Brain (2001), **124,** 962–973

The Multiple Sclerosis Impact Scale (MSIS-29)
A new patient-based outcome measure

Jeremy Hobart,[1] Donna Lamping,[2] Ray Fitzpatrick,[3] Afsane Riazi[1] and Alan Thompson[1]

[1]*Neurological Outcome Measures Unit, Institute of Neurology, London,* [2]*Health Services Research Unit, London School of Hygiene and Tropical Medicine and* [3]*Department of Public Health and Primary Care, University of Oxford, 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

Changes in health policy have underlined the importance of evidence-based clinical practice and rigorous evaluation of patient-based outcomes. As patient-based outcome measurement is particularly important in treatment trials of multiple sclerosis, a number of disease-specific instruments have been developed recently. One limitation of these instruments is that none was developed using the standard psychometric approach of reducing a large item pool generated from people with multiple sclerosis. Consequently, an outcome measure for clinical trials of multiple sclerosis that is disease specific and combines patient perspective with rigorous psychometric methods will complement existing instruments. The aim of this study was to develop such a measure. Standard psychometric methods were used. A pool of 129 questionnaire items was generated from interviews with 30 people with multiple sclerosis, expert opinion and literature review. The questionnaire was administered by postal survey to 1530 people selected randomly from the Multiple Sclerosis Society membership database. Redundant items and those with limited measurement properties were removed. The remaining items (n = 41) were grouped into scales using factor analysis, and then refined to form the Multiple Sclerosis Impact Scale (MSIS-29), an instrument measuring the physical (20 items) and psychological (nine items) impact of multiple sclerosis. Five psychometric properties of the MSIS-29 (data quality, scaling assumptions, acceptability, reliability and validity) were examined in a separate postal survey of 1250 Multiple Sclerosis Society members. A preliminary responsiveness study of the MSIS-29 was undertaken in 55 people admitted for rehabilitation and intravenous steroid treatment of relapses. The MSIS-29 satisfied all psychometric criteria. Data quality was excellent, missing data were low (maximum 3.9%), item test–re-test reliability was high (r = 0.65–0.90) and scale scores could be generated for >98% of respondents. Item descriptive statistics, item convergent and discriminant validity, and factor analysis indicated that it was legitimate to generate scores for MSIS-29 scales by summing items. MSIS-29 scales showed good variability, small floor and ceiling effects, high internal consistency (Cronbach's alpha ≤0.91) and high test–re-test reliability (intraclass correlation ≤0.87). Correlations with other measures and the analysis of group differences provided evidence that the MSIS-29 measures the physical and psychological impact of multiple sclerosis. Effect sizes (physical scale = 0.82, psychological scale = 0.66) demonstrated preliminary evidence of good responsiveness. These results indicate the MSIS-29 is a clinically useful and scientifically sound patient-based outcome measure of the impact of multiple sclerosis suitable for clinical trials and epidemiological studies.

Keywords: multiple sclerosis; Multiple Sclerosis Impact Scale (MSIS-29); health outcomes measurement; clinical trials; psychometric methods

Abbreviations: BI = Barthel Index; EDSS = Expanded Disability Status Scale; EQ5D = EuroQol; ES = effect size; FAMS = Functional Assessment of Multiple Sclerosis; GHQ-12 = 12-item version of the General Health Questionnaire; MSIS-29 = Multiple Sclerosis Impact Scale; NHNN = National Hospital for Neurology and Neurosurgery; SF-36 = Medical Outcomes Study 36-item Short-Form Health Survey

Introduction

Changes in health policy have underlined the importance of evidence-based clinical practice and the need to evaluate outcomes that are important to patients. These changes are particularly relevant to multiple sclerosis, a chronic, disabling, condition of young people for which a number of costly interventions are available that purport to improve quality of life. As decisions about the effectiveness of these treatments influence patient welfare and the expenditure of public funds, it is essential that evaluations are based on scientifically rigorous outcome measures. If treatments are to be evaluated using outcomes that are important to patients, and are intended to incorporate their perspective, the instruments used should be developed from and completed by patients.

