ. Performance-based gait tests for acute stroke patients. Am J Phys Med Rehabil 2002;81:848-856.

Cup EHC, Scholte op Reimer WJM, Thijssen MCE, van Kuyk-Minis MAH. Reliability and validity of the Canadian Occupational Performance Measure in stroke patients. Clinical Rehabilitation 2003;17:402-409.

Cusick CP, Berhart KA, Mellick DC. Participant-proxy reliability in traumatic brain injury outcome research. J Head Trauma Rehabil 2000;15:739-749.

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

Czechowsky D, Hill MD. Neurological outcome and quality of life after stroke due to vertebral artery dissection. Cerebrovasc Dis 2002;13:192-197.

D'Ath P, Katona P, Mullan E, Evans S, Katona C. Screening, detection and management of depression in elderly primary care attenders 1: The acceptability and performance of the 15-item Geriatric Depression Scale (GDS15) and the development of short versions. Family Practice 1994;266.

D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies in stroke. Stroke 1996;27:2331-2336.

da CI, Jr., Lim PA, Qureshy H, Henson H, Monga T, Protas EJ. Gait outcomes after acute stroke rehabilitation with supported treadmill ambulation training: a randomized controlled pilot study. Arch Phys Med Rehabil 2002;83:1258-1265.

Damiano DL, Quinlivan JM, Owen BF, Payne P, Nelson KC, Abel MF. What does the Ashworth Scale really measure and are instrumented measures more valid and precise? Developmental Medicine and Child Neurology 2002;44:112-118.

Daneski K, Coshall C, Tilling K, Wolfe CD. Reliability and validity of a postal version of the Reintegration to Normal Living Index, modified for use with stroke patients. Clin Rehabil 2003;17:835-839.

Daving Y, Andren E, Nordholm L, Grimby G. Reliability of an interview approach to the Functional Independence Measure. Clin Rehabil 2001;15:310.

Davis J, Kayser J, Matlin P, Mower S, Tadano P. Clinical analysis. Nine-hole peg tests: are they all the same? OT Practice1999 Apr; 4:59-61.

Dean C, Mackey F. Motor assessment scale scores as a measure of rehabilitation outcome follwoing stroke. Australian Journal of Physiotherapy 1992;38:31-35.

de Bruin AF, de Witte LP, Stevens F, Diederiks JPM. Sickness Impact Profile: The state of the art of a generic functional status measure. Soc Sci Med 1992;35:1003-1014.

de Haan R, Horn J, Limburg M, Van Der MJ, Bossuyt P. A comparison of five stroke scales with measures of disability, handicap, and quality of life. Stroke 1993;24:1178-1181.

de Haan RJ, Limburg M, Bossuyt P, van der Meulen J, Aaronson NK. The clinical meaning of Rankin 'handicap' grades after stroke. Stroke 1995;26:2027-2030.

de Haan R, Aaronson N, Limburg M, Langton-Hewer R et al. Measuring quality of life in stroke. Stroke 1993; 24: 320 – 327.

de Koning I, van Kooten F, Koudstaal PJ. Value of screening instruments in the diagnosis of post-stroke dementia. Haemostasis 1998;28:158-166.

Desrosiers J, Noreau L, Rochette A, Bravo G, Boutin C. Predictors of handicap situations following post-stroke rehabilitation. Disability and Rehabilitation 2002;24:774-785.

Dettman MA, Linder MT, Sepic SB. Relationships among walking performance, postural stability and functional assessments of the hemiplegic patient. American Journal of Physical Medicine and Rehabilitation 1987;66:77-90.

deWeerdt WJ, Harrison M. 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 1985;37:65-70.

Dewey HM, Donnan GA, Freeman EJ, et al. Interrater reliability of the National Institutes of Health Stroke Scale: rating by neurologists and nurses in a community-based stroke incidence study. Cerebrovasc Dis 1999;9:323-327.

Dick JPR, Builoff RJ, Stewart A, et al. Mini-mental State Examination in neurological patients. J Neurol Neurosurg Psychiatry 1984;47:499.

Dodds TA, Martin DP, Stolov WC, Deyo R. A validation of the Functional Independence Measurement and its performance among rehabilitation inpatients. Arch Phys Med Rehabil 1993;74:531-536.

Dorevitch MI, Cossar RM, Bailey FJ, et al. The accuracy of self and informant ratings of physical functional capacity in the elderly. J Clin Epidemiol 1992;45:791-798.

Dorman PJ, Slattery J, Farrell B, Dennis M, Sandercock P, for the United Kingdom Collaborators in the International Stroke Trial (IST). A randomised comparison of the EuroQol and Short Form 36 after stroke. BMJ 1997;315:461.

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

Dorman PJ, Dennis M, Sandercock P, for the United Kingdom Collaborators in the International Stroke Trial (IST). How do scores on the EuroQol relate to scores on the SF-36 after stroke? Stroke 1999;30:2146-2151.

Dromerick AW, Edwards DF, Diringer MN. Sensitivity to changes in disability after stroke: A comparison of four scales useful in clinical trials. Journal of Rehabilitation Research and Development 2003;40:1-8.

Dubuc N, Haley S, Ni P, Kooyoomjian J, Jette A. 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 2004;26:362-370.

Duncan PW, Samsa G, Weinberger M, et al. Health status of individuals with mild stroke. Stroke 1997;28:740-745.

Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The Stroke Impact Scale version 2.0. Evaluation of reliability, validity and sensitivity to change. Stroke 1999;30:2131-2140.

Duncan PW, Jorgensen HS, Wade DT. Outcome measures in acute stroke trials. A systematic review and some recommendations to improve practice. Stroke 2000;31:1429-1438.

Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. Stroke 2002;33:2593-2599.

Duncan PW, Reker DM, Horner RD, et al. Performance of a mail-administered version of a stroke-specific outcome measure: the Stroke Impact Scale. Clin Rehabil 2002;16:505.

Duncan PW, Lai SM, Bode RK, Perera S, De Rosa J, GAIN Americas Investigators. Stroke Impact Scale-16: A brief assessment of physical function. Neurology 2003;60:291-296.

Duncan PW, Bode RK, Lai SM, Perera S. Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. Arch Phys Med Rehabil 2003;84:950-963.

Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Physical Therapy 1983;63:1606-1610.

Duncan PW, Goldstein LB, Matchar D, Divine GW, Feussner J. Measurement of motor recovery after stroke: outcome assessment and sample size requirements. Stroke 1992;23:1084-1089.

Duncan PW, Goldstein LB, Horner RD, Landsman PB. Similar motor recovery of upper and lower extremities after stroke. Stroke 1994;25:1181-1188.

Ebrahim S, Barer D, Nouri F. Use of the Nottingham Health Profile with patients after a stroke. Journal of Epidemiology and Community Health 1986;40:166-169.

Edwards B, O'Connell B. Internal consistency and validity of the Stroke Impact Scale 2.0 (SIS 2.0) and SIS-16 in an Australian sample. Quality of Life Research 2003;12:1135.

Eng JJ RS, McLaren LM. Mobility status during inpatient rehabilitation: A comparison of patients with stroke and traumatic brain injury. Arch Phys Med Rehabil 2002;83:483-490.

English CK, Hillier SL, Stiller K, Warden-Flood A. The sensitivity of three commonly used outcome measures to detect change amongst patients receiving inpatient rehabilitation following stroke. Clin Rehabil 2006;20:52-55.

Essink-Bot ML, Drabbe PF, Bonsel GJ, Aaronson NK. 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. Med Care 1997;35:522-537.

Fabrigoule C, Lechevallier N, Crasborn L, Dartigues JF, Orgogozo JM. Inter-rater reliability of scales and tests used to measure mild cognitive impairment by general practitioners and psychologists. Current Medical Research and Opinion 2003;19:603-608.

Filiatrault J, Arsenault AB, Dutil E, Bourbonnais D. Motor function and activities of daily living assessments: a study of three tests for persons with hemiplegia. Am J Occup Ther 1991;45:806-810.

