Journal of
NEUROLOGY
NEUROSURGERY
& PSYCHIATRY

Editorial

Evaluating neurological outcome measures: the bare essentials

Chronic impairment is a common consequence of neurological disease. Many of these disorders affect the younger age group and are often progressive over many decades. Measurement of health related outcomes such as disability, handicap, and quality of life are therefore important in the evaluation of therapeutic efficacy. To ensure sound measurement of these outcomes, it is essential that the instruments used have been comprehensively evaluated not only in terms of clinical appropriateness but also, and perhaps more importantly, with respect to their scientific properties. Clinicians are often unfamiliar with the rigorous scientific techniques required to design and evaluate health measurement tools largely because the theoretical foundations and methodological concepts, which originated in the social sciences, have been slow to transfer to medicine. This editorial introduces these concepts and provides a basis of knowledge for informed decision making in the evaluation of outcome measures.

Instruments for measurement of outcome must be evaluated in terms of both clinical usefulness and scientific soundness. If an instrument is to be clinically useful and acceptable such that it can be incorporated into daily practice, it must be appropriate to the patient group being studied, brief, user friendly, practical to administer, and cost effective. Unwieldy, time consuming, and resource consuming instruments have limited use in clinical practice. Clinical utility, however, does not guarantee scientific soundness in terms of rigorous measurement.

The second and perhaps more important step in instrument evaluation is the assessment of three scientific properties which ensure reliable and valid measurement of the health outcome of interest:

(1) *Reliability* considers the question of whether an instrument measures outcome in a way that is accurate, consistent, stable over time, and reproducible.

(2) *Validity* considers whether an instrument measures what it purports or is intended to measure.

(3) *Responsiveness* determines whether an instrument is sensitive to and can detect clinically important change.

The basic principles and methods of these three scientific concepts were developed in the social sciences, particularly psychology, where the need for rigorous measurement of abstract entities such as intelligence and personality stimulated conceptual and methodological advances that led to the establishment of psychometrics—the science of measurement. The foundations of this science were laid in the mid-1800s and were followed by extensive developments in the 1930s to '50s.

Clinical medicine has traditionally been concerned with simple and easy to measure outcomes such as mortality, presence or absence of disease, duration of survival, and duration of disease free interval. Developments in health care and changing social conditions have resulted in an increasing prevalence of chronic illnesses and led to a broader World Health Organisation definition of health as "a complete state of physical, mental, and social wellbeing and not merely the absence of disease or infirmity". These changes, coupled with recent developments including diagnostic advances,[1] emergence of new treatments,[2] and the importance of incorporating the patient perspective,[3–5] have highlighted the inadequacy of traditional outcomes and have pointed to the need for the assessment of more pertinent but abstract concepts such as disability, handicap, and quality of life. The knowledge required to ensure that these complex entities are being measured with the necessary scientific rigour has yet to transfer from the social sciences and is generally unavailable to most clinicians. This may explain why, even though many instruments exist,[6–8] measurement of disability and handicap is viewed as being in its infancy from a scientific point of view.[7 9 10]

Scientific evaluation of outcome measurement instruments

Scientific evaluation of outcome measurement instruments involves assessment of reliability, validity, and responsiveness.[11–20] Standard techniques, described below, are available to determine these properties. As these are not all or nothing concepts, the aim of assessment is to evaluate the strength of evidence for each property.

RELIABILITY

A reliable measure is one which produces results that are accurate, consistent, stable over time, and reproducible. There are four different types of reliability:

Internal consistency (interitem consistency)

Internal consistency is the extent to which items comprising a scale measure the same concept—that is, a measure of the homogeneity of the scale. This is assessed by a statistical technique, Cronbach's alpha in the case of continuous measurement scales,[21] or, in the case of dichotomous measurement scales, by the Kuder-Richardson 20 statistic.[22]

Test-retest reliability
Test-retest reliability is the stability of a measuring instrument over time. It is assessed by giving the measure to the same subjects on two different occasions and examining the correlation between the two scores. It is pertinent for self report instruments.