Over the last two decades, outcome measurement in multiple sclerosis has relied heavily on the Expanded Disability Status Scale (EDSS) (Kurtzke, 1983). Although the EDSS evaluates disability, it was developed before psychometric methods became widely available to clinicians, was not based on recognized techniques of scale construction (Nunnally and Bernstein, 1994) and did not directly involve people with multiple sclerosis. Moreover, the EDSS is rated by neurologists rather than by patients themselves and has limited measurement properties (Sharrack *et al.*, 1999; Hobart *et al.*, 2000*b*).

The lack of validated multiple sclerosis-specific measures has led to the use of generic measures, such as the Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) (Ware *et al.*, 1993), which have the advantage of enabling comparisons across diseases. However, generic measures may fail to address important areas of impact that are disease specific (Peto *et al.*, 1995) or may have limited responsiveness (Patrick and Deyo, 1989). Psychometric limitations of the SF-36 in multiple sclerosis include significant floor and ceiling effects (Freeman *et al.*, 2000), limited responsiveness (Freeman *et al.*, 2000), underestimation of mental health problems (Nortvedt *et al.*, 2000) and a failure to satisfy assumptions about scaling summary scores (Hobart *et al.*, 2000*a*).

A number of multiple sclerosis-specific measures have been developed in the last 5 years. These include the Functional Assessment of Multiple Sclerosis (FAMS) (Cella *et al.*, 1996), the MSQOL-54 (Vickrey *et al.*, 1995), the Multiple Sclerosis Functional Composite (MSFC) (Rudick *et al.*, 1997), the Guy's (now UK) Neurological Disability Scale (GNDS/UKNDS) (Sharrack and Hughes, 1999), the Multiple Sclerosis Quality of Life Inventory (MSQLI) (LaRocca *et al.*, 1996) and the Health-Related Quality of Life Questionnaire for Multiple Sclerosis (HRQOL-MS) (Pfennings *et al.*, 1999). One of the limitations of these disease-specific measures is that none was developed using the standard psychometric approach of reducing an item pool generated *de novo* from people with multiple sclerosis. The FAMS and MSQOL-54 were developed by adding multiple sclerosis-specific items to existing measures, an approach that has some limitations (Freeman *et al.*, 1999). The HRQOL-MS was developed from the statistical analysis of items from two generic and one multiple sclerosis-specific measure, whilst the MSQLI combines a large number of existing disease-specific and generic instruments. Items for the GNDS were developed through expert clinical opinion rather than on the basis of interviews with people with multiple sclerosis. Consequently, an outcome measure for clinical trials that is multiple sclerosis specific and combines patient perspective with rigorous psychometric methods will complement existing instruments. The aim of this study was to develop such a measure.

Method

Overview

The MSIS-29 was developed in three stages. First, a 129-item questionnaire was generated from 30 patient interviews, expert opinion and literature review. The questionnaire was then administered by postal survey to 1530 randomly selected members of the Multiple Sclerosis Society of Great Britain and Northern Ireland; standard item reduction techniques were used to develop a 29-item scale measuring the physical (20 items) and psychological (nine items) impact of multiple sclerosis (see Appendix I). Finally, the psychometric properties of the MSIS-29 (i.e. data quality, scaling assumptions, acceptability, reliability and validity) were evaluated in an independent sample of 1250 members of the Multiple Sclerosis Society. A preliminary study of the responsiveness of the MSIS-29 has been conducted in 55 in-patients at the National Hospital for Neurology and Neurosurgery (NHNN). The ethics committee of the NHNN approved the study.

Item generation

An initial pool of 129 items concerning the health impact of multiple sclerosis was generated from three sources: semi-structured interviews of people with multiple sclerosis, multidisciplinary expert opinion and a comprehensive literature review. All 30 people with multiple sclerosis who were invited for interview, selected to represent the diversity of the illness, agreed to participate. Interviews lasted an average of 1 h, were tape recorded, transcribed and then content analysed. Statements concerning the health impact of multiple sclerosis were extracted, grouped into themes and examined for redundancy by the study team. A total of 3750 health impact statements were extracted from the interviews (mean 125; range 64–212). These statements generated 91 questionnaire items. Although no new themes appeared after the first 20 interviews, all 30 were analysed. A further 38 items were generated from interviews with health professionals at the NHNN (i.e. neurologists, neuropsychologists, nurses, occupational therapists, physiotherapists, social workers, and speech and language therapists) who were involved in the care of people with multiple sclerosis and from a comprehensive literature review.

