Kurtzke scales revisited: the application of psychometric methods to clinical intuition

Jeremy Hobart, Jenny Freeman and Alan Thompson

Neurological Outcome Measures Unit, Institute of Neurology, London, UK

Correspondence to: Dr Jeremy Hobart, Neurological Outcome Measures Unit, Institute of Neurology, Queen Square, London WC1N 3BG, UK
E-mail: J.Hobart@ion.ucl.ac.uk

Summary
When developing his disability scales for multiple sclerosis, Kurtzke demonstrated perception and insight. However, 45 years later, the evaluation of his clinically derived scales remains limited, particularly for more disabled patients. Indeed, many of Kurtzke’s assumptions underpinning the development of the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) are untested. This study aims to build on previous work and provide a more detailed examination using psychometric methods of the EDSS and FS. There are three study objectives: (i) to examine comprehensively the psychometric properties of the EDSS in more disabled people with multiple sclerosis undergoing in-patient rehabilitation; (ii) to examine the reliability of the FS and test Kurtzke’s assumptions that they measure different aspects of the neurological examination and measure different constructs from that measured by the EDSS; and (iii) to examine whether the FS can be summed to generate a summary score. The EDSS was examined for its acceptability (score distributions), reliability (inter- and intra-rater reproducibility, standard error of measurement), validity (convergent and discriminant validity, measurement precision, discrimination between individuals) and responsiveness (effect size). The FS were examined for their reliability (inter- and intra-rater reproducibility), intercorrelations, correlations with the EDSS and the extent to which they satisfy Likert’s criteria as a summed rating scale. In this more disabled sample of people with multiple sclerosis, the EDSS is an acceptable measure but demonstrates limited variability. Inter-rater reproducibility (intraclass correlation coefficient; ICC = 0.78) is adequate for group comparison studies, but intra-rater reproducibility is variable (ICC = 0.62–0.94). Convergent and discriminant validity for the EDSS is supported, but its measurement precision relative to the Functional Independence Measure is limited (56%). Also, the EDSS has a limited ability to distinguish between individuals in terms of their disability and its responsiveness is poor (effect size = 0.10). Results indicate that the FS measure constructs distinct from each other (intercorrelations = -0.23 to +0.52) and from the EDSS (correlations = -0.10 to +0.59). Intra-rater, but not inter-rater reproducibility is adequate for group comparison studies. The FS do not satisfy criteria as an eight-, seven- or six-item summed rating scale. Despite being based on sound clinical intuition, the lack of psychometric input into the development of the EDSS and FS has limited their usefulness as evaluative outcome measures in multiple sclerosis.

Keywords: Kurtzke scales; psychometric evaluation; health measurement; multiple sclerosis; health status

Abbreviations: BI = Barthel Index; CI = confidence interval; DSS = Disability Status Scale; EDSS = Kurtzke Expanded Disability Status Scale; FIM = Functional Independence Measure; FS = Kurtzke Functional Systems; GHQ = General Health Questionnaire; ICC = intraclass correlation coefficient; LHS = London Handicap Scale; MSFC = Multiple Sclerosis Functional Composite; SEM = standard error of measurement; SF-36 = Medical Outcomes Study Short-Form 36-Item Health Survey; SF-36 PCS = SF-36 Physical Component Summary Score; SF-36 MCS = SF-36 Mental Component Summary Score; SHO = senior house officer.

Introduction
When developing his disability scales, Kurtzke demonstrated perception and insight. He recognized that the effectiveness of therapeutic interventions could be determined only if disease severity could be quantified accurately (Kurtzke, 1961). As no direct measure of disease severity existed for multiple sclerosis, he proposed an indirect method: measuring disability using a quantitative neurological examination. Arguing that the absence of generally accepted disability...
scales for multiple sclerosis was a major factor hindering the evaluation of therapeutic interventions (Kurtzke, 1955), but acknowledging that disability can bear only a loose relationship to the volume of diseased tissue present (Kurtzke, 1961), Kurtzke developed his own measurement instruments.

There are three Kurtzke scales: (i) the Disability Status Scale (DSS) (Kurtzke, 1955) is an 11-point scale (0 = normal neurological examination; 10 = death due to multiple sclerosis) measuring overall disability, developed to evaluate the effectiveness of isoniazid as a treatment for multiple sclerosis (Kurtzke and Berlin, 1954); (ii) a modification of the DSS, which has replaced it, called the 20-point Expanded Disability Status Scale (EDSS) (Kurtzke, 1983); (iii) the Functional Groups [more commonly known as Functional Systems (FS)] are eight scales representing different functions of the CNS (Kurtzke, 1961). Each system is rated on a five-point (three systems) or six-point (four systems) response scale except ‘Other Functions’ which is rated dichotomously (0 = none, 1 = any other neurological findings attributed to multiple sclerosis).

Although the DSS (1954) was published 7 years before the FS (1961), Kurtzke’s writings indicate that these two instruments are intimately related. The DSS was designed to represent the sum of a person’s neurological dysfunction (Kurtzke and Berlin, 1954), and measure their maximal function given the neurological deficits (Kurtzke, 1955). It was devised to depict, in the fewest number of steps (Kurtzke, 1961), the progression of multiple sclerosis as it usually occurs, with an emphasis on ambulation (Kurtzke and Berlin, 1954), but with the provision of grades for less involved individuals (Kurtzke, 1961). In contrast, the FS were intended to delineate the type and severity of eight neurological impairments (Kurtzke, 1961). Together, the DSS and FS were intended to form a complementary measurement method for quantifying the objective and verifiable deficits due to multiple sclerosis, as elicited from the neurological examination, in a reasonable and reproducible manner (Kurtzke, 1961). Moreover, the FS were designed to provide two useful checks on the DSS. For less disabled patients, the DSS score should be greater than or equal to the highest score in any individual FS. No change in DSS number should occur without a change in one or more FS (Kurtzke, 1961).

The DSS was developed to enable comparisons of disability within and between patients. Kurtzke argued that an overall index of disability was required for studies in multiple sclerosis as the eight FS could not be summed to indicate the total disorder (Kurtzke, 1961). He cited five reasons in support of this argument. First, the eight FS are independent. Secondly, they are not equivalent to each other in representing the amount of disease present. Thirdly, clinical signs may follow different courses allowing fluctuations to be obscured by an unchanging total. Fourthly, a total score can improve while the patient worsens (e.g. when increasing weakness obscures cerebellar signs). Finally, Kurtzke believed it was not appropriate to sum scores from ordinal scales (Kurtzke, 1983).