Rater reliability
Rater reliability is the agreement between raters or within an individual rater. It is therefore pertinent when the measurement process involves a rater, but is not relevant for self report instruments. There are two types: *Interrater reliability* is the agreement between two or more raters. This is assessed by examining the correlation between ratings obtained from independent observers using a correlation coefficient or, alternatively, the intraclass correlation coefficient in the case of continuous scales,[23] or Cohen's kappa for dichotomous scales.[24-26] *Intrarater reliability* is the agreement between two ratings made by a single observer on the same patient. It is assessed by examining the correlation between these ratings using the same statistical methods as above.

Parallel forms reliability (alternate forms reliability)
Parallel forms reliability is the degree of agreement between two identically constructed (parallel) forms of the same measure. In some cases, there is a need for two similar versions of a measure. This would be true when measuring memory—for example, when a learning effect from the initial administration of a memory test might bias the measurement of memory on the second occasion if the same measure were used. In such circumstances, two memory measures, which are identical in terms of level of difficulty etc but different in content, would be needed to reliably measure memory. These two versions of a measure are called parallel (or alternate) forms.

Reliability is therefore a generic term; each of the four types contributes to the overall evaluation of the reliability of a measure. As test-retest, interrater, and intrarater reliability are concerned with the degree of consistency or agreement between independently derived sets of scores; they can all be expressed in terms of a correlation coefficient. Bland and Altman argue that correlation coefficients are misleading for this purpose because they measure the strength of the relation between two variables and not the agreement between them; they are affected by the range of the true quantity in the sample such that the wider the range the higher the correlation even if the agreement remains the same; and considerable lack of agreement may be associated with high correlation.[27 28] In view of these theoretical arguments, Bland and Altman propose a graphical representation to indicate the degree of agreement between two ratings. Needless to say their method has its critics.[20] The use of both techniques is therefore advisable.

At a conceptual level, reliability assessment also reflects the degree of random error associated with a measurement instrument (high reliability = low error). Random error refers to all chance factors that produce variations on repeated measurement. Reliability assessment, therefore, which evaluates the consistency of repeated measurements, determines the extent of random error associated with the measurement process. The potential sources of random error are protean, but can be attributed to errors in the measure itself, the person doing the measuring, or the person being measured.[18] As the four different types of reliability consider different sources of random error, it becomes apparent why it is necessary to examine each type of reliability for a comprehensive evaluation. It is a common misconception that providing evidence for one type of reliability is sufficient.

By way of an example, if a new disability scale was developed for stroke patients requiring rating by a neurologist an appropriate reliability study would consist of the following: (a) assessment of internal consistency by statistical analysis of the ratings produced on a predetermined number of patients; (b) assessment of interrater reliability by comparing agreement between ratings obtained from independent neurologists on the same stroke patients. Many studies of interrater reliability are undertaken using two raters. Whereas this is acceptable more generalisable results are obtained when a greater number of raters are involved; (c) assessment of intrarater reliability by comparing agreement between ratings obtained from the same neurologist on the same stroke patients. Results from multiple rather than single neurologists will give more generalisable results. The time between the two ratings should be short enough to exclude the possibility of change occurring in the patient's clinical state but long enough to prevent the neurologist remembering the initial ratings. In practice this interval will be variable; (d) test-retest reliability would only be appropriate if the instrument involved self reported disability rather than rater observed disability. For example, the recently published postal version of the Barthel ADL Index.[29] Test-retest reliability would be determined by giving the measure on two occasions to the same patients. Again, the interval between the two ratings (test-retest interval) should be short enough to exclude the possibility of change occurring in the patient's clinical state but long enough to prevent the patient remembering the initial ratings; (e) Parallel forms reliability could only be undertaken if two identical forms of the measure exist. Such parallel forms seldom exist for the measurement of disability; however, if they did the two measures would be given to the same patient at the same time and the agreement between ratings compared.

VALIDITY

Measuring instruments that have been shown to be reliable have satisfied only one of the criteria for achieving scientific acceptance, they must also be shown to be valid. Whereas an instrument must be reliable to be shown to be valid, evidence for reliability does not guarantee evidence for validity.