Examination of the content of the 129 items indicated that two distinct question stems and response scales were required. The majority of items ($n = 97$) were best represented by the stem 'How much have you been bothered by . . .' with a five-point response option (1 = not at all; 5 = extremely). The remaining items ($n = 32$) that referred specifically to activity limitations were best represented by the stem 'How much has your multiple sclerosis limited your ability to . . .' with a six-point response option (1 = not at all limited; 6 = unable to do this activity). The time frame for all questions was the preceding 2 weeks.

The preliminary 129-item questionnaire was reviewed for content, wording and clinical appropriateness by patients and clinicians who were involved in its development. It was then pre-tested formally in an independent and heterogeneous sample of 20 people with multiple sclerosis who were attending the NHNN. They identified items and instructions that were unclear, ambiguous, irrelevant, misleading or offensive, and made suggestions for alterations to the questionnaire.

Item reduction and development of scales (first field test)

The 129-item questionnaire was administered by postal survey to 1530 people, randomly selected and geographically stratified, from the membership database of the Multiple Sclerosis Society of Great Britain and Northern Ireland. This sampling frame has the advantage of being truly representative. The disadvantage is that not all members have multiple sclerosis. Therefore, based on results of a pilot study (Hobart *et al.*, 2000*c*), we chose a target sample size of 1530 to ensure 500 completed questionnaires with no missing data. A subsample of 400 people was randomly selected from the larger sample to study item test–retest reproducibility to ensure 125 completed questionnaires on two occasions with no missing data. Patients in the test–retest sample received two questionnaires in the same envelope: one to complete immediately (time 1) and a second in a sealed envelope with instructions to open and complete 10 days later (time 2). A postcard reminder to complete the time 2 questionnaire was sent at day 7. Non-responders received reminders (letter and questionnaire) at 2 and 4 weeks (Dillman, 1978). In the test–retest subsample, non-responders to the time 2 questionnaire did not receive a reminder.

Item reduction

The following psychometric properties (descriptive statistics and reliability estimates) were examined for each item: percentage missing data; frequency distributions for each response option; maximum endorsement frequency (response option most frequently endorsed); mean score, standard deviation, skewness, floor and ceiling effects; and test–retest reproducibility (product–moment correlations). Items with >10% missing data were eliminated (WHOQOL Group, 1998). Correlations among the items were then examined to identify redundant items (item–item correlations ⩾0.70; Juniper *et al.*, 1997). For each item–item correlation ⩾0.70, the item with the least favourable psychometric properties was eliminated. When items had similar psychometric properties, a consensus clinical decision determined which item to retain. Finally, the psychometric properties of the remaining items were examined. Items were eliminated if: floor effects, ceiling effects or maximum endorsement frequencies exceeded 40%; the sum of the endorsement frequencies for any two adjacent item response categories was <10% (WHOQOL Group, 1998); or if item test–retest reproducibility was <0.50 (Duruoz *et al.*, 1996).

Development of scales

Scales were developed using an iterative process. First, all items were entered into a principal components analysis without rotation to determine whether there were any rogue items that should be eliminated (Ferguson and Cox, 1993). Next, principal axis factoring with varimax rotation was undertaken (Fayers and Machin, 1998). Multiple criteria were used to determine how many factors to rotate: Eigenvalues exceeding unity (Guttman, 1954); the scree test (Cattell, 1966); the 5% rule (Guertin and Bailey, 1970); and trial rotations (Ware *et al.*, 1980). All potential factor solutions were examined for cross loading [items loading on two or more factors by >0.40, and items loading on two or more factors within 0.1 of each other (Ferguson and Cox, 1993)], clinical interpretability of item content and replicability of results in random split half samples. Item groups modelled through factor analysis were then examined to determine if they satisfied recommended criteria for summed rating scales and were acceptable, reliable and valid (methods described below).