Users of the DSS argued that it was too insensitive in its middle range (especially step 7) and too limited for studies of chronic multiple sclerosis. These criticisms prompted Kurtzke to develop the EDSS by dividing each DSS step (except the first) in half. He reasoned that the Gaussian distribution of DSS scores in his study samples provided empirical evidence that no step was discrepant (Kurtzke, 1983). Although the EDSS has also attracted criticism (Willoughby and Paty, 1988; Noseworthy et al., 1990; Goodkin et al., 1992; Sharrack and Hughes, 1996), it currently is the most widely used disability measure in clinical trials of multiple sclerosis (The IFNB Multiple Sclerosis Study Group, 1993; Jacobs et al., 1996).

As well as advocating disability measurement, Kurtzke recognized that rating scales must fulfil clinical and scientific criteria. He noted they should be user-friendly (Kurtzke, 1955), applicable to all patients (Kurtzke and Berlin, 1954; Kurtzke, 1961), and reproducible (Kurtzke, 1955). He stated that the sum total of any patient’s disabilities should fit them into a suitable category on a scale (Kurtzke, 1955) and, that any change in disability should be reflected in a change of status on that scale (Kurtzke, 1955). In fact, he says that there is only one good reason for quantitative schemes in multiple sclerosis: to document change (Kurtzke, 1965). In short, Kurtzke appreciated that scales should be clinically useful (capable of being incorporated into routine clinical practice) and scientifically sound (acceptable, reliable, valid and responsive).

Despite his insight, Kurtzke did not examine the measurement properties of his scales as the expertise necessary was not readily accessible. While methods for developing and evaluating measures of psychological constructs (psychometric methods) were already established in the social sciences (Guilford, 1936), these methods were only applied to medicine in the 1970s (Brook et al., 1979). Textbooks on health measurement did not appear until the late 1980s (McDowell and Newell, 1987; Streiner and Norman, 1989), and formal guidelines for the development and evaluation of health measures have been proposed only recently (Scientific Advisory Committee of the Medical Outcomes Trust, 1995; McDowell and Jenkinson, 1996; Fitzpatrick et al., 1998).

In the last decade, increasing awareness of psychometric methods has resulted in a number of studies examining the measurement properties of the EDSS and the FS (e.g. Amato et al., 1987, 1988; Noseworthy et al., 1990; Francis et al., 1991; Verdier-Taillefer et al., 1991; Goodkin et al., 1992; Marolf et al., 1996; Sharrack et al., 1996b, 1999). Most of these studies have concentrated on reliability and not examined validity or responsiveness. This approach is limited because psychometric properties are sample and not instrument dependent (Cronbach, 1949), reliability is a necessary but not sufficient condition for validity (Nunnally, 1978), and adequate reliability and validity do not guarantee responsiveness (Guyatt et al., 1987). Consequently, all relevant measurement properties must be examined comprehensively in patient samples. In a recent article,
Sharrack and colleagues (Sharrack et al., 1999) published the first study to examine the reliability, validity and responsiveness of the EDSS, and the reliability of the FS in the same sample and provides a good base for further, more detailed psychometric evaluations of these instruments.

In this article, the psychometric properties of the EDSS are examined comprehensively. In contrast to other studies: the sample is patients who are more disabled undergoing rehabilitation; acceptability is examined; reliability is determined in the clinical context in which the instrument is used routinely; the validation strategy addresses specifically the purposes for which the EDSS was designed; and responsiveness is determined prospectively. The reliability of the FS is examined and Kurtzke’s assumptions that the FS measure eight different aspects of the neurological examination and address different constructs from that measured by the EDSS are tested. Finally, the issue of whether the FS satisfy criteria as a summed rating scale is addressed.

Method

Samples

Patients were recruited with clinically definite multiple sclerosis (Poser et al., 1983) undergoing in-patient neurorehabilitation at the National Hospital for Neurology and Neurosurgery in London. The ethical committee of this hospital gave approval, and informed consent was obtained from all subjects. Details of patient selection criteria and the rehabilitation process are reported elsewhere (Freeman et al., 1997).

Outcome measures

The following outcome measures were administered on admission and discharge in strict accordance with the developer’s guidelines.

Barthel Index (BI) (Mahoney and Barthel, 1965)
The BI measures disability as independence in 10 activities of daily living. Items are rated from behavioural observation on a two-point (two items), three-point (six items) or four-point (two items) response scale and summed to generate a total score (low scores indicate greater disability). Wade’s version of the BI was used (Collin et al., 1988), and has been shown to be reliable, valid and responsive (Wade, 1992; van Bennekom et al., 1996).

Functional Independence Measure (FIM) (Granger et al., 1986)
The FIM measures disability as the assistance required to perform 18 tasks. Items are rated on a seven-point scale and summed to generate a total score (low scores indicate greater disability). The FIM has been shown to be reliable, valid and responsive in people with multiple sclerosis (Granger et al., 1990; Brosseau, 1994).

London Handicap Scale (LHS) (Harwood and Ebrahim, 1995)
The LHS is a six-item self-report measure of handicap. Items are rated on a six-point weighted scale and summed to generate a total score (high scores indicate greater handicap). The LHS is reliable, valid and responsive.

Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) (Ware, 1993)
The SF-36 measures self-reported health status in eight dimensions or two summary scores [physical component summary score (PCS) and mental component summary score (MCS)] (Ware et al., 1994). Low scores indicate worse health. The reliability and validity of the SF-36 are the subject of numerous studies. In this study, summary scores are reported.

General Health Questionnaire (GHQ) (Goldberg and Hillier, 1979)
The GHQ is a 28-item self-report measure of psychological distress. Items are rated on a two-point scale and summed to generate a total score (high scores indicate greater distress). Evidence supports its reliability and validity (McDowell and Newell, 1987).

Staff-rated transition question (Fitzpatrick et al., 1993)
The transition question used in this study is a four-point staff rating of change in disability at discharge (1 = no change; 4 = marked improvement).

EDSS

Four psychometric properties were studied: acceptability, reliability, validity and responsiveness.

Acceptability

This concerns the extent to which the spectrum of health status measured by an instrument is representative of the distribution of health status in the sample (Ware et al., 1978). It is determined by examining the score distributions. An instrument is considered acceptable when: scores demonstrate good variability and span the full scale range (Ware et al., 1980; Stewart and Ware, 1992); mean scores are situated near the scale mid-point (Eisen et al., 1979); floor and ceiling effects, calculated as the percentage of responses for the
Reliability

The scores generated by a measurement instrument represent the sum of two components: true score and random error (Guilford, 1936). Reliability refers to the extent to which an instrument is free from random error (Nunnally, 1978). As reliability increases (or decreases), scores are more (or less) consistent and, therefore, measured variance reflects true variance (or random error) in the construct (Stewart and Ware, 1992). In keeping with this definition, reliability coefficients are estimates of the proportion of total score variance that is due to true score variance.

Reliability studies aim to quantify the main sources of random error associated with an instrument (Stanley, 1971). Variability within (intra-rater) and between (inter-rater) observers are the major sources of random error associated with observer-rated instruments such as the EDSS and FS (Shumaker et al., 1997). The most appropriate method of estimating reliability of single item measures such as the EDSS and individual FS is the test–retest reproducibility method (Stewart and Ware, 1992) which examines the agreement between paired ratings of patients and attributes measured variance to random error (Nunnally, 1978).