Validity concerns the relation between the concept being measured and the instrument used to assess that concept. It can be broadly defined as the extent to which the instrument measures the concept it purports or is intended to measure.[17 30-32] There are three types of validity:

Content validity
Content validity is the extent to which a measure is representative of the conceptual domain it is intended to cover;

Criterion related validity
Criterion related validity is the degree to which a measure correlates with a gold standard (the criterion);

Construct validity
Construct validity is a process used to establish the validity of a measurement instrument through a series of studies examining the relation between the measure and other measures or behaviours.

Whereas there is some conceptual overlap between the three types of validity, each takes a somewhat different approach to assessing the extent to which an instrument

measures what it purports. Therefore, all three types of validity must be considered for an overall evaluative judgement of the adequacy of inferences drawn from a test.[33] By way of an analogy, the validation process can be likened to a court trial.[19] Firstly, validity must be formally proved by persuasive evidence. Secondly, evidence for validity must be gathered using specific methods from multiple sources. Thirdly, validity in one context or setting does not suppose validity in a different situation.

Evidence for content validity is commonly obtained through comprehensive sampling of the domain of interest using various sources including comprehensive reviews of the medical literature, consensus expert opinion, qualitative patient interviews, and examination of existing measures of the same or a similar concept. Unlike other types of validity, evidence for content validity is logical rather than statistical.

Evidence for criterion related validity is provided by examining the correlation between the measure and a gold standard (the criterion). There are two types of criterion related validity: concurrent and predictive. The distinction between the two refers to whether the measure is compared with a gold standard measured at the same time (concurrent) or in the future (predictive).

When no gold standard exists with which to compare a measure, a different type of evidence is required to support the supposition that the measure is valid for a particular purpose. This is often the case with health outcomes such as disability, handicap, and quality of life. Under these circumstances, construct rather than criterion related validity is evaluated. Construct validation considers the questions of whether an instrument: (a) measures what it is supposed to measure; (b) does not measure what it does not intend to measure; (c) distinguishes between groups in predictable ways; and (d) produces results consistent with theoretical expectation.[30 31] In practice, evidence for construct validity is gathered by undertaking a series of studies to determine:

(1) convergent validity—the extent to which the measure correlates with other measures of related entities.

(2) Discriminant validity—the extent to which the measure does not correlate with measures of different entities.

(3) Group differences—the extent to which the measure is able to detect differences in groups known to differ in the concept being measured;

(4) Hypothesis testing—the extent to which hypotheses generated on the basis of some theoretical notion about the construct are supported by results obtained from the measure.

Construct validity is therefore achieved by the accumulation of all of these types of evidence. It is an ongoing process—that is, no single piece of evidence is sufficient to show validity. Rather, a series of findings from several studies lends weight to the body of evidence for the validity of a measure. In view of the lack of gold standards in areas such as disability and handicap assessment, construct validity is the cornerstone of the evaluation of the validity of many health measurement instruments.

To illustrate validity assessment consider again our new disability measure for stroke mentioned above. Let us assume that satisfactory evidence for all types of reliability has been determined. We now need to provide evidence that the instrument measures disability. An appropriate validity study might include the following:

Content validity

Ideally this would have been determined during scale construction, but assuming it has not been it must be demonstrated. Copies of the scale would be sent to a cross section of experts within the field of disability measurement including neurologists, rehabilitation consultants, physiotherapists, occupational therapists, speech therapists, neuropsychologists, nurses, and appropriate others to obtain their opinions as to the suitability as a measure of disability. These professionals could be asked whether, in their opinion, they think that any items should be added to, or removed from, the scale. In addition a literature review and comparison with existing measures would provide wider evidence of the domain of disability.

Criterion related validity

Concurrent validity—As no accepted gold standard measure of disability exists this is considered under construct validity.

Predictive validity—If we thought that the scale results at seven days after stroke could predict degree of disability at one year after stroke, data collected at these two instances in time must be correlated.