Psychometric evaluation of the MSIS-29 (second field test)

Item reduction analyses produced a 29-item measure that includes two scales: physical impact (20 items) and psychological impact (nine items). All items could be referenced back to statements made by patients during the interviews. The two summary scores are generated by summing individual items and then transformed to a 0–100 scale. High scores indicate worse health. For respondents with missing data, but where at least 50% of the items in a scale had been completed, a respondent-specific mean score computed from the completed items was imputed (Ware *et al.*, 1993).

The psychometric properties of the MSIS-29 were evaluated comprehensively in two independent samples. A second and separate postal survey of randomly selected and geographically stratified members of the Multiple Sclerosis Society ($n = 1250$) was undertaken to evaluate data quality, scaling assumptions, acceptability, reliability and validity. Responsiveness was evaluated in 55 people with multiple sclerosis admitted to the NHNN for in-patient rehabilitation or intravenous steroids for multiple sclerosis relapses.

Postal survey

The postal survey sample was divided randomly into three subsamples (n = 500, 500 and 250). Respondents in the two larger subsamples completed the MSIS-29, demographic questions and three other health measures. Respondents in sample 1 completed the SF-36, EuroQol (EQ-5D) (EuroQol Group, 1990) and postal Barthel Index (BI) (Gompertz *et al.*, 1994), whilst respondents in sample 2 completed the FAMS, EQ-5D and 12-item version of the General Health Questionnaire (GHQ-12) (Goldberg and Hillier, 1979). Respondents in the smaller test–retest subsample completed the MSIS-29 on two occasions separated by a 10-day interval. The survey methods were the same as those used in the first field test.

Five psychometric properties of the MSIS-29 were evaluated using standard methods (Nunnally and Bernstein, 1994; Streiner and Norman, 1995; Lohr *et al.*, 1996). Data quality (McHorney *et al.*, 1994) was determined by calculating the percentage missing data for items, percentage computable scale scores and item test–retest reproducibility (intraclass correlation coefficient, ICC; Bartko, 1966). Scaling assumptions examine whether it is legitimate to generate scores by summing items without weighting or standardization, and whether items are grouped correctly into scales. Items can be summed to generate scores when items have similar response option frequency distributions, equivalent mean scores and variances, and substantial ($r > 0.30$) and equivalent item–total correlations (Likert, 1932). Items are grouped correctly into scales when item–own scale correlations exceed item–other scale correlations by at least two standard errors ($1/\sqrt{n}$; Ware *et al.*, 1997), and when the results of factor analysis support hypothesized item groups.

Acceptability was determined by examining score distributions. Acceptability is supported when observed scores are well distributed (Stewart and Ware, 1992), mean scores are near the scale mid-point (Eisen *et al.*, 1979), floor and ceiling effects are <20% (McHorney and Tarlov, 1995) and skewness statistics range from −1 to +1 (Holmes *et al.*, 1996). Two types of reliability, internal consistency (Cronbach's alpha coefficients; Cronbach, 1951) and scale test–retest reproducibility (ICC), were examined. Estimates should exceed 0.80 (Nunnally and Bernstein, 1994).

The aim of the validity studies was to examine evidence that the MSIS-29 was a measure of the physical and psychological impact of multiple sclerosis. Three types of validity were examined. Internal validity (Bohrnstedt, 1983) was determined by examining the intercorrelation between MSIS-29 scales. A moderate correlation (r = 0.30–0.70) was predicted. Convergent and discriminant validity (Cronbach and Meehl, 1955) was determined by examining the extent to which correlations between MSIS-29 scales and other measures (SF-36, BI, EQ-5D, FAMS and GHQ) and variables (age, sex and duration of multiple sclerosis) were consistent with predictions. For example, we predicted that the MSIS-29 physical impact scale would correlate highly ($r > 0.70$) with other measures of physical health (e.g. SF-36 physical functioning dimension, BI, FAMS mobility scale and EQ-5D mobility dimension). Group differences validity was determined by examining MSIS-29 scores for groups of patients. We predicted that: people who were retired due to their multiple sclerosis would have higher scores than people who were still employed; people with increasing difficulties in mobility and self-care as defined by the EQ-5D would have greater differences in their physical scores than their psychological scores; people with increasing anxiety or depression as defined by the EQ-5D would have greater differences in their psychological scores than their physical scores; men and women would have similar scores; and people with or without a degree would have similar scores.