Finch E, Brooks D, Stratford PW, Mayo NE. Physical Rehabilitations Outcome Measures. A Guide to Enhanced Clinical Decision-Making., 2nd ed. Toronto, Ontario: Canadian Physiotherapy Association, 2002.

Fiorelli M, Alperovitch A, Argentino C, et al. Prediction of long-term outcome in the early hours following acute ischemic stroke. Italian Acute Stroke Study Group. Arch Neurol 1995;52:250-255.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluation of patient-based outcome measures for use in clinical trials. Health Technology Assessment 1998;2:1-74.

Flansbjer UB, Holmback AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. J Rehabil Med 2005;37:75-82.

Folstein MF, Folstein SE, McHugh PR. A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:198.

Fortinsky RH, Granger CV, Seltzer GB. The use of functional assessment in understanding home care needs. Medical Care 1981;19:489-497.

Freidl W, Schmidt R, Stronegger WJ, Fazekas F, Reinhart B. Sociodemographic predictors and concurrent validity of the Mini Mental State Examination and the Mattis Dementia Rating Scale. Eur Arch Psychiatry Clin Neurosci 1996;246:317-319.

Friedman PJ. Clock drawing in acute stroke. Age & Ageing 1991;20:140-145.

Frytak J. Measurement. Journal of Rehabilitation Outcomes Measurement 2000;4:15-31.

Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient: I. A method for evaluation of physical performance. Scandinavian Journal of Rehabilitation Medicine 1975;7:13-31.

Fugl-Meyer AR. Post-stroke hemiplegia assessment of physical properties. Scandinavian Journal of Rehabilitation Medicine 1980;7:85-93.

Furukawa TA, Goldberg DP, Rabe-Hesketh S, Ustun TB. Stratum-specific likelihood ratios of two versions of the general health questionnaire. Psychol Med 2001;31:519-529.

Gladstone DJ, Danells CJ, Black SE. The Fugl-Meyer Assessment of Motor Recovery after Stroke: A critical review of its measurement properties. Neurorehabilitation and Neural Repair 2002;16:232-240.

Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. Psychol Med 1979;9:139-145.

Goldberg DP, Gater R, Sartorius N, et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol Med 1997;27:191-197.

Goldberg DP, Oldehinkel T, Ormel J. Why GHQ threshold varies from one place to another. Psychol Med 1998;28:915-921.

Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. Arch Neurol 1989;46:660-662.

Goldstein LB, Chilukuri V. Retrospective assessment of initial stroke severity with the Canadian Neurological Scale. Stroke 1997;28:1181-1184.

Goldstein LB, Samsa GP. Reliability of the National Institutes of Health Stroke Scale. Extension to non-neurologists in the context of a clinical trial. Stroke 1997;28:307-310.

Golomb BA, Vickrey BG, Hays RD. A review of health-related quality-of-life measures in stroke. Pharmacoeconomics 2001;19:155-185.

Gompertz P, Pound P, Ebrahim S. The reliability of stroke outcome measures. Clinical Rehabilitaiton 1993; 7: 290 – 296.

Goodchild ME, Duncan-Jones P. Chronicity and the General Health Questionnaire. Br J Psychiatry 1985;146:55-61.

Gosman-Hedstrom G, Svensson E. Parallel reliability of the Functional Independence Measure and the Barthel ADL Index. Disabil Rehabil 2000;22:702-715.

Gowland C, Stratford PW, Ward M, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. Stroke 1993;24:58-63.

Gowland C, Van Hullenaar S, Torresin W, Moreland J, et al. Chedoke-McMaster Stroke Assessment. Development, validation and administration manual. Hamilton, Ontario: Chedoke-McMaster Hospitals and McMaster University, 1995.

Grace J, Nadler JD, White DA, et al. Folstein vs. Modified Mini-Mental State Examination in geriatric stroke. Arch Neurol 1995;52:477-484.

Granger CV, Devis LS, Peters MC, Sherwood CC, et al. Stroke rehabilitation: analysis of repeated Barthel index measures. Arch Phys Med Rehabil 1979;60:14-17.

Granger CV, Cotter AC, Hamilton BB, Fiedler RC. Functional Assessment Scales: a study of persons after stroke. Arch Phys Med Rehabil 1993;74:133-138.

Green J, Forster A, Young I. 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. Disabil Rehabil 2001;23:670-676.

Gregson JM, Leathley M, Moore AP, Sharma AK, Smith TL, Watkins CL. Reliability of the Tone Assessment Scale and the Modified Ashworth Scale as clinical tool for assessing poststroke spasticity. Archives of Physical Medicine and Rehabilitation 1999;1999:1013-1016.

Gregson JM, Leathley M, Moore AP, Smith TL, et al. Reliability of measurement of muscle tone and muscle power in stroke patients. Age and Aging 2000;29:223-228.

Haas BM, Bergstrom E, Jamous A, Bennie A. The inter rater reliability of the original and of the modified Ashworth scale for the assessment of spasicity in patients with spinal cord injury. Spinal Cord 1996;34:560-564.

Hachisuka K, Ogata H, Ohkuma H, Tanaka S, Dozono K. Test-retest and inter-method reliability of the self-rating Barthel Index. Clinical Rehabilitation 1997;11:28-35.

Hajek VE, Gagnon S, Ruderman JE. Cognitive and functional assessments of stroke patients: an analysis of their relation. Archives of Physical Medicine and Rehabilitation 1997;78:1331-1337.

Hamilton BB, Granger CV, Sherwin FS, et al. A uniform national data system for medical rehabilitation. In:Fuhrer MJ, ed. Rehabilitation Outcomes: Analysis and Measurement. Baltimore, Maryland: Paul H. Brooks, 1987:137-147.

Hammond FM, Grattan KD, Sasser H, Corrigan JD, Bushnik T, Zafonte RD. Long-term recovery course after traumatic brain injury: A comparison of the Functional Independence Measure and Disabilty Rating Scale. J Head Trauma Rehabil 2001;16:318-329.

Harwood RH, Gompertz P, Ebrahim S. Handicap one year after a stroke: validity of a new scale. J Neurol Neurosurg Psychiatry 1994;57:825-829.

Harwood RH, Rogers A, Dickinson E, Ebrahim S. Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. Qual Health Care 1994;3:11-16.

Harwood RH, Ebrahim S. Measuring the outcomes of day hospital attendance: a comparison of the Barthel Index and London Handicap Scale. Clin Rehabil 2000;14:527-531.

Harwood RH, Ebrahim S. A comparison of the responsiveness of the Nottingham extended activities of daily living scale, London handicap scale and SF-36. Disabil Rehabil 2000;22:786-793.

Hayes V, Morris J, Wolfe C, Morgan M. The SF-36 Health Survey Questionnaire: Is it suitable for use with older adults? Age and Aging 1995;24:120-125.

Heinemann AW, Harvey RL, McGuire JR, et al. Measurement properties of the NIH Stroke Scale during acute rehabilitation. Stroke 1997;28:1174-1180.

Heinik J, Solomesh I, Berkman P. Correlation between the CAMCOG, the MMSE and three clock drawing tests in a specialized outpatient psychogeriatric service. Arch Gerontol Geriatr 2004;38:77-84.

Heller A, Wade DT, Wood VA, Sunderland A, Hewer RL, Ward E. Arm function after stroke: measurement and recovery over the first three months. J Neurol Neurosurg Psychiatry 1987;50:714-719.

Helm HW, Boward MD. Factor structure of the Beck Depression Inventory in a university sample. Psychological Reports 2003;92:53-61.

Herrmann N, Mittmann N, Silver IL, et al. A validation study of the Geriatric Depression Scale short form. International Journal of Geriatric Psychiatry 1996;11:457-460.

Herrmann N, Black SE, Lawrence J, Szekely C, Szalai JP. The Sunnybrook Stroke Study. A prospective study of depressive symptoms and functional outcome. Stroke 1998;29:618-624.