Construct validity

Convergent construct validity—This is determined on the basis of a strong correlation between the results obtained with the new scale and other measures of disability—for example the Barthel activities of daily living (ADL) index;

Discriminant construct validity—This is determined on the basis of the new scale correlating less well with measures of impairment, handicap, and quality of life than those measuring disability.

Group differences construct validity—This is determined on the basis of the degree of disability in tetraplegic patients being shown to be greater than that for paraplegic patients.

Hypothesis testing—Here we would set out to confirm a series of hypotheses. Our hypotheses could include: stroke patients with dominant hemisphere lesions should have more communication difficulties than patients with non-dominant hemisphere lesions; patients with head injury should have lower cognitive scores than patients with spinal cord trauma; disability scale scores should vary in accordance with the type of discharge destination and care needs.

RESPONSIVENESS

Whereas reliability and validity are the major determinants of the scientific robustness of a measure, the ability of an instrument to detect clinically significant change is also essential when evaluating the relative benefits of different interventions. This is particularly important when treatments are associated with small but significant differences (a feature of current day interventions) which may be undetected by measures that are unresponsive. In such cases a clinically appropriate, reliable, and valid but unresponsive instrument is of limited value.

The term "sensitivity to change" has been used as an alternative to responsiveness for describing the measurement of clinically significant change. Whereas both terms are equally valid we would encourage the use of responsiveness for two reasons. Firstly, the specialist literature on measurement predominantly uses the term responsiveness[34–38]; and secondly, using the term sensitivity can result in confusion with its use in the epidemiological sense to mean the ability to detect a high proportion of true cases.[39]

Historically responsiveness has been considered as an aspect of validity. Many of the traditional outcome measures developed in the social sciences and borrowed for use in medical settings, however, although reliable and

valid, have failed to detect subtle but significant clinical change. The ability of an instrument to measure clinically significant change is so salient to clinical practice and the evaluation of medical interventions, that responsiveness has emerged as a separate concept distinct from validity.

The responsiveness of an instrument can be determined in several ways, including: serial administration of the measure at different points over time when clinical changes are expected to occur (for example, before and after treatment of known efficacy); comparison against other criteria of change (for example, staff and patient perceptions of change); and comparison with other measures of the same concept to assess relative responsiveness.[34–37 40]

Conclusions

Very few of the many instruments available for measuring neurological outcomes have undergone comprehensive evaluation. One notable exception is the medical outcomes study short form health survey (SF-36), which measures health related quality of life.[41] Many of the commonly used instruments—for example the Kurtzke expanded disability status scale (EDSS)—provide limited data about their scientific properties. Even the Barthel ADL index, considered by many to be a benchmark for measurement of disability, has incomplete psychometric data. Of more concern there remain many without any, or with incorrect evaluation of their scientific properties. Any instrument to be used in health care evaluation and technology assessment must be formally and comprehensively evaluated according to scientific criteria.

New and better instruments can only be developed from the proved shortcomings of existing measures. These must be designed and comprehensively evaluated using the standard methods derived from measurement theory described in this article before they can be considered for widespread use. Evidence about the reliability and validity of new and existing measures should be peer reviewed and easily available, along with guidelines for the appropriate use of the instrument. Such information is generally provided in a user manual or technical report which accompanies the publication of a new measure.

The quality of outcome data is determined by the quality of the measurement instruments used to produce it. Using poorly evaluated instruments may lead to misleading results, and thereby affect important clinical decisions both at patient and population level. It is essential, therefore that clinicians be aware of the need for comprehensive scientific evaluation of the instruments used to measure health outcomes. This implies a commitment to using instruments that have been fully evaluated, as well as knowledge about standard techniques for developing and validating new instruments. Significant methodological expertise is required to design and comprehensively evaluate health measurement instruments, as well as to analyse and interpret data. Clinicians may benefit from collaboration with social science measurement experts when working in this area.

JCH is funded by the Central Audit Unit of the NHS National Executive. The Neuro-Rehabilitation Section of the Institute of Neurology is generously supported by the Brain Research Trust.