Inter-rater reproducibility was estimated by calculating the agreement between independent ratings of patients generated within 48 h of admission by J.H. and the neurology senior house officer (SHO). Intra-rater reproducibility was estimated for both raters on the same subsample of patients by calculating the agreement between paired ratings generated 48–72 h apart. All reproducibility results are reported as intraclass correlation coefficients (ICC; Bartko, 1966), for which the estimates of variance were obtained from a repeated measures analysis of variance table using a fixed effects model (Shrout and Fleiss, 1979). Minimum recommended standards for reproducibility are 0.70 for group comparisons and 0.90–0.95 for individual comparisons (Nunnally, 1978).

As the outcomes of some studies of multiple sclerosis are based on individual EDSS score changes, reproducibility estimates are also reported as standard errors of measurement (SEM). These estimate the standard deviation (SD) of scores obtained if an instrument was administered to the same individual multiple times. Consequently, they are used to determine confidence intervals (CI) around individual patient scores (Guilford, 1954), reflect an instrument’s accuracy for individual patient assessment and clinical decision-making (Williams and Naylor, 1992), and gauge the likelihood that an individual patients’ change in score is attributable to true change (McHorney and Tarlov, 1995). The following formulae are used (Anastasi and Urbina, 1997):

\[
\text{SEM} = \frac{\text{SD}}{\sqrt{n}}
\]

95% CI around an individual score = true score ± 1.96 SEM.

Validity

This is defined as the extent to which an instrument measures the construct it purports to measure (American Educational Research Association, National Council on Measurement used in Education, 1955). There are many methods of gathering evidence for the validity of a measure, but construct validation is used when no criterion (gold standard) or content domain is accepted as entirely adequate to define the attribute being measured. Construct validity is defined as the extent to which empirical data support hypotheses concerning the construct the instrument is purported to measure (Cronbach and Meehl, 1955). As there are many types of evidence under the rubric of construct validity, and validity is not a fixed measurement property, a strategy is required that examines the validity of an instrument for the specific purposes and specific setting in which it is being used (McDowell and Jenkinson, 1996). In accordance with Kurtzke’s recommendation (Kurtzke, 1961), this study examines the validity of the EDSS as an overall measure of disability, and evaluates its ability to detect differences between groups and individuals with multiple sclerosis on account of their disability.

The extent to which the EDSS is a measure of disability, and not a measure of related health constructs, is determined by examining convergent and discriminant validity (Cronbach and Meehl, 1955). In this analysis, correlations are examined between the EDSS and other measures and variables. The extent to which their direction, magnitude and pattern conform with a priori hypotheses indicates the strength of this validity evidence. If the EDSS is a measure of overall disability, four findings are predicted.

(i) Correlations with disability measures (BI and FIM) should be high \((r > 0.80)\).

(ii) Correlations with measures of handicap (LHS), mental health status (SF-36 MCS) and psychological distress (GHQ) should be low \((r < 0.30)\). Correlations between the EDSS and are also predicted to be low as the SF-36 PCS scale measures physical role limitations, pain and general health perceptions as well as physical functioning.

(iii) Correlations between the EDSS and the LHS and SF-36 PCS should exceed correlations between the EDSS and the SF-36 MCS and GHQ. This is because handicap and physical health status are more closely related conceptually to disability than mental health status and psychological distress.

(iv) The EDSS should be uncorrelated with age \((r < 0.10)\) as disability in this sample of multiple sclerosis patients is not expected to be biased by this variable.

The extent to which the EDSS is capable of discriminating between groups of multiple sclerosis patients on the basis of their disability is determined by comparing its measurement precision, relative to the FIM and BI, for detecting change in disability due to rehabilitation. On discharge from the
Consequently, effect sizes for the EDSS, BI and FIM are if they vary (Ware sizes provide limited information about the responsiveness (Nunnally, 1978; Scienti rehabilitation. examined. First, all eight FS are examined. Next, because compared in order to determine the relative ability of these The extent to which the FS satisfy scaling criteria as instrument to detect change (Norman induced by the intervention, as well as the ability of the exceeding 0.30, the criteria of equivalent item means, variances deviation of the admission scores (Kazis et al. et al. The higher the F-statistic, the greater the measurement precision. By comparing the EDSS, BI and FIM in the same sample, relative measurement precision is estimated as the ratio of pairwise F-statistics (F for one measure divided by F for another) and indicates, as a percentage, how much more (or less) precise one measure is compared with another at detecting group differences (McHorney et al., 1992). In this study, the instrument with the largest F-statistic is chosen as the arbitrary standard and assigned a relative measurement precision of 1.

The extent to which the EDSS is able to discriminate between individuals on account of their disability is determined by examining the variability of BI and FIM scores for patients scoring at each EDSS level. The ability of the EDSS to discriminate between individuals is inversely related to the variability of BI and FIM scores.

**Responsiveness**

This is defined as the ability of an instrument to detect clinically significant change in the construct measured, even if that change is small (Guyatt et al., 1987). It is often determined by comparing scores before and after an intervention expected to alter the quantity being measured, and calculating an effect size (standardized change score) (Scientific Advisory Committee of the Medical Outcomes Trust, 1995). There are many effect size calculations. In this study, it is defined as the mean change score (admission minus discharge EDSS scores) divided by the standard deviation of the admission scores (Kazis et al., 1989): larger effect sizes indicate greater responsiveness. However, effect sizes provide limited information about the responsiveness of measures as they reflect the magnitude of the change induced by the intervention, as well as the ability of the instrument to detect change (Norman et al., 1997). Consequently, effect sizes for the EDSS, BI and FIM are compared in order to determine the relative ability of these instruments to detect change in patients undergoing rehabilitation.

**FS Reliability**

The inter- and intra-rater reproducibility of the FS was examined using the same method described above for the EDSS.

**Relationships among the FS and between the FS and EDSS**

Relationships among the FS are determined by examining their intercorrelations. The magnitude of the correlations indicates the strength of the relationships. Kurtzke’s assumption that the FS measure independent aspects of the neurological examination is supported when their intercorrelations are positive and low to moderate (<0.60). Relationships between the FS and EDSS are determined by examining their intercorrelations. Kurtzke’s assumption that the EDSS measures a different construct from each of the individual FS is supported when correlations are positive and low to moderate (<0.60). It is predicted that correlations between the EDSS and FS should exceed intercorrelations among the FS as the EDSS is purported to represent the sum of the disabilities defined by patients’ neurological deficits.