Hesse S, Bertelt C, Schaffrin A, Malezic M, Mauritz KH. Restoration of gait in nonambulatory hemiparetic patients by treadmill training with partial body-weight support. Arch Phys Med Rehabil 1994;75:1087-1093.

Hill K, Ellis P, Bernhardt J, Maggs P, Hull S. Balance and mobility outcomes for stroke patients: a comprehensive audit. Aust J Physiother 1997;43:173-180.

Hobart JC, Lamping DL, Freeman JA, et al. Evidence-based measurement. Which disability scale for neurologic rehabilitation? Neurology 2001;57:644.

Hobart JC, Thompson AJ. The five item Barthel Index. J Neurol Neurosurg Psychiatry 2001;71:225-230.

Hobart JC, Williams LS, Moran K, Thompson AJ. Quality of life measurement after stroke. Uses and abuses of the SF-36. Stroke 2002;33:1348-1356.

Hocking C, Williams M, Broad J, Baskett J. Sensitivity of Shah, Vanclay and Cooper's Modified Barthel Index. Clin Rehabil 1999;13:141-147.

Holbrook M, Skilbeck CE. An activities index for use with stroke patients. Age and Aging 1983;12:166-170.

Holden MK, Gill KM, Magliozzi MR, Nathan J, Piehl-Baker L. Clinical gait assessment in the neurologically impaired. Reliability and meaningfulness. Phys Ther 1984;64:35-40.

Holden MK, Gill KM, Magliozzi MR. Gait assessment for neurologically impaired patients. Standards for outcome assessment. Phys Ther 1986;66:1530-1539.

Horgan F, Cunningham CJ, O'Neill D, et al. Validating the Orpington Prognostic Score in an Irish in-patient stroke population. Irish Medical Journal 2005;98:172-175.

House A, Dennis M, Mogridge L, Warlow C, Hawton K, Jones L. Mood disorders in the year after first stroke. The British Journal of Psychiatry 1991;158:83-92.

Hsieh CL, Hsueh IP, Mao HF. Validity and responsiveness of the Rivermead Mobility Index in stroke patients. Scan J Rehab Med 2000;32:140-142.

Hsieh CH, Hsueh IP, Chiang FU, Lin PH. Inter-rater reliability and validity of the action Research Arm test in stroke patients. Age Ageing 1998;27:107-113.

Hsueh IP, Lee MM, Hsieh CL. Psychometric characteristics of the Barthel activities of daily living index in stroke patients. Journal of the Formosa Medical Association 2001;100:526-532.

Hsueh IP, Lin JH, Jeng JS, Hsieh CL. 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 2002;73:188-190.

Hsueh IP, Lee MM, Hsieh CL. The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? Clin Rehabil 2002;16:382-388.

Hsueh IP, Hsieh CL. Responsiveness of two upper extremity function instruments for stroke inpatients receiving rehabilitation. Clin Rehabil 2002;16:617-624.

Hsueh IP, Wang CH, Sheu CF, Hsieh CL. Comparison of psychometric properties of three mobility measures for patients with stroke. Stroke 2003;34:1741-1745.

Huijbregts M. The physiotherapy clinical outcome variables (COVS) reliability testing videotape. Physiotherapy Canada 1996;48:285.

Hunt SM, McKenna SP, McEwen J, Backett EM, et al. A quantitative approach to perceived health status: a validation study. Journal of Epidemiology and Community Health 1980; 34: 281-286.

Hunt SM, McKenna SP, McEwen J, Williams J et al. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med 1981;15:221– 29.

Hunt SM, McEwan J, McKenna SP. Perceived health: age and sex comparisons in a community. Journal of Epidemiology and Community Health 1984;34:281–286.

Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985;35:85–188.

Hunt SM, McKenna SP, McEwan J. The Nottingham Health Profile: user's manual. Revised ed. Manchester, England, 1989.

Incalzi RA, Cesari M, Pedone C, Carbonin PU. Construct validity of the 15-item geriatric depression scale in older medical inpatients. J Geriatr Psychiatry Neurol 2003;16:23-28.

Jacob-Lloyd HA, Dunn OM, Brain ND, Lamb SE. Effective measurement of the functional progress of stroke clients. British Journal of Occupational Therapy2005 Jun; 68:253-259.

Jansa J, Pogacnik T, Gompertz P. An evaluation of the extended Barthel Index with acute ischemic stroke patients. Neurorehabilitation and Neural Repair 2004;18:37-41.

Jenkinson C, Fitzpatrick R, Argyle M. The Nottingham Health Profile: An analysis of its sensitivity in differentiating illness groups. Soc Sci Med 1988; 27(12): 1411-1414.

Jenkinson C, Fitzpatrick R. Measurement of health status in patients with chronic illness: comparison of the Nottingham Health Profile and the General Health Questionnaire. Family Practice 1990;7(2):121– 24.

Jenkinson C. Why are we weighting? A critical examination of the use of item weights in a health status measure. Soc Sci Med 1991;32(12):1413–1416.

Jenkinson C, Mant J, Carter J, Wade D, Winner S. The London handicap scale: a re-evaluation of its validity using standard scoring and simple summation. J Neurol Neurosurg Psychiatry 2000;68:365-367.

Johnson G, Burvill PW, Anderson CW, Jamrozik K, Stewart-Wynne EG, Chakera TMH. Screening instruments for depression and anxiety following stroke: experience in the Perth community stroke study. Acta Psychiatrica Scandinavica 1995;91:252-257.

Jones RN, Gallo JJ. Dimensions of the Mini-Mental State Examination among community dwelling older adults. Psychological Medicine 2000;30:605-618.

Jorgensen HS, Nakayama H, Raaschou HO, Vive-Larsen J, Stoier M, Olsen TS. Outcome and time course of recovery in stroke. Part II: time course of recovery. The Copenhagen stroke study. Archives of Physical Medicine and Rehabilitation 1995;76:406-412.

Juby A. Correlation between the Folstein Mini-Mental State Examination and three methods of clock drawing scoring. J Geriatr Psychiatry Neurol 1999;12:87-91.

Juneja G, Czyrny JJ, Linn RT. Admission balance and outcomes of patients admitted for acute inpatient rehabilitation. Am J Phys Med Rehabil 1998;77:388-393.

Jutai J, Teasell R. The necessity and limitations of evidence-based practice in stroke rehabilitation. Topics in Stroke Rehabilitation 2003;10:71-78.

Kalra L, Crome P. The role of prognostic scores in targeting stroke rehabilitation in elderly patients. Journal of the American Geriatrics Society 1993;41:396-400.

Kalra L, Dale P, Crome P. Evaluation of a clinical score for prognostic stratification of elderly stroke patients. Age and Ageing 1994;23:492-497.

Kasner SE, Chalela JA, Luciano JM, et al. Reliability and validity of estimating the NIH stroke scale score from medical records. Stroke 1999;30:1534-1537.

Keller SD, Bayliss MS, Ware JE Jr, Hsu MA, Damiano AM, Goss TF. Comparison of responses to SF-36 health survey questions with one-week and four-week recall periods. HSR: Health Services Research 1997;32:367-384.

Kellor M, Frost J, Silberberg N, Iversen I, Cummings R. Hand strength and dexterity. Am J Occup Ther 1971;25:77-83.

Kelly-Hayes M. Stroke outcomes measurment: new developments. J Rehabil Outcomes Meas 2000;4:57-63.

Kidd D, Stewart G, Baldry J, et al. The Functional Independence Measure: a comparative validity and reliability study. Disabil Rehabil 1995;17:10-14.

Kilic C, Rezaki M, Rezaki B, et al. General Health Questionnaire (GHQ12 & GHQ28): psychometric properties and factor structure of the scales in a Turkish primary care sample. Soc Psychiatry Psychiatr Epidemiol 1997;32:327-331.