JEREMY C HOBART
Department of Clinical Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, and Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

DONNA L LAMPING
Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

ALAN J THOMPSON
University Department of Clinical Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, UK

Correspondence to: Dr A J Thompson.

1 Thompson AJ, Miller DH. Magnetic resonance imaging in clinical practice. In: Kennard C, ed. *Recent advances in clinical neurology*. Vol 7. London: Churchill Livingstone, 1992:199–219.
2 McDonald WI. New treatments for multiple sclerosis. *BMJ* 1995;**310**: 345–6.
3 Reiser SJ. The Era of the patient. *JAMA* 1993;**269**:1012–7.
4 Hopkins A. Economic change and health service reform: likely impact on teaching, practice, and research in neurology. *J Neurol Neurosurg Psychiatry* 1994;**57**:667–71.
5 Devinsky O. Outcomes research in neurology: incorporating health-related quality of life. *Ann Neurol* 1995;**37**:141–2.
6 Bowling A. *Measuring disease: a review of disease specific quality of life measurement scales*. Buckingham: Open University Press, 1995.
7 McDowell I, Newell C. *Measuring Health: a guide to rating scales and questionnaires*. Oxford: Oxford University Press, 1987.
8 Wade DT. *Measurement in neurological rehabilitation*. Oxford: Oxford University Press, 1992.
9 Frey WD. Functional assessment in the '80's: a conceptual enigma, a technical challenge. In: Halpern AS, Fuhrer MJ, eds. *Functional assessment in rehabilitation*. Baltimore: Paul H. Brookes, 1984:
10 Keith RA. Functional assessment measures in medical rehabilitation: current status. *Arch Phys Med Rehabil* 1984;**65**:74–8.
11 Lord FM, Novick MR. *Statistical theories of mental test scores*. Reading, MA: Addison Wesley, 1968.
12 Brown FG. *Principles of educational and psychological testing*. Hinsdale, IL: Dryden Press, 1970.
13 Nunnally JC Jr. *Introduction to psychological measurement*. 2nd ed. New York: McGraw Hill, 1970.
14 Nunnally JC, Jr. *Psychometric theory*. 2nd ed. New York: McGraw-Hill, 1978.
15 Allen MJ, Yen WM. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole, 1979.
16 Carmines EG, Zeller RA. *Reliability and validity assessment*. London: Sage, 1979.
17 American Education Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for education and psychological testing*. Washington DC: American Psychological Association, 1985.
18 Anastasi A. *Psychological testing*. 6th ed. New York: Macmillan, 1988.
19 Kaplan RM, Saccuzzo DP. Psychological testing: principles, applications, and issues. 3rd ed. Pacific Grove, CA: Brooks/Cole, 1993.
20 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1995.
21 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
22 Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika* 1937;**2**:151–60.
23 Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychol Bull* 1979;**86**:420–8.
24 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;**20**:37–46.
25 Cohen J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213–20.
26 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**:378–82.
27 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 1983;**32**:307–17.
28 Bland JM, Altman DG. Statistical methods for assessing the agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.
29 Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. *Clinical Rehabilitation* 1994;**8**:233–39.
30 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
31 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;**56**:81–105.
32 Campbell DT. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *Am Psychol* 1960;**15**:546–553.
33 Messick S. Test validity and the ethics of assessment. *Am Psychol* 1980;**35**:1012–27.
34 Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronic Diseases* 1986;**39**:897–906.
35 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clin Trials* 1991;**12**:142–58S.
36 Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases* 1987;**40**:171–8.
37 Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989;**42**:403-8.
38 Kirshner B, Guyatt G. A methodological framework for assessing health indices. *Journal of Chronic Diseases* 1985;**38**:27–36.
39 Wilkin D, Hallam L, Doggett M-A. *Measures of need and outcome for primary health care*. Oxford: Oxford University Press, 1994.
40 Norman GR. Issues in the use of change scores in randomized trials. *J Clin Epidemiol* 1989;**42**:1097–105.
41 Ware JE. *Short form 36 health survey manual and interpretation guide*. Boston, MA: Nimrod Press, 1993.