**Do the FS constitute a summed rating scale?**

Kurtzke hypothesized that the FS could not be summed to indicate the total disorder (Kurtzke, 1961). However, it is appropriate to examine this untested assumption because the FS were developed to measure different aspects of the same underlying construct (the neurological examination), and summed rating scales have superior measurement properties to single item measures (Nunnally, 1978). Likert demonstrated that items with ordinal response categories can be summed to generate total scores when they: measure the same underlying construct; measure at similar points on a scale; have similar variances; and contribute equal proportions of information to the total score (Likert, 1932). These four criteria are satisfied when items are internally consistent (inter-related) and have equivalent mean scores, variances and corrected item–total correlations. A group of items is internally consistent when: the mean item intercorrelation exceeds 0.30 (Eisen et al., 1979); correlations between each item and the total score computed from the sum of the remaining items (corrected item–total correlation; Howard and Forehand, 1962) exceed 0.40 (McHorney et al., 1994); and when the Cronbach’s alpha coefficient (Cronbach, 1951) exceeds 0.70 (Nunnally, 1978; Scientific Advisory Committee of the Medical Outcomes Trust, 1995). When item–total correlations exceed 0.30, the criteria of equivalent item means, variances and item–total correlations can be considered satisfied, even if they vary (Ware et al., 1997).

The extent to which the FS satisfy scaling criteria as eight-item, seven-item and six-item summed rating scales is examined. First, all eight FS are examined. Next, because ‘Other Functions’ is scored dichotomously rather than on a multi-point response scale and does not measure a specific aspect of the neurological examination, scaling criteria are examined for the remaining seven FS. Finally, ‘Mental Functions’ is also removed, as it is rarely reported in studies of the FS, and scaling criteria are examined for the remaining six items.
Table 1 Characteristics of patient samples

<table>
<thead>
<tr>
<th>Samples</th>
<th>n</th>
<th>Age mean (SD)</th>
<th>Females n (%)</th>
<th>MS disease type</th>
<th>EDSS score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>n (%)</td>
<td>RR*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n (%)</td>
<td>Progressive†</td>
<td>Range</td>
</tr>
<tr>
<td>MS population</td>
<td>311</td>
<td>42.9 (11.7)</td>
<td>212 (68.2)</td>
<td>42 (13.5)</td>
<td>5.0–9.0</td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>125</td>
<td>43.8 (11.6)</td>
<td>80 (64.0)</td>
<td>9 (7.2)</td>
<td>5.0–9.0</td>
</tr>
<tr>
<td>Intra-rater reliability</td>
<td>40</td>
<td>45.8 (11.4)</td>
<td>29 (72.5)</td>
<td>4 (10.0)</td>
<td>5.5–9.0</td>
</tr>
<tr>
<td>Validity/responsiveness</td>
<td>64</td>
<td>43.0 (11.9)</td>
<td>41 (64.1)</td>
<td>5 (7.8)</td>
<td>5.0–9.0</td>
</tr>
</tbody>
</table>

*Relapsing–remitting MS; †both primary and secondary progressive multiple sclerosis; ‡multiple sclerosis patients admitted between 1993 and 1998.

Data collection
Consecutive admissions were enrolled between April 1, 1994 and April 1, 1996. Inter-rater reproducibility of the EDSS and FS was determined for all patients, together with intra-rater reproducibility on consecutive admissions between May 1995 and February 1996. BI, FIM, LHS, SF-36, GHQ and staff-rated transition question scores were collected on patients randomly selected to participate in a larger study evaluating the measurement properties of disability measures. Admission and discharge EDSS and FS scores were collected independently by J.H. and the neurology SHO at the Neurorehabilitation Unit. BI and FIM ratings were undertaken by the treating multi-disciplinary team for each patient. All raters received training in scoring Kurtzke scales and the FIM.

Results
Samples
Of the 137 patients admitted with clinically definite multiple sclerosis during the study period, none declined to participate. Inter-rater reproducibility was examined in 125 patients (91%) and intra-rater reproducibility in 40. Sixty-four patients participated in the evaluation of the measurement properties of disability measures. Table 1 presents age, sex, disease type and EDSS scores for the samples used in this study and compares them with all patients admitted between 1993 and 1998. Results indicate that all samples are representative of the multiple sclerosis population treated at the Neurorehabilitation Unit.

EDSS
Acceptability
Table 2 demonstrates that EDSS scores span only the upper half of the scale range (5.0–9.0), their standard deviation is small (1.0) and the mean score (7.1) is above the scale midpoint. There are no floor or ceiling effects, and the skewness statistic (0.164) lies within the recommended range. These results indicate that this sample is moderate to severely disabled, that the range of disability defined by the EDSS is greater than that of the study sample, but that scores are not skewed excessively. The FIM and BI demonstrate greater score variability than the EDSS and still have small floor and ceiling effects. These findings indicate the FIM and BI have acceptability superior to the EDSS in this sample.

Reliability
Eleven raters were used, i.e. J.H. and 10 consecutive SHOs. Table 3 demonstrates that EDSS inter-rater reproducibility satisfies the recommended criterion for group (0.70), but not individual (0.90–0.95) comparison studies. Table 4 demonstrates that EDSS intra-rater reproducibility is variable and ranges from satisfying criteria for individual and group comparison studies (J.H.) to satisfying neither criteria (SHO). Ninety-five per cent CI vary accordingly: when different raters are involved, they span almost two EDSS points (four levels). Even when reliability is very high, CI are almost one EDSS point, while for a less experienced rater CI are larger.

Validity
Convergent and discriminant validity (Table 5). The EDSS correlates highly with the BI and the FIM, and poorly with the LHS, SF-36 PCS, SF-36 MCS, GHQ and age. The direction, pattern and magnitude of correlations are consistent with a priori hypotheses. These findings provide evidence for convergent and discriminant validity and indicate that the EDSS measures overall disability, and discriminates disability from related health constructs.

Measurement precision (Table 6). Transition question ratings are available for 59 of the 64 patients (92%). As the FIM has the largest F-statistic, it is nominated as the arbitrary standard and assigned a measurement precision of 1. The relative measurement precision of the BI (74%) and EDSS (56%) are notably less, indicating that the EDSS is least able to discriminate between patients in terms of disability.