Koeter MW. Validity of the GHQ and SCL anxiety and depression scales: a comparative study. J Affect Disord 1992;24:271-279.

Korner-Bitensky N, Wood-Dauphinee S, Siemiatycki J, Shapiro S, Becker R. Health-related information postdischarge: telephone versus face-to-face interviewing. Arch Phys Med Rehabil 1994;75:1287-1296.

Korner-Bitensky N, Wood-Dauphinee S. Barthel information elicited over the telephone. Is it reliable? Am J Phys Med Rehabil 1995;74:9-18.

Korner-Bitensky N, Mazer BL, Sofer S, et al. Visual testing for readiness to drive after stroke: a multicenter study. Americal Journal of Physical Medicine and Rehabilitation 2000;79:259.

Kornetti DL, Fritz SL, Chiu Y-P, Light KE, Velozo CA. Rating scale analysis of the Berg Balance Scale. Archives of Physical Medicine and Rehabilitation 2004;85:1128-1135.

Kotila M, Numminen H, Waltimo O, Kaste M. Depression After Stroke : Results of the FINNSTROKE Study. Stroke 1998;29:368-372.

Kotila M, Numminen H, Waltimo O, Kaste M. Post-stroke depression and functional recovery in a population-based stroke register. The Finnstroke study. European Journal of Neurology 1999;6:309-312.

Kwon S, Hartzema AG, Duncan PW, Lai SM. Disability measures in stroke. Relationship among the Barthel Index, the functional ndependence Measure and the Modified Rankin Scale. Stroke 2004;35:918-923.

Lai SM, Duncan PW. Stroke recovery profile and the Modified Rankin assessment. Neuroepidemiology 2001;20:26-30.

Lai SM, Studenski S, Duncan PW, Perera S. Persisting consequences of stroke measured by the Stroke Impact Scale. Stroke 2002;33:1844.

Lai SM, Perera S, Duncan PW, Bode RK. Physical and social functioning after stroke: Comparison of the Stroke Impact Scale and Short Form-36. Stroke 2003;34:488-493.

Lai SM, Duncan PW, Keightly J. Prediction of functional outcome after stroke. Comparison of the Orpington Prognostic Scale and the NIH Stroke Scale. Stroke 1998;29:1838-1842.

Lannin N. Reliability, validity and factor structure of the upper limb subscale of the Motor Assessment Scale (UL-MAS) in adults following stroke. Disabil Rehabil 2004;26:109-116.

Law M. Measurement in occupational therapy: Scientific criteria for evaluation. Canadian Journal of Occupational Therapy 1987;54:133-138.

Law M. Evidence-based Rehabilitation. A Guide to Practice. Thorofare, NJ: Slack Press, 2002.

Lee KC, Carson L, Kinnin E, Patterson V. The Ashworth Scale: A reliable and reproducible method of measuring spasticity. J Neuro Rehabil 1989;3:205-209.

Lin J-H, Hseueh IP, Sheu CF, Hsieh C-L. Psychometric properties of the sensory scale of the Fugl-Meyer Assessment in stroke patients. Clin Rehabil 2004;18:391-397.

Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994;75:127-132.

Lincoln NB, Nicholl CR, Flannaghan T, Leonard M, Van der GE. The validity of questionnaire measures for assessing depression after stroke. Clin Rehabil 2003;17:840-846.

Lindesay J. Validity of the General Health Questionnaire (GHQ) in detecting psychiatric disturbance in amputees with phantom pain. J Psychosom Res 1986;30:277-281.

Liston R, Brouwer BJ. Reliability and validity of measures obtained from stroke patients using the balance master. Arch Phys Med Rehabil 1996;77:425-430.

Lo R, Harwood R, Woo J, Yeung F, Ebrahim S. Cross-cultural validation of the London Handicap Scale in Hong Kong Chinese. Clin Rehabil 2001;15:177-185.

Lobo A, Perez-Echeverria MJ, Jimenez-Aznarez A, Sancho MA. Emotional disturbances in endocrine patients. Validity of the scaled version of the General Health Questionnaire (GHQ-28). Br J Psychiatry 1988;152:807-812.

Loewen SC, Anderson BA. Reliability of the Modified Motor Assessment Scale and the Barthel Index. Phys Ther 1988;68:1077-1081.

Loewen SC, Anderson BA. Predictors of stroke outcome using objective measurement scales. Stroke 1990;21:78-81.

Lord SE, McPherson K, McNaughton HK, Rochester L, Weatherall M. Community ambulation after stroke: how important and obtainable is it and what measures appear predictive? Arch Phys Med Rehabil 2004;85:234-239.

Lorentz WJ, Scanlan JM, Borson S. Brief screening tests for dementia. Canadian Journal of Psychiatry 2002;47:723-733.

Low Choy N, Kuys S, Richards M, Isles R. Measurement of functional ability following traumatic brain injury using the Clinical Outcomes Variable Scale: A reliability study. Australian Journal of Physiotherapy 2002;48:35-39.

Lukkarinen H, Hentinen M. Assessment of quality of life with the Nottingham Health Profile among patients with coronary heart disease. Journal of Advanced Nursing 1997;26:73 – 84.

Lyden P, Brott T, Tilley B, et al. Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. Stroke 1994;25:2220-2226.

Lyden P, Lu M, Jackson C, et al. Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. NINDS tPA Stroke Trial Investigators. Stroke 1999;30:2347-2354.

Lyden PD, Lu M, Levine SR, Brott TG, Broderick J. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. Stroke 2001;32:1310-1317.

Lykouras L, Adrachta D, Kalfakis N, et al. GHQ-28 as an aid to detect mental disorders in neurological inpatients. Acta Psychiatr Scand 1996;93:212-216.

Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981;4:483-492.

Lyness JM, Noel TK, King DA, Conwell Y, Caine ED. Screening for depression in elderly primary care patients. A comparison of the Center for Epidemiologic Studies - Depression Scale and the Geriatric Depression Scale. Arch Internal Med 1997;157:449-454.

MacNeill SE, Lichtenberg PA. The MacNeill-Lichtenberg Decision Tree: A unique method of triaging mental health problems in older medical rehabilitation patients. Arch Phys Med Rehabil 2000;81:618-622.

Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Maryland State Med J 1965;14:61-65.

Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 health status questionnaire. Soc Sci Med 2002;54:11-21.

Malouin F, Pichard L, Bonneau C, Durand A, Corriveau D. Evaluating motor recovery early after stroke: Comparison of the Fugl-Meyer assessment and the Motor Assessment Scale. Archives of Physical Medicine and Rehabilitation 1994;75:1206-1212.

Malt UF. The Validity of the General Health Questionnaire in A Sample of Accidentally Injured Adults. Acta Psychiatrica Scandinavica 1989;80:103-112.

Mao HF, Hsueh IP, Tang PF, Sheu CF, Hsieh CL. Analysis and comparison of the psychometric properties of three balance measures for stroke patients. Stroke 2002;33:1022-1027.

Mathias S, Nayak USL, Isaacs B. Balance in elderly patients: the "Get up and Go" test. Arch Phys Med Rehabil 1986;67:387-389.

Mathiowetz VW. Adult Norms for the Nine Hole Peg Test of Finger Dexterity. The Occupational Therapy Journal of Research 5[1], 25-38.

Mayo N, Wood-Dauphinee S, Cote R, Durcan L, Carlton J. Activity, participation & quality of life 6 months post-stroke. Arch Phys Med Rehabil 2002;83:1035-1042.

Mayo NE, Wood-Dauphinee S, Ahmed S, et al. Disablement following stroke. Disability and Rehabilitation 1999;21:258-268.

Mazer BL, Korner-Bitensky N, Sofer S. Predicting ability to drive after stroke. Archives of Physical Medicine and Rehabilitation 1998;79:743-750.

McColl MA, Paterson M, Davies D, Doubt L, Law M. Validity and community utility of the Canadian Occupational Performance Measure. Can J Occup Ther 2000;67:22-30.