Responsiveness study

A preliminary responsiveness study has been undertaken in consecutive admissions to the NHNN between February 1 and August 1, 2000 for rehabilitation and intravenous steroid treatment. People were excluded if they appeared to have severe cognitive impairment substantiated by neuropsychological testing. People admitted for rehabilitation completed the MSIS-29 on admission and discharge, whilst those admitted for intravenous steroid treatment completed the MSIS-29 on admission and 6 weeks later. Responsiveness was determined by calculating effect sizes (ES; Kazis *et al.*, 1989), mean change score (admission minus discharge) divided by the standard deviation of admission scores. These are interpreted (Cohen, 1969) as either small (ES < 0.20), medium (ES = 0.50) or large (ES > 0.80). The statistical significance of the change scores was determined using paired sample t tests (Deyo *et al.*, 1991).

Results

Item generation, item reduction and development of scales

The characteristics of the 30 people with multiple sclerosis interviewed covered the diversity of the illness (Table 1). From the first field test ($n = 1530$), a total of 1202 (78.6%) questionnaires were returned of which 436 were returned blank (change of address or deceased $n = 113$, did not have multiple sclerosis $n = 207$, did not wish to participate $n = 97$, no reason given $n = 19$). The response rate was 63.3%

Table 1 *Characteristics of samples*

Variable*	Sample Semi-structured interviews	First field test	Second field test	Responsiveness
n[†]	30	766	713	55[‡]
Gender				
Female	56	74	71	66
Age				
Mean (SD)	41 (12)	51 (12)	52 (12)	45 (13)
Range	23–70	23–87	18–82	23–83
Ethnicity				
White	100	98	98	95
Years since MS onset				
Mean (SD)	12 (11)	19 (12)	19 (11)	16 (12)
Range	1–36	1–56	1–59	1–60
Mobility indoors				
Walks unaided	40	–[§]	32	24
Walks with an aid	23	–	40	49
Uses a wheelchair	37	–	28	27
Mobility				
Can walk	N/A	79	–	
Cannot walk	N/A	21	–	
Marital status				
Married	77	66	70	64
Living with others		83	81	82
Employment status				
Retired due to MS	63	54	56	31
Employed		18	19	44
Type of MS (%)				
Primary progressive	13.3	Unknown	Unknown	5.5
Secondary progressive	43.4	Unknown	Unknown	47.3
Relapsing–remitting	43.3	Unknown	Unknown	47.3

*All values are percentages unless specified otherwise; [†]for whom both physical and psychological scale scores could be computed; [‡]$n = 27$ admitted for in-patient rehabilitation and $n = 28$ admitted for intravenous steroids; [§]question not asked.

[response rate: 1202 – 436/(1530 – 113 – 207) = 63.3%]. Therefore, item analyses were performed on data for 766 people with multiple sclerosis (Table 1). None of the items failed the criteria for missing data or test–retest reproducibility. Forty-seven items were eliminated on the basis of item redundancy, and 41 items failed the other criteria (floor and ceiling effects, etc.)

The remaining 41 items were entered into a principal components analysis. All items loaded onto the first component by >0.40, indicating a common underlying dimension. Neither principal components analysis nor principal axis factoring indicated a clear solution. Therefore, all solutions with two to seven factors were evaluated. The two-factor solution was judged to be the most appropriate. However, three items that loaded on both factors by >0.40, indicating a limited ability to discriminate between the two factors, were removed. Whilst the clinical interpretation of the two factors led these to be labelled the physical (25 items) and psychological (13 items) impact of multiple sclerosis, the consensus opinion of the investigators was that five items were not entirely consistent with this interpretation. These items were thus removed, resulting in a 33-item instrument with two scales: physical impact (22 items) and psychological impact (11 items).

Preliminary psychometric evaluation of the MSIS-33 indicated that three items had similar correlations with the two scales and, therefore, were considered probable scaling failures. These items were removed to minimize measurement overlap between the two scales (Ware *et al.*, 1997). When the psychometric properties of the 30-item measure were re-tested, all criteria were satisfied except one item that was classified as a probable scaling failure. This item was therefore removed to produce the final 29-item MSIS (20-item physical scale; nine-item psychological scale). The MSIS-29 includes 26 items with five-point response options and three items with six-point response options. The latter three items were re-scaled (category 5 combined with 6) so that all items have the same number of response options. Preliminary psychometric analyses, based on data collected in the first field test, indicated that the MSIS-29 satisfied standard criteria for acceptability, reliability and validity (results not reported).