Discrimination between individuals. Table 2 demonstrates that the BI and FIM have greater score variability than the EDSS, suggesting that they are better able to discriminate between individual patients on the basis of disability. This is confirmed by Table 7 which demonstrates...
Table 2 Descriptive statistics for EDSS, BI and FIM admission scores (n = 137)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Score range</th>
<th>Mean (SD)</th>
<th>Floor effect n (%)</th>
<th>Ceiling effect n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDSS</td>
<td>0–10</td>
<td>5.0–9.0</td>
<td>7.1 (1.0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>BI</td>
<td>0–20</td>
<td>0–20</td>
<td>12.02 (5.7)</td>
<td>3 (1.5)</td>
</tr>
<tr>
<td>FIM</td>
<td>18–126</td>
<td>24–122</td>
<td>89.42 (23.6)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Table 3 Inter-rater reproducibility of the EDSS and FS (n = 125)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Mean score (SD)</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDSS</td>
<td>7.09 (1.01)</td>
<td>6.98 (1.11)</td>
</tr>
</tbody>
</table>

Functional System

<table>
<thead>
<tr>
<th>System</th>
<th>Functional System</th>
<th>Rater 1*</th>
<th>Rater 2†</th>
<th>ICC</th>
<th>SEM (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyramidal</td>
<td>3.74 (0.91)</td>
<td>3.57 (1.2)</td>
<td>0.67</td>
<td>0.52 (±1.02)</td>
<td></td>
</tr>
<tr>
<td>Cerebellar</td>
<td>2.42 (1.25)</td>
<td>2.44 (1.31)</td>
<td>0.61</td>
<td>0.99 (±1.94)</td>
<td></td>
</tr>
<tr>
<td>Brainstem</td>
<td>1.78 (0.88)</td>
<td>1.38 (1.24)</td>
<td>0.61</td>
<td>0.56 (±1.08)</td>
<td></td>
</tr>
<tr>
<td>Visual</td>
<td>1.82 (1.54)</td>
<td>1.43 (1.63)</td>
<td>0.56</td>
<td>1.02 (±2.00)</td>
<td></td>
</tr>
<tr>
<td>Bladder + bowels</td>
<td>2.66 (1.48)</td>
<td>2.45 (1.71)</td>
<td>0.72</td>
<td>0.78 (±1.53)</td>
<td></td>
</tr>
<tr>
<td>Sensory</td>
<td>2.12 (1.28)</td>
<td>2.30 (1.26)</td>
<td>0.38</td>
<td>1.29 (±2.53)</td>
<td></td>
</tr>
<tr>
<td>Mental</td>
<td>0.99 (1.09)</td>
<td>1.30 (1.30)</td>
<td>0.51</td>
<td>0.76 (±1.50)</td>
<td></td>
</tr>
</tbody>
</table>

*J.H.; †SHO.

Table 4 Intra-rater reproducibility of the EDSS and FS (n = 40)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Mean score (SD)</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDSS</td>
<td>7.08 (0.99)</td>
<td>7.11 (0.94)</td>
</tr>
</tbody>
</table>

FS scale

<table>
<thead>
<tr>
<th>System</th>
<th>Functional System</th>
<th>Time 1</th>
<th>Time 2</th>
<th>ICC</th>
<th>SEM (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyramidal</td>
<td>3.70 (0.99)</td>
<td>3.8 (0.88)</td>
<td>0.78</td>
<td>0.46 (±0.91)</td>
<td></td>
</tr>
<tr>
<td>Cerebellar</td>
<td>2.15 (0.99)</td>
<td>2.26 (1.01)</td>
<td>0.72</td>
<td>0.52 (±1.03)</td>
<td></td>
</tr>
<tr>
<td>Brainstem</td>
<td>1.63 (0.81)</td>
<td>1.65 (0.92)</td>
<td>0.79</td>
<td>0.37 (±0.73)</td>
<td></td>
</tr>
<tr>
<td>Visual</td>
<td>2.00 (1.50)</td>
<td>2.05 (1.58)</td>
<td>0.67</td>
<td>0.86 (±1.69)</td>
<td></td>
</tr>
<tr>
<td>Bladder + bowels</td>
<td>2.58 (1.48)</td>
<td>2.78 (1.69)</td>
<td>0.70</td>
<td>0.81 (±1.59)</td>
<td></td>
</tr>
<tr>
<td>Sensory</td>
<td>2.05 (1.18)</td>
<td>2.35 (1.08)</td>
<td>0.57</td>
<td>0.77 (±1.52)</td>
<td></td>
</tr>
<tr>
<td>Mental</td>
<td>0.75 (1.03)</td>
<td>0.50 (0.88)</td>
<td>0.28</td>
<td>0.87 (±1.71)</td>
<td></td>
</tr>
</tbody>
</table>

*J.H.; †SHO.

Table 5 Product–moment correlations between EDSS, other health measures and age (n = 64)

<table>
<thead>
<tr>
<th>Disability</th>
<th>Handicap*</th>
<th>Health status*</th>
<th>Psychological well-being†</th>
<th>Age well-being†</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>FIM</td>
<td>LHS SF-36 PCS</td>
<td>SF-36 GHQ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.89‡ -0.84</td>
<td>-0.33</td>
<td>-0.22 -0.06 0.06</td>
</tr>
</tbody>
</table>

* n = 60; † n = 53; ‡negative correlations with the EDSS occur for measures where low scores indicate worse health.

that each EDSS score is associated with a wide range of BI and FIM scores. These findings indicate that patients have variable levels of disability that are not distinguished by the EDSS. This analysis was undertaken on the total sample of patients admitted between 1993 and 1998 so that the sample sizes for each EDSS level were maximized.

Responsiveness Table 8 demonstrates that the EDSS has a small effect size in this sample, indicating poor responsiveness. Effect sizes for the BI and FIM are considerably greater.

FS Reliability Table 3 demonstrates that only the inter-rater reproducibility of the Bladder and Bowels FS satisfies the criterion for group
Table 6 Relative measurement precision estimates for the EDSS, FIM and BI

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean change score (SD)</th>
<th>F-statistic (P)*</th>
<th>Relative precision†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No or minimal change in disability (n = 29)</td>
<td>Moderate or marked change in disability (n = 30)</td>
<td></td>
</tr>
<tr>
<td>FIM</td>
<td>–4.76 (7.65)</td>
<td>–7.40 (8.21)</td>
<td>1.63 (0.21)</td>
</tr>
<tr>
<td>BI</td>
<td>–1.69 (1.95)</td>
<td>–2.33 (2.50)</td>
<td>1.21 (0.28)</td>
</tr>
<tr>
<td>EDSS</td>
<td>–0.02 (0.25)</td>
<td>0.12 (0.72)</td>
<td>0.91 (0.34)</td>
</tr>
</tbody>
</table>

*Ratio of between-groups variability to within-group variability; †pairwise F-statistics compared with FIM.

Table 7 BI and FIM score ranges and means for each EDSS score (n = 311)

<table>
<thead>
<tr>
<th>EDSS score</th>
<th>n</th>
<th>BI score</th>
<th>FIM score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>Mean</td>
</tr>
<tr>
<td>5.0</td>
<td>3</td>
<td>19–20</td>
<td>N/A*</td>
</tr>
<tr>
<td>5.5</td>
<td>6</td>
<td>14–19</td>
<td>N/A</td>
</tr>
<tr>
<td>6.0</td>
<td>49</td>
<td>8–20</td>
<td>17.9</td>
</tr>
<tr>
<td>6.5</td>
<td>80</td>
<td>8–20</td>
<td>15.6</td>
</tr>
<tr>
<td>7.0</td>
<td>48</td>
<td>5–20</td>
<td>12.3</td>
</tr>
<tr>
<td>7.5</td>
<td>37</td>
<td>2–18</td>
<td>10.1</td>
</tr>
<tr>
<td>8.0</td>
<td>54</td>
<td>2–16</td>
<td>8.1</td>
</tr>
<tr>
<td>8.5</td>
<td>23</td>
<td>0–11</td>
<td>3.7</td>
</tr>
<tr>
<td>9.0</td>
<td>11</td>
<td>0–5</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Not applicable due to small sample.