McDowell I, Newell C. Measuring Health. A Guide to Rating Scales and Questionnaires., 2nd ed. New York: Oxford University Press, 1996.

McGinnis GE, Seward ML, DeJong G, Osberg S. Program evaluation of physical medicine and rehabilitation departments using self-report Barthel. Archives of Physical Medicine and Rehabilitation 1986;67:125.

McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-item short form health survey (SF-36) II: Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care 1993;31:247-263.

McHorney CA. Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 health survey. The Gerontologist 1996;36:571-583.

McPhee SD. Functional hand evaluations: a review. Am J Occup Ther 1987;41:158-163.

Meldrum D, Pittock SJ, Hardiman O, Ni DC, O'Regan M. Recovery of the upper limb post ischaemic stroke and the predictive value of the Orpington Prognostic Score. Clin Rehabil 2004;18:694-702.

Mercier L, Hebert R, Gauthier L. Motor free visual perception test: impact of vertical answer cards position on performance of adults with hemispatial neglect. Occup Ther J Res 1995;15:223-226.

Mercier L, Desrosier J, Hebert R, Rochette A, Dubois MF. Normative data for the Motor-Free Visual Perception Test - Vertical. Physical and Occupational Therapy in Geriatrics 2001;19:39-50.

Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. Stroke 2002;33:1261-1266.

Molloy DW, Standish TIM. A guide to the standardized Mini Mental State Examination. International Psychogeriatrics 1997;9:87-94.

Morris S, Morris ME, Iansek R. Reliability of measurements obtained with the Timed "Up & Go" Test in people with Parkinson Disease. Phys Ther 2001;81:810-818.

Morris PL, Goldberg RJ. Validity of the 28-item General Health Questionnaire in hospitalized gastroenterology patients. Psychosomatics 1989;30:290-295.

Mossberg K, McFarland C. A patient-oriented health status measure in outpatient rehabilitation. Am J Phys Med Rehabil 2001;80:896-902.

Muir KW, Weir CJ, Murray GD, Povey C, Lees KR. Comparison of neurological scales and scoring systems for acute stroke prognosis. Stroke 1996;27:1817-1820.

Murphy MA, Roberts-Warrior D. A review of motor performance measures and treatment interventions for patients with stroke. Topics in Geriatric Rehabilitation2003 Jan-Mar; 19:3-42.

Muus I, Ringsberg KC. Stroke Specific Quality of Life Scale: Danish adaptation and a pilot study for testing psychometric properties. Scand J Caring Sci 2005;19:140-147.

New PW, Buchbinder R. Critical appraisal and review of the Rankin scale and its derivatives. Neuroepidemiology 2006;26:4-15.

Newkirk LA, Kim JM, Thompson JM, Tinklenberg JR, Yesavage JA, Taylor JL. Validation of a 26-Point Telephone Version of the Mini-Mental State Examination. Journal of Geriatric Psychiatry and Neurology 2004;17:81-87.

Ng SS, Hui-Chan CW. 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 2005;86:1641-1647.

Nikolaus T, Bach M, Oster P, Schlierf G. Prospective value of self-report and performance-based tests of functional status for 18-month outcomes in elderly patients. Aging Clin Exp Res 1996;8:271-276.

Nordin E, Rosendahl E, Lundin-Olsson L. Timed "Up & Go" test: reliability in older people dependent in activities of daily living--focus on cognitive state. Phys Ther 2006;86:646-655.

Nuyens G, de Weerdt W, Ketelaer P, et al. Inter-rater reliability of the Ashworth Scale in multiple sclerosis. Clinical Rehabilitation 1994;8:286-292.

Nys GM, van Zandvoort MJ, de Kort PL, Jansen BP, Kappelle LJ, de Haan EH. Restrictions of the Mini-Mental State Examination in acute stroke. Arch Clin Neuropsychol 2005;20:623-629.

O'Mahoney PG, Rodgers H, Thomson RG, Dobson R, et al. Is the SF-36 suitable for assessing health status of older stroke patients? Age and Aging 1998;27:19-23.

O'Rourke S, MacHale S, Signorini D, Dennis M. Detecting psychiatric morbidity after stroke: comparison of the GHQ and the HAD Scale. Stroke 1998;29:980-985.

Oczkowski W, Barreca S. The Functional Independence Measure: Its use to identify rehabilitation needs in stroke survivors. Arch Phys Med Rehabil 1993;74:1291-1294.

Ormel J, Koeter MW, van den BW, Giel R. Concurrent validity of GHQ-28 and PSE as measures of change. Psychol Med 1989;19:1007-1013.

Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the Functional Independence Measure: A quantitative review. Arch Phys Med Rehabil 1996;77:1226-1232.

Oxford GK, Vogel KA, Le V, Mitchell A, Muniz S, Vollmer MA. Adult norms for a commercially available Nine Hole Peg Test for finger dexterity. Am J Occup Ther 2003;57:570-573.

Ozdemir F, Birtane M, Tabatabaei R, Ekuklu G, Kokino S. Cognitive evaluation and functional outcome after stroke. Americal Journal of Physical Medicine and Rehabilitation 2001;80:410-415.

Pandyan AD, Johnson GR, Price CIM, Curless RH, Barnes MP, Rodgers H. A review of the properties and limitatons of the Ashworth and modified Ashworth scales as measures of spasticity. Clinical Rehabilitation 1999;13:373-383.

Pandyan AD, Price CIM, Rodgers H, Barnes MP, Johnson GR. Biomechanical examination of a commonly used measure of spasticity. Clinical Biomechanics 2001;16:859-865.

Pandyan AD, Price CIM, Barnes MP, Johnson GR. A biomechanical investigation into the validity of the modified Ashworth Scale as a measure of elbow spasticity. Clinical Rehabilitation 2003;17:290-294.

Paolucci S, Antonucci G, Gialloreti LE, et al. Predicting stroke inpatient rehabilitation outcome: the prominent role of neuropsychological disorders. Eur Neurol 1996;36:385-390.

Parker VM, Wade DT, Langton HR. Loss of arm function after stroke: measurement, frequency, and recovery. Int Rehabil Med 1986;8:69-73.

Patrick E, Ada L. The Tardieu Scale differentiates contracture from spasticity whereas the Ashworth Scale is confounded by it. Clin Rehabil 2006;20:173-182.

Pedersen PM, Jorgensen HA, Nakayama H, Raaschou HO, Olsen TS. Comprehensive assessment of activities of daily living in stroke. The Copenhagen Stroke Study. Arch Phys Med Rehabil 1997;78:161-165.

Perenboom RJM, Chorus AMJ. Measuring participation according to the International Classification of Functioning, Disability and Health (ICF). Disability and Rehabilitation 2003;25:577-587.

Petrella RJ, Overend T, Chesworth B. FIM after hip fracture: Is telephone administration valid and sensitive to change? Am J Phys Med Rehabil 2002;81:639-644.

Piercy M, Carter J, Mant J, Wade DT. Inter-rater reliability of the Frenchay Activities Index in patients with stroke and their carers. Clin Rehabil 2000;14:433-440.

Pierre U, Wood-Dauphinee S, Korner-Bitensky N, Gayton D, Hanley J. Proxy use of the Canadian SF-36 in rating health status of the disabled elderly. J Clin Epidemiol 1998;51:983-990.

Platz T, Pinkowski C, van Wijck F, Kim IH, di Bella P, Johnson G. 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 2005;19:404-411.

Podsiadlo D, Richardson S. The timed "Up and Go" test. Arch Phys Med Rehabil 1989;67:387-398.

Poole JL, Whitney SL. Motor Assessment Scale for stroke patients: concurrent validity and interrater reliability. Archives of Physical Medicine and Rehabilitation 1988;69:195-197.

Poole JL, Whitney SL. Assessments of motor function post stroke: A review. Physical and Occupational Therapy in Geriatrics 2001;19:1-22.