Table 8 Admission, discharge, change scores and responsiveness indices for EDSS, BI and FIM (n = 64)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean score (SD)</th>
<th>Effect size*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admission</td>
<td>Discharge</td>
</tr>
<tr>
<td>EDSS</td>
<td>7.1 (1.0)</td>
<td>7.0 (1.1)</td>
</tr>
<tr>
<td>BI</td>
<td>12.7 (5.7)</td>
<td>14.9 (5.7)</td>
</tr>
<tr>
<td>FIM</td>
<td>91.2 (24.1)</td>
<td>97.6 (24.2)</td>
</tr>
</tbody>
</table>

*Calculated as mean change score divided by the standard deviation of admission scores; †change scores = admission score minus discharge score; negative change scores arise as high scores on the BI and FIM indicate less disability, the reverse of the EDSS.

comparisons. However, Table 4 demonstrates that intra-rater reliability for most FS satisfy this criterion. In contrast to the EDSS, intra-rater reliability of the FS is generally highest for the SHO.

Intercorrelations among the FS

Table 9 demonstrates that intercorrelations among the eight FS range from –0.23 to +0.52, indicating that all relationships are weak or moderate. The majority of correlations (89%) are ≤0.37, indicating weak relationships. These findings support Kurtzke’s hypothesis that these eight scales measure different constructs. The correlation between Pyramidal and Cerebellar Functions is –0.23, indicating a weak negative relationship.

Correlations between the FS and EDSS

Table 9 demonstrates that correlations between the EDSS and all FS range from –0.10 to 0.59. Five of the eight correlations are in the moderate range (0.30–0.60). It is notable that Other and Mental Functions have weak correlations with the EDSS. For 50% of the FS, correlations with the EDSS exceed correlations with other FS.

The FS as a summed rating scale

Table 10 demonstrates that the FS fail to satisfy criteria as eight-, seven- or six-item summed rating scales because mean item intercorrelations are <0.30 and alpha coefficients are <0.70. Furthermore, the alpha coefficients do not increase substantially when each FS is omitted (alpha with item deleted), indicating that none of them is compromising the internal consistency of the scales. These findings support Kurtzke’s assumption that the FS cannot be summed to generate a total score.

Discussion

It is increasingly recognized that the evaluation of therapeutic interventions for multiple sclerosis should include measures of patient-oriented outcomes such as disability. If these outcomes are to influence health policy, neurologists must be confident that the data are rigorous. Psychometric methods enable clinicians to develop scientifically sound measures and evaluate the measurement properties of instruments, such as the Kurtzke scales, that were developed before these techniques were available. To date, the psychometric evaluation of the EDSS and FS has been limited.

The findings of this study serve to highlight both the strengths and weaknesses of the EDSS and FS in disability measurement. The EDSS addresses a broader spectrum of disability than other measures, has inter-rater reproducibility adequate for group comparison studies, measures overall disability and discriminates disability from other health constructs. However, its intra-rater reproducibility is variable and, compared with other disability measures, the EDSS has
a limited ability to distinguish between individuals or groups on the basis of disability and has poor responsiveness. The FS measure eight distinct constructs that are different from the construct measured by the EDSS. Although the inter-rater reproducibility of most of the FS satisfies criteria for group comparison studies, their inter-rater reproducibility does not. They do not constitute an eight-, seven- or six-item summed rating scale.

Whilst it is common for studies to document a measure of central tendency for scores, it is uncommon for acceptability data to be reported comprehensively. However, these descriptive statistics indicate the probable relevance of an instrument to a clinical setting before a full psychometric evaluation is undertaken. For example, the ceiling effect of admission scores represents a subsample of patients with no potential to improve their score regardless of any change induced by the intervention. Consequently, ceiling effects adversely influence responsiveness. Furthermore, acceptability data aid the interpretation of other psychometric analyses. For example, correlations between measures are affected by the distribution of their data (Nunnally, 1975). When the score distributions for two measures are different (e.g. one skewed, the other distributed more normally), their correlation will be attenuated. As the selection criteria for studies of multiple sclerosis inevitably limit the spectrum of disease severity included, acceptability data are particularly relevant.

Previous reliability studies of the EDSS and FS have generated variable results. Comparison of these studies is limited as they have used different statistics, methods and samples. Results have been reported as ICCs (Goodkin et al., 1992; Sharrack et al., 1999), Kappa (K) statistics (Amato et al., 1988; Noseworthy et al., 1990; Francis et al., 1991; Sharrack et al., 1996b, 1999) and weighted Kappa (Kw) statistics (Amato et al., 1987). Sharrack and colleagues demonstrate that ICCs and K generate different results (Sharrack et al., 1999). Even studies using recommended methods (ICC or Kw; Scientific Advisory Committee of the Medical Outcomes Trust, 1995; Streiner and Norman, 1995; McDowell and Jenkinson, 1996) report variable results. This finding may indicate that EDSS reliability is not stable across samples with different disability levels (e.g. the EDSS range 1.0–3.5 was studied by Goodkin et al., 1992; the EDSS ranges 0–9.5 and 0–7.5 were studied by Sharrack et al., 1999). Alternatively, inconsistent reliability results may reflect the different numbers of raters used, their level of training or expertise, or the sample size. Whilst there are no absolute guidelines for these variables, it is essential that the method used to determine reliability, and the sample in which it is studied, are representative of those in which the instrument is to be used.

SEM's have not been reported before for Kurtzke scales. Instead, most studies report percentage agreement within given EDSS and FS ranges (e.g. Amato et al., 1987; Francis et al., 1991; Goodkin et al., 1992). Although this method seems intuitively sensible, questions have been raised about its usefulness as an index of reliability because it does not correct for chance agreement (Cohen, 1960). SEM's have the additional advantage of generating CI which are more familiar to clinicians than reliability coefficients (Streiner and Norman, 1995).

Relatively little attention has focused on the validity of

Table 9 Intercorrelations among the FS, and between the FS and EDSS* (n = 137)

<table>
<thead>
<tr>
<th>FS</th>
<th>FS</th>
<th>EDSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>1.0</td>
<td>0.59</td>
</tr>
<tr>
<td>CF</td>
<td>-0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>BF</td>
<td>0.52</td>
<td>0.33</td>
</tr>
<tr>
<td>BBF</td>
<td>0.15</td>
<td>0.50</td>
</tr>
<tr>
<td>SF</td>
<td>0.29</td>
<td>0.31</td>
</tr>
<tr>
<td>VF</td>
<td>0.31</td>
<td>0.34</td>
</tr>
<tr>
<td>MF</td>
<td>1.0</td>
<td>0.17</td>
</tr>
</tbody>
</table>

*All correlations are product–moment correlations.

Table 10 Examination of the FS as an eight-, seven- and six-item summed rating scale (n = 137)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Item mean score range</th>
<th>Item variance range</th>
<th>Item intercorrelation mean (range)</th>
<th>Corrected item–total correlation range (mean)</th>
<th>Alpha coefficient if item deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eight-item*</td>
<td>0.07–3.74</td>
<td>0.06–2.55</td>
<td>0.14 (–0.23 to 0.52)</td>
<td>0.04–0.44 (0.29)</td>
<td>0.59</td>
</tr>
<tr>
<td>Seven-item†</td>
<td>1.01–3.37</td>
<td>0.77–2.55</td>
<td>0.18 (–0.23 to 0.52)</td>
<td>0.22–0.44 (0.32)</td>
<td>0.60</td>
</tr>
<tr>
<td>Six-item‡</td>
<td>1.74–3.74</td>
<td>0.77–2.55</td>
<td>0.18 (–0.23 to 0.52)</td>
<td>0.19–0.45 (0.31)</td>
<td>0.56</td>
</tr>
</tbody>
</table>

*All eight Functional Groups; †excludes Other Functions; ‡excludes Other Functions and Mental Functions.
the EDSS. Previous studies report high correlations with the FIM (Brosseau, 1994; Marolf et al., 1996; Sharrack et al., 1999) and BI (Sharrack et al., 1999), supporting our evidence for the convergent validity of the EDSS. Only Sharrack and colleagues examined convergent and discriminant validity together (Sharrack et al., 1999). The importance of this approach to validity examination is that a measurement instrument is far more valuable if it measures only the intended concept, i.e. if it is both specific and sensitive (McDowell and Jenkinson, 1996). In contrast to our findings, Sharrack and colleagues report that the EDSS correlates similarly with the BI (–0.74) and LHS (–0.69) (Sharrack et al., 1999). They also report a similar correlation between the EDSS and EuroQoL quality of life thermometer (–0.69).

The implication of these findings is that, in their less disabled sample, the EDSS has a limited ability to discriminate disability from handicap or quality of life and that measurement of these conceptually distinct health constructs is confounded.

There is consensus that instruments used to evaluate therapeutic interventions should be able to detect change (Fitzpatrick et al., 1998). The importance of responsiveness lies in the trade-off between sample size and statistical power: instruments with greater responsiveness have higher power for a fixed sample size, or require fewer patients to achieve a fixed level of statistical power (Liang et al., 1985). Despite this fact, the responsiveness of the EDSS, evaluated using recommended methods, is not well documented. Nevertheless, results of previous studies allude to its limited responsiveness. Marolf and colleagues studied multiple sclerosis patients before and after in-patient rehabilitation, and demonstrated that 95% of patients had unchanged EDSS scores, whereas 68% had unchanged FIM scores (Marolf et al., 1996). Kurtzke reports that 50% of patients ‘whose exacerbation leading to admission had persisted for 2 years or less’ (Kurtzke, 1955), and 61% of patients admitted for ‘an early bout of’ multiple sclerosis (Kurtzke, 1983) had unchanged DSS scores at discharge (length of stay and details of treatment not reported). Finally, in the study of isoniazid (Veterans Administration Multiple Sclerosis Study Group, 1957), 53% of patients categorized as having active disease had the same DSS scores 9 months later.

Only Sharrack and colleagues report the responsiveness of the EDSS as an effect size and compare this with other instruments (Sharrack et al., 1999). Although their results lead to a similar conclusion to this study, Sharrack determines responsiveness retrospectively by calculating an effect size only for those patients deemed to have changed. In the present study, responsiveness is determined prospectively by comparing the EDSS scores for the whole sample pre- and post-rehabilitation. Recently, Norman and colleagues have compared prospective and retrospective methods of determining responsiveness. They provide empirical evidence that there is no consistent relationship between the two methods and that responsiveness calculated retrospectively does not indicate the ability of an instrument to detect change (Norman et al., 1997).

The measurement properties of the FS have been the subject of a number of studies. As with the EDSS, variable reliability results are reported. Two studies (Kurtzke, 1970; Weihenken et al., 1996) document correlations for the FS. Kurtzke examined the relationships among six FS (Visual and Other Functions not reported), and between these FS and the DSS. Results are similar to ours, with one notable exception, a correlation between Pyramidal and Cerebellar Functions of +0.32 (–0.23 in the present study). Sample differences probably explain this discrepancy. Kurtzke studied less disabled patients (80% DSS < 6.0) in which a weak positive correlation is predicted; we have studied more disabled patients in whom increasing weakness obscures cerebellar signs (Kurtzke, 1961), and a weak negative correlation is predicted. Weihenken and colleagues (EDSS < 8.0) report three FS intercorrelations, all similar to ours: Brainstem with Cerebellar FS (0.33), Brainstem with Cerebral Functions (0.25) and Pyramidal Functions with the EDSS (0.54) (Weihenken et al., 1996). The results of these two studies, along with our findings, support Kurtzke’s hypotheses that the eight FS and DSS/EDSS measure different constructs, although these results alone do not tell us the nature of those constructs.

The extent of these similarities between these three studies is perhaps surprising for two reasons. First, it is widely believed that the EDSS is an impairment scale in its lower range and a disability scale in its upper range (Sharrack and Hughes, 1996). Secondly, low EDSS scores are fully defined by the FS grades, whereas high scores are more independent of them (Kurtzke 1983). Whilst there is evidence that the validity of an instrument is not constant throughout the range of scores (Lee and Foley, 1986), the extent to which these assumptions influence the measurement properties of the EDSS and its relationships with the FS are empirical questions that have yet to be answered.

This study has a number of limitations in relation to the generalizability of the results and the reliability of the methodology. The measurement properties of the EDSS and FS have been examined in a sample of moderately and severely disabled multiple sclerosis patients, largely in the progressive phase of the disease, individually selected for in-patient rehabilitation. As this sample is not representative of all multiple sclerosis patients (natural history studies suggest that only 55% of multiple sclerosis patients have EDSS scores >5.0), our results should be generalized with caution to dissimilar samples. This is particularly relevant as results from different studies suggest that some of the measurement properties of the EDSS are not stable across samples.

Suboptimal reproducibility methods are used in this study as we have studied only the agreement between one rater and a group of 10 others. Are the results biased by the fact that one rater was constant throughout the study? Subgroup analyses suggest that this is not the case, although it must be noted that the sample sizes involved are small. Whilst
J.H. generated EDSS ratings higher than the SHOs in seven of 10 comparisons, only two were statistically significant ($P < 0.05$). For the FS, J.H. generated higher ratings than the SHOs in 34 of 70 comparisons.