Poole JL, Whitney SL. Motor Assessment Scale for stroke patients: concurrent validity and interrater reliability. Archives of Physical Medicine and Rehabilitation 1988;69:195-197.

Post MWM, de Witte LP. Good inter-rater reliability of the Frenchay Activities Index in stroke patients. Clinical Rehabilitation 2003; 17: 548-552.

Post MWM, Gerrisen J, Diederiks JPM, de Witte LP. Measuring health status of people who are wheelchair-dependent: validity of the Sickness Impact Profile 68 and the Nottingham Health Profile. Disability and Rehabilitation 2001; 23(6): 245-253.

Prescott RJ, Garraway WM, Akhtar AJ. Predicting functional outcome following acute stroke using a standard clinical examination. Stroke 1982;13:641-647.

Price CI, Curless RH, Rodgers H. Can stroke patients use visual analogue scales? Stroke 1999;30:1357-1361.

Rabins PV, Brooks BR. Emotional disturbance in multiple sclerosis patients: validity of the General Health Questionnaire (GHQ). Psychol Med 1981;11:425-427.

Rankin J. Cerebral vascular accidents in patients over the age of 60 II: prognosis. Scott Med J 1957;2:200-215.

Richard C, Lussier MT, Gagnon R, Lamarche L. GHQ-28 and cGHQ-28: implications of two scoring methods for the GHQ in a primary care setting. Soc Psychiatry Psychiatr Epidemiol 2004;39:235-243.

Richardson HE, Glass JN. A comparison of scoring protocaols on the clock drawing test in relation to ease of use, diagnostic group and correlations with mini-mental state examination. Journal of the American Geriatrics Society 2002;50:169-173.

Riddle DL, Stratford PW. Interpreting validity indexes for diagnostic tests: an illustration using the Berg Balance Test. Physical Therapy 1999;79:939-948.

Rieck M, Moreland J. The Orpington Prognostic Scale for patients with stroke: reliability and pilot predictive data for discharge destination and therapeutic services. Disabil Rehabil 2005;27:1425-1433.

Rinaldi P, Mecocci P, Benedetti C, et al. Validation of the five-item geriatric depression scale in elderly subject in three different settings. Journal of the American Geriatrics Society 2003;51:694-698.

Ring H, Feder M, Schwartz J, Samuels G. Functional measures of first-stroke rehabilitation inpatients: Usefulness of the Functional Independence Measure total score with a clinical rationale. Arch Phys Med Rehabil 1997;78:630-635.

Roberts L, Counsell R. Assessment of clinical outcomes in acute stroke trials. Stroke 1998;28:986-991.

Robinson RG. Poststroke depression: prevalence, diagnosis, treatment and disease progression. Biol Psychiatry 2003;54:376-387.

Roccaforte WH, Burke WJ, Bayer BL, Wengel SP. Validation of a telephone version of the mini-mental state examination. Journal of the American Geriatrics Society 1992;40:697-702.

Rockwood K, Awalt E, Carver D, MacKnight C. Feasibility and measurement properties of the Functional Reach and the timed Up and Go tests in the Canadian study of health and aging. Journal of Gerontology 2000;55A:M70-M73.

Rossier P, Wade DT. Validity and reliability comparison of 4 mobility measures in patients presenting with neurologic impairment. Arch Phys Med Rehabil 2001;82:9-13.

Ruchinskas RA, Curyto KJ. Cognitive screening in geriatric rehabilitation. Rehabilitation Psychology 2003;48:14-22.

Sabari JS, Lim AL, Velozo CA, Lehman L, Kieran O, Lai JS. 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 2005;86:1609-1615.

Salbach N, Mayo N, Higgins J, Ahmed S, Finch LE, Richards CL. Responsiveness and predictability of gait speed and other disability measures in acute stroke. Arch Phys Med Rehabil 2001;82:1204-1212.

Sanford J, Moreland J, Swanson LR, Stratford PW, Gowland C. Reliability of the Fugl-Meyer Assessment for testing motor performance in patients following stroke. Physical Therapy 1993;73:447-454.

Scanlan JM, Brush M, Quijano C, Borson S. Comparing clock tests for dementia screening: naive judgments vs. formal systems -- what is optimal? Int J Geriatric Psychiatry 2002;17:14-21.

Schindl MR, Forstner C, Kern H, Hesse S. Treadmill training with partial body weight support in nonambulatory patients with cerebral palsy. Arch Phys Med Rehabil 2000;81:301-306.

Schindl MR, Forstner C, Kern H, Hesse S. Treadmill training with partial body weight support in nonambulatory patients with cerebral palsy. Arch Phys Med Rehabil 2000;81:301-306.

Schlegel DJ, Tanne D, Demchuk AM, Levine SR, Kasner SE. Prediction of hospital disposition after thrombolysis for acute ischemic stroke using the National Institutes of Health Stroke Scale. Arch Neurol 2004;61:1061-1064.

Schmulling S, Grond M, Rudolf J, Kiencke P. Training as a prerequisite for reliable use of NIH Stroke Scale. Stroke 1998;29:1258-1259.

Schuling J, de Haan RJ, Limburg M, Groenier KH. The Frenchay Activities Index; assessment of functional status in stroke patients. Stroke 1993;24:1173-1177.

Seaby L, Torrance G. Reliability of a physiotherapy functional assessment used in a rehabilitation setting. Physiotherapy Canada 1989;41:264-271.

Seaby L and Torrance G. Measurement properties of the physiotherapy clinical outcomes variables scale (COVS): a mobility funcitonal assessment. Physiotherapy Canada 1994;46:76.

Segal ME, Schall RR. Determining funcitonal/health status and its relation to disability in stroke survivors. Stroke 1994;25:2391-2397.

Segal ME, Gillard M, Schall RR. Telephone and in-person proxy agreement between stroke patients and caregivers for the Functional Independence Measure. Americal Journal of Physical Medicine and Rehabilitation 1996;75:208-212.

Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. J Clin Epidemiol 1989;42:703-709.

Shah S, Cooper B, Maas F. The Barthel Index and ADL evaluation in stroke rehabilitation in Australia, Japan, the US and the USA. The Australian Occupational Therapy Journal 1992;39:5-13.

Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clin Gerontol 1986;5:165-172.

Shelton F, Volpe BT, Reding M. Motor impairment as a predictor of functional recovery and guide to rehabilitation treatment after stroke. Neurorehabilitation and Neural Repair 2001;15:237.

Shinar D, Gross CR, Bronstein KS, et al. Reliability of the activities of daily living scale and its use in telephone interview. Arch Phys Med Rehabil 1987;68:723-728.

Shinohara Y, Minematsu K, Amano T, Ohashi Y. Modified Rankin scale with expanded guidance scheme and interview questionnaire: interrater agreement and reproducibility of assessment. Cerebrovasc Dis 2006;21:271-278.

Shumway-Cook A, Baldwin M, Polissar NL, Gruber W. Predicting the probability for falls in community-dwelling older adults. Phys Ther 1997;79:819.

Siggeirsdottir K, Jonsson BY, Jonsson H Jr, Iwarsson S. The Timed "Up and Go" is dependent on chair type. Clin Rehabil 2002;16:609-616.

Simondson JA, Goldie P, Greenwood KM. The Mobility Scale for Acute Stroke Patients: concurrent validity. Clin Rehabil 2003;17:558-564.

Singh A, Black SE, Herrmann N, et al. Functional and neuroanatomic correlations in poststroke depression. The Sunnybrook Stroke Study. Stroke 2000;31:637-644.

Sloan RL, Sinclair E, Thompson J, Taylor S, Pentland B. Inter-rater reliability of the modified Ashworth Scale for spasticity in hemiplegic patients. International Journal of Rehabilitation Research 1992;15:161.

Smith PM, Bennett Illig S, Fielder RC, Hamilton BB, Ottenbacher KJ. Intermodal agreement of follow-up telephone funcitonal assessment using the functional independence measure in patients with stroke. Arch Phys Med Rehabil 1996;77:431-435.