Reproducibility studies using a small number of raters also have limited generalizability. The optimum method would have examined the agreement of EDSS and FS scores within and between raters randomly selected from a large pool (Shrout and Fleiss, 1979). However, other EDSS and FS reproducibility studies have used fewer raters (e.g. Amato et al., 1987; Francis et al., 1991; Goodkin et al., 1992; Sharrack et al., 1999), which may explain why some inter-rater reproducibility results are so variable. Most importantly, we have examined the reliability of EDSS and FS in a manner representative of the clinical setting in which they are used.

Results from this and other studies have important implications for the use of the EDSS and FS. The finding that reliability is variable and has often failed to satisfy the criterion for group comparisons supports recommendations that raters of Kurtzke scales should be trained (Lechner-Scott et al., 1997), and that this improves reliability and is likely to improve responsiveness. It is, however, notable in the present study that the more experienced rater has higher reproducibility for the EDSS but lower reproducibility for most of the FS. Although the FS result is not expected, it may be explained by the limited reliability of single-item measures, an issue that is discussed later. This study has too few raters to determine the relationship between experience and reproducibility.

Perhaps more importantly, the finding that EDSS reliability rarely satisfies the criterion for individual comparisons calls into question the use of treatment failure as the primary outcome of trials for therapeutic interventions in multiple sclerosis. In this method, EDSS scores for individuals are compared in order to determine unequivocal deterioration in disability (Weinshenker et al., 1996). The demonstration in this study that a change of 0.5 EDSS points in the range 5.0–9.0 can be attributed to measurement error questions the recommendation that this degree of change is significant. Our results suggest that reliability must exceed 0.93 for this criterion to be adopted, which is consistent with the recommendations of others (Nunnally, 1978; Scientific Advisory Committee of the Medical Outcomes Trust, 1995).

The implications of the validity and responsiveness results are particularly important. Although we have demonstrated that the EDSS in the range 5.0–9.0 measures disability and is able to discriminate disability from related health constructs, we have also demonstrated that it has a limited ability to discriminate between individuals and groups and has poor responsiveness. As Kurtzke pointed out, when measures are used to evaluate therapeutic effectiveness, the ability to differentiate between people and groups, and to detect change over time is essential. When these properties are not demonstrated, the finding that a therapeutic intervention appears ineffective may represent limitations of the measure rather than of the treatment.

Criticisms of the EDSS have resulted in research directed towards the development of new measures (Whitaker et al., 1995), such as the Multiple Sclerosis Functional Composite (MSFC; Cutter et al., 1999). This is a three-item measure developed by analysing longitudinal data sets from the placebo arms of clinical trials and natural history studies that contained both clinical and functional measures. From these data, clinically relevant variables with good measurement properties were selected. The three items are stand-alone measures of different clinical dimensions: the nine hole peg test (arm dimension); the timed walk (leg dimension); and the 3 min version of the paced auditory serial addition test (PASAT3; cognitive dimension). Raw scores for each item are transformed into $Z$-scores (standard scores) so that they have a common metric, and then summed to generate a composite score. Selecting items from a pool on the basis of empirical performance is consistent with psychometric methods (Spector, 1992), and research is now required to examine whether the three items of the MSFC satisfy criteria as a summed rating scale and, if so, is it reliable, valid and responsive.

The MSFC represents a development in measurement methods that is well recognized by psychometricians—the superiority of multi-item over single-item measures. The EDSS and each FS, like the Rankin and Ashworth scales (Rankin, 1957; Ashworth, 1964), are single-item measures. An alternative method of scaling the attributes of people is to combine multiple items, each of which measures a different aspect of the underlying construct. This is the basis of Likert or summed rating scales (Likert, 1932), examples of which are the BI, FIM, SF-36 and EuroQoL (EuroQoL Group, 1990). Likert demonstrated that scores based on the simple assignment of integral weights to items’ responses (i.e. 0, 1, 2, 3, etc.) correlated 0.99 with Thurstone’s more complicated scoring system (Thurstone, 1928) in which the step intervals for scaling items were equalized. Empirical evidence has demonstrated repeatedly that Likert’s method of summing items rated on ordinal scales is a valid method of measurement. Furthermore, multi-item measures have been shown to be more reliable, valid and precise than single-item measures of psychological (Nunnally, 1978; McIver and Carmines, 1981; Spector, 1992) and health (McHorney et al., 1992) constructs. This study provides further support for the limitations of single-item health measures by demonstrating that the EDSS has measurement properties inferior to the BI and FIM.

The superiority of multi-item measures is easily explained. Conceptually, single items are unlikely fully to represent complex theoretical concepts (Nunnally, 1978) such as disability. This is exemplified by criticisms that the EDSS in the range studied here is too focused on ambulation and does not account for other aspects of disability (Sharrack and Hughes, 1996). Furthermore, single items are unable to make the fine differentiations among people that are desirable for
most measurement problems (Nunnally, 1978). Consider the EDSS and the FIM. The EDSS defines 20 levels of disability. In contrast, each of the 18 FIM items measures a different aspect of disability on a seven-point response scale (1 = maximum disability, 7 = minimum disability). Item scores are summed to generate a total score which defines 108 disability levels.

Another limitation of single-item measures is that they are unreliable (Nunnally, 1978). This is the most likely explanation for the variable and largely poor reliability results of the EDSS and FS that the different studies have reported. However, unreliability averages out when scores on numerous items are summed to obtain a total score (measurement error is random). Perhaps the most important limitation of single-item measures is that their measurement properties are difficult to examine (McIver and Carmines, 1981). This is exemplified by the FS: determining their validity is difficult. Indeed, we have not attempted this.

In conclusion, Kurtzke’s scales represent an important milestone in the history of the measurement of disease severity in multiple sclerosis. Widespread international use over 45 years confirms their clinical usefulness. However, limitations in their scientific rigour, predictable from psychometric theory, severely question their role as evaluative outcome measures in future therapeutic studies of multiple sclerosis. Recently, a number of multi-item measures for multiple sclerosis have been developed using psychometric techniques (Vickrey et al., 1995; Cella et al., 1996; LaRocca et al., 1996; Sharrack et al., 1996a; Ford et al., 1997). Studies are required to examine and compare the measurement properties of these instruments so that choice for future treatment trials can be empirically led.

Acknowledgements
We wish to thank the staff of the Neurorehabilitation Unit and the 10 rotating neurology SHOs for collecting outcomes data routinely, and to Professor Ian McDonald for his helpful comments on an earlier draft. Between September 1995 and November 1996, Dr Hobart was funded by a Wellcome Training Fellowship in Health Services Research.

References


McHorney CA, Ware JE Jr, Raczer 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. Med Care 1993; 31: 247–63.


Thurstone LL. Attitudes can be measured. Am J Sociol 1928; 33: 529–54.