Smith PS, Hembree JA, Thompson ME. Berg Balance Scale and Functional Reach: determining the best clinical tool for individuals post acute stroke. Clin Rehabil 2004;18:811-818.

Sneeuw KC, Aaronson NK, de Haan RJ, Limburg M. Assessing quality of life after stroke. The value and limitations of proxy ratings. Stroke 1997;28:1541-1549.

Snyder AG, Stanley MA, Novy DM, Averill PM, Beck JG. Measures of depression in older adults with generalized anxiety disorder: a psychometric evaluation. Depression and Anxiety 2000;11:120.

Sommerfeld D, von Arbin MH. Disability test 10 days after acute stroke to predict early discharge home in patients 65 years and older. Clin Rehabil 2001;15:528-534.

South MB, Greve KW, Bianchini KJ, Adams D. Interrater reliability of three clock drawing test scoring systems. Applied Neuropsychology 2001;8:174-179.

Stansfeld SA, Roberts R, Foot SP. Assessing the validity of the SF-36 general health survey. Qual Life Res 1997;6:217-244.

Stark SL, Edwards DF, Hollingsworth H, Gray DB. Validation of the Reintegration to Normal Living Index in a population of community-dwelling people with mobility limitations. Arch Phys Med Rehabil 2005;86:344-345.

Stavem K, Lossius M, Ronning OM. Reliability and validity of the Canadian Neurological Scale in retrospective assessment of initial stroke severity. Cerebrovasc Dis 2003;16:286-291.

Steffen TM, Hacker TA, Mollinger L. 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 2002;82:128-137.

Steiner A, Raube K, Stuck AE, et al. Measuring psychosocial aspects of well-being in older community residents: performance of four short scales. Gerontologist 1996;36:54-62.

Stevenson TJ. Using impairment inventory scores to determine ambulation status in individuals with stroke. Physiotherapy Canada 1999;51:168-174.

Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. Australian Journal of Physiotherapy 2001;47:29-38.

Stiles PG, McGarrahan JE. The Geriatric Depression Scale: A comprehensive review. Journal of Clinical Geropsychology 1998;4:89-109.

Stone SP, Ali B, Auberleek I, Thompsell A, Young A. The Barthel Index in clinical practice: use on a rehabilitation ward for elderly people. Journal of the Royal College of Physicians of London 1994;28:419-423.

Studenski S, Wallace D, Duncan PW, Rymer M, Lai SM. Predicting stroke recovery: Three and six-month rates of patient-centred functional outcomes based on the Orpington Prognostic Scale. Journal of the American Geriatrics Society 2001;49:308-312.

Sturm JW, Osborne RH, Dewey HM, Donnan GA, Macdonell RA, Thrift AG. Brief comprehensive quality of life assessment after stroke: the assessment of quality of life instrument in the north East melbourne stroke incidence study (NEMESIS). Stroke 2002;33:2888-2894.

Su CY, Chang JJ, Chen HM, Su CJ, Chien TH, Huang MH. Perceptual differences between stroke patients with cerebral infarction and intracerebral hemorrhage. Archives of Physical Medicine and Rehabilitation 2000;81:714.

Suhr JA, Grace J, Allen J, Nadler JD, McKenna M. Quantitatie and qualitative performance of stroke versus normal elderly on six clock drawing systems. Archives of Clinical Neuropsychology 1998;13:495-502.

Suhr JA, Grace J. Brief cognitive screening of right hemisphere stroke: relation to functional outcome. Archives of Physical Medicine and Rehabilitation 1999;80:773-776.

Sulter G, Steen C, deKeyser J. Use of the Barthel Index and Modified Rankin Scale in acute stroke trials. Stroke 1999;30:1538-1541.

Sumer MM, Ozdemir I, Tascilar N. Predictors of outcome after acute ischemic stroke. Acta Neurol Scand 2003;107:276-280.

Sunderland A, Tinson D, Bradley L, Hewer RL. Arm function after stroke. An evaluation of grip strength as a measure of recovery and a prognostic indicator. J Neurol Neurosurg Psychiatry 1989;52:1267-1272.

Sveen U, Bautz-Holter E, Sodring KM, Wyller TB, Laake K. Association between impairments, self-care abiilty and social activities 1 year after stroke. Disabil Rehabil 1999;21:372-377.

Tang WK, Mok V, Chan SS, et al. 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 2005;18:3-7.

Teasell R ie. Stroke Rehabilitation Evidence-Based Review: Part I. Topics in Stroke Rehabilitation 10[1], 1-78.

Teasell R ie. Stroke Rehabilitation Evidence-Based Review: Part II. Topics in Stroke Rehabilitation 2003;10:1-129.

Teng EL, Chui HC. The modified Mini-Mental State (3MS) Examination. Journal of Clinical Psychiatry 1987;48:314-318.

Tepper S, Beatty P, DeJong G. Outcomes in traumatic brain injury: self-report versus report of significant others. Brain Injury 1996;10:575-581.

Thompson M, Medley A. Performance of community dwelling elderly on the Timed Up and Go Test. Physical and Occupational Therapy in Geriatrics 1995;13:17-30.

Tombaugh TN, McIntyre NJ. The Mini-Mental State Examination: A comprehensive review. Journal of the Americal Geriatric Society 1992;40:922-935.

Tooth LR, McKenna KT, Smith M, O'Rourke P. Further evidence for the agreement between patients with stroke and their proxies on the Frenchay Activities Index. Clinical Rehabilitation 2003;17:656-665.

Tooth LR, McKenna KT, Smith M, O'Rourke PK. Reliability of scores between stroke patients and significant others on the Reintegration to Normal Living (RNL) Index. Disabil Rehabil 2003;25:433-440.

Torrance GM and Seaby LS. Rasch analysis of the physiotherapy clinical outcome variables (COVS). Physiotherapy Canada 1994;46:4.

Trombly CA, Vining-Radomski M, Schold-Davis E. Achievement of self-identified goals by adults with traumatic brain injury: Phase I. American Journal of Occupational Therapy 1998;52:810-818.

Tyson SF, DeSouza LH. Reliability and validity of functional balance tests post stroke. Clin Rehabil 2004;18:916-923.

Tyson S, DeSouza L. A systematic review of methods to measure balance and walking post-stroke. Part 1: ordinal scales. Physical Therapy Reviews2002 Sep; 7:173-186.

Uyttenboogaart M, Stewart RE, Vroomen PC, De Keyser J, Luijckx GJ. Optimizing cutoff scores for the Barthel index and the modified Rankin scale for defining outcome in acute stroke trials. Stroke 2005;36:1984-1987.

Valach L, Signer S, Hartneier A, Hofer K, Steck GC. Chedoke-McMaster stroke assessment and modified Barthel Index self-assessment in patients with vascular brain damage. Int J Rehabil Res 2003;26:93-99.

Van-Gorp W, Marcotte T, Sultzer D, Hinkin C, Mahler M, Cummings J. Screening for dementia: comparison of three commonly used instruments. Journal of Clinical and Experimental Neuropsychology 1999;21:29-38.

Van de Port IGL, Kelelaar M, Schepers VPM, van den Bos GAM, Lindeman E. Monitoring the functional health status of stroke patients: the value of the Stroke-Adpated Sickness Impact Profile-30. Disabil Rehabil 2004;26:635-640.

Van der Lee JH, De G, V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001;82:14-19.

Van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. J Rehabil Med 2001;33:110-113.

Van der Lee JH, Roorda LD, Beckerman H, Lankhorst GJ, Bouter LM. Improving the Action Research Arm test: a unidimensional hierarchical scale. Clin Rehabil 2002;16:646-653.

van der Putten JMF, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry 1999;66:480-484.

van Marwijk H, Arnold I, Bonnema J, Kaptein A. Self-report depression scales for elderly patients in primary care: A preliminary study. Family Practice 1993;10:63-65.

van Marwijk H, Wallace P, De Bock GH, Hermans J, Kaptein A, Mulder JD. Evaluation of the feasibility, reliability & diagnostic value of shortened versions of the geriatric depression scale. British Journal of General Practice 1995;45:195-199.

van Straten A, de Haan RJ, Limburg M, Schuling J, Bossuyt PM, van den Bos GAM. A stroke-adapted 30-item version of the Sickness Impact Profile to assess quality of life (SA-SIP-30). Stroke 1997;28:2155-2161.

van Straten A, de Haan RJ, Limburg M, van den Bos GAM. Clinical meaning of the Stroke-Adapted Sickness Impact Profile-30 and the Sickness Impact Profile-136. Stroke 2000;31:2615.

Van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJA, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1988;19:604-607.

van Tuijl JH, Janssen-Potten YJ, Seele HA. Evaluation of upper extremity motor function tests in tetraplegics. Spinal Cord 2002;40:51-64.

van Wijck FMJ, Pandyan AD, Johnson GR, Barnes MP. Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. Neurorehabilitation and Neural Repair 2001;15:23-30.

Visser MC, Koudstaal PJ, Erdman RAM, Deckers JW, et al. Measuring quality of life in patients with myocardial infarction or stroke: a feasibility study of four questionnaires in The Netherlands. J Epidemiol Community Health 1995; 49: 513-517.

Wade DT, Leigh-Smith J, Langton-Hewer R. Social activities after stroke: measurement and natual history using the Frenchay Activities Index. Int Rehabil Med 1985;1:176-181.

Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability? Int Disabil Stud 1988;10:64-67.

Wade DT. Measurement in neurological rehabilitation. New York: Oxford University Press, 1992.

Wade DT. Measuring arm impairment and disability after stroke. Int Disabil Stud 1989;11:89-92.

Wagenaar RC, Meijer OG, van Wieringen PC, et al. The functional recovery of stroke: a comparison between neuro-developmental treatment and the Brunnstrom method. Scand J Rehabil Med 1990;22:1-8.

Wallace D, Duncan PW, Lai SM. 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 reponsiveness. J Clin Epidemiol 2002;55:922-928.

Walters SJ, Munro JF, Brazier JE. Using the SF-36 with older adults: a cross-sectional community-based survey. Age and Aging 2001;30:337-343.

Wang CH, Hsueh IP, Sheu CF, Yao G, Hsieh CL. Psychometric properties of 2 simplified 3-level balance scales used for patients with stroke. Phys Ther 2004;84:430-438.

Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF36) I: Conceptual framework and item selection. Medical Care 1992;30:473-483.

Watson YI, Arfken CL, Birge SJ. Clock completion: An objective screening test for dementia. Journal of the American Geriatrics Society 1993;41:1235-1240.

Wee JYM, Bagg SD, Palepu A. The Berg Balance Scale as a predictor of length of stay and discharge in an acute stroke rehabilitation setting. Arch Phys Med Rehabil 1999;80:448-452.

Wee JYM, Wong H, Palepu A. Validation of the Berg Balance Scale as a predictor of length of stay and discharge destination in stroke rehabilitation. Arch Phys Med Rehabil 2003;84:731-735.

Weimar C, Kurth T, Kraywinkel K, et al. Assessment of Functioning and Disability After Ischemic Stroke. Stroke 2002;33:2053-2059.

Weinberger M, Nagle B, Hanlon JT, et al. Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. Journal of the American Geriatric Society 1994;42:1295-1299.

Weinberger M, Oddone EZ, Samsa G, Landsman PB. Are health-related quality of life measures affected by the mode of administration? J Clin Epidemiol 1996;49:135-140.

Werneke U, Goldberg DP, Yalcin I, Ustun BT. The stability of the factor structure of the General Health Questionnaire. Psychol Med 2000;30:823-829.

Whaley CJ, Morrison DL, Payne RL, Fritschi L, Wall TD. Chronicity of psychological strain in occupational settings and the accuracy of the General Health Questionnaire. J Occup Health Psychol 2005;10:310-319.

Whitney SL, Poole JL, Cass SP. A review of balance instruments for older adults. Am J Occup Ther 1998;52:666-671.

Williams CL, Rittman MR, Boylstein C, Faircloth C, Haijing Q. Qualitative and quantitative measurement of depression in veterans recovering from stroke. J Rehabil Res Dev 2005;42:277-290.

Williams LS, Weinberger M, Harris LE, Biller J. Measuring quality of life in a way that is meaningful to stroke patients. Neurology 1999;53:1839-1843.

Williams LS, Weinberger M, Harris MD, Clark DO, Biller J. Development of a stroke-specific quality of life scale. Stroke 1999;30:1362-1369.

Williams LS, Redmon G, Martinez B, and Weinberger M. Proxy ratings of stroke specific quality of life (SSQOL) scores. Stroke 2000;31:301.

Williams LS, Yilmaz EY, Lopez-Yunez AM. Retrospective assessment of initial stroke severity with the NIH Stroke Scale. Stroke 2000;31:858-862.

Willmott SA, Boardman JA, Henshaw CA, Jones PW. Understanding General Health Questionnaire (GHQ-28) score and its threshold. Soc Psychiatry Psychiatr Epidemiol 2004;39:613-617.

Wilson JTL, Hareendran A, Grant M, et al. Improving the assessment of outcomes in stroke: Use of structured interview to assign grades on the Modified Rankin Scale. Stroke 2002;33:2243-2246.

Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. Stroke 2005;36:777-781.

Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ. Assessment of scales of disability and handicap for stroke patients. Stroke 1991;22:1242-1244.

Woo D, Broderick JP, Kothari RU, et al. Does the National Institutes of Health Stroke Scale favor left hemisphere strokes? NINDS t-PA Stroke Study Group. Stroke 1999;30:2355-2359.

Wood-Dauphinee S, Williams JI. Reintegration to Normal Living as a proxy to quality of life. J Chronic Dis 1987;40:491-502.

Wood-Dauphinee SL, Opzoomer MA, Williams JI, Marchand B, Spitzer WO. Assessment of global function: The Reintegration to Normal Living Index. Arch Phys Med Rehabil 1988;69:583-590.

Wood-Dauphinee S, Williams JI, Shapiro SH. Examining outcome measures in a clinical study of stroke. Stroke 1990;21:731-739.

Wood-Dauphinee S, Berg KO, Bravo G, Williams JL. The Balance Scale: responsiveness to clinically meaningful changes. Can J Rehab 1997;10:35-50.

World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva: World Health Organization, 2001.

World Health Organization. Towards a Common Language for Functioning, Disability and Health: ICF.

Wright J, Cross J, Lamb S. Physiotherapy outcome measure for rehabilitation of elderly people: responsiveness to change of the Rivermead Mobility Index and Barthel Index. Physiotherapy 1998;84:216-221.

Wright CJ, Swinton LC, Green TL, Hill MD. Predicting final disposition after stroke using the Orpington Prognostic Score. Can J Neurol Sci 2004;31:494-498.

Wyller TB, Sveen U, Bautz-Holter E. The Frenchay Activities Index in stroke patients: agreement between scores by patients and by relatives. Disabil Rehabil 1996;18:454-459.

Yeo D, Faleiro R, Lincoln NB. Barthel ADL Index: a comparison of administration methods. Clin Rehabil 1995;9:34-39.

Yesavage JA, Brink TL. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 1983;17:37-49.

York CD, Cermak SA. Visual perception and praxis in adults after stroke. American Journal of Occupational Therapy 1995;49:550.

Zung WWK. A self-rating depression scale. Archives of General Psychiatry 1965;12:63-70.

Zwick D, Rochelle A, Choksi A, Domowicz J. Evaluation and treatment of balance in the elderly: A review of the efficacy of the Berg Balance Test and Tai Chi Quan. Neurorehabilitation 2000;15:49-56.