Psychometric evaluation of the MSIS-29

Postal survey

A total of 1023 (81.8%) questionnaires were returned, of which 310 were returned blank (change of address or deceased

Table 2 *Data quality, scaling assumptions, acceptability, reliability and responsiveness of the MSIS-29*

Psychometric property	MSIS-29 scale	
	Physical impact	Psychological impact
Data quality (n = 713)		
Item missing data %	1.7–3.6	1.1–1.8
Item test–retest reproducibility*: range (mean)	0.65–0.90 (0.81)	0.72–0.82 (0.78)
Computable scale scores %	98.0	98.7
Scaling assumptions (n = 703)		
Item mean scores: range	2.54–3.83	2.57–3.28
Item SD: range	1.20–1.56	1.27–1.37
Item skewness: range	–0.86 to +0.41	–0.29 to +0.40
Definite scaling successes[†]	100%	100%
Item–own factor loading: range	0.58–0.85	0.47–0.79
Item–other factor loading: range	0.19–0.38	0.19–0.36
Acceptability (n = 703)		
Possible score range	0–100	0–100
Observed score range	0–100	0–100
Mean score (SD)	56.0 (26.6)	45.5 (25.2)
Floor/ceiling effect %	0.9/3.9	1.7/1.9
Skewness	–0.285	+0.172
Reliability		
Cronbach's alpha (n = 703)	0.96	0.91
Scale test–re-test reproducibility (n = 128)*	0.94	0.87
Responsiveness (n = 55)		
Time 1 score: mean (SD)	64.4 (23.0)	48.4 (26.7)
Time 2 score: mean (SD)	45.6 (23.4)	30.7 (22.3)
Change score[‡]: mean (SD); P	18.8 (19.6); <0.001	17.7 (24.6); <0.001
Effect size[§]	0.82	0.66

*Intraclass correlation coefficient; [†]percentage of times where item–own scale correlation exceeds item–other scale correlation by at least 2 SE ($2 \times 1/\sqrt{n}$); [‡]time 1 minus time 2; [§]mean change score divided by standard deviation of time 1 score.

n = 63, did not have multiple sclerosis n = 155, did not wish to participate n = 64, no reason given n = 28). The second postal survey generated data for 713 people giving a response rate of 69.1% (1023 – 310/1250 – 63 – 155) that was similar to the first field test. In the test–retest subsample, 90.6% (n = 136) of people who returned the time 1 questionnaire returned the time 2 questionnaire. The characteristics of samples for the first and second field tests were similar (Table 1). There were no significant differences in demographic characteristics between patients in the three subsamples.

Data quality (Table 2). Missing data for items were low (range 1.1–3.6%). Eighty-four per cent of respondents endorsed all 29 items (100% complete data), 8.4% of respondents missed out one item and 3.2% of respondents missed out two items. Ninety-seven per cent of respondents had ⩾90% complete data. Therefore, MSIS-29 scale scores could be computed for 703 respondents (98.6%). Item test–retest reproducibility was high. These results indicate that data quality was high.

Scaling assumptions (Table 2). Frequency distributions for item response scales were quite symmetrical and not unduly skewed (range –0.86 to +0.41), and items within each scale had similar mean scores and standard deviations. All item–own scale correlations were high (range 0.49–0.84) and exceeded item–other scale correlations by at least two standard errors (range 0.12–0.39). Principal axis factoring of the 29 items, cross-validated in random split half samples, generated two factors whose item contents were consistent with the hypothesized physical and psychological scales. These results indicate that the MSIS-29 satisfied tests of scaling assumptions.

Acceptability and reliability (Table 2). Scale scores spanned the entire scale range and were not notably skewed, mean scores were near the scale mid-point, and floor and ceiling effects were negligible (maximum 3.9%). Internal consistency and test–retest reproducibility exceeded the recommended criterion for group comparisons of 0.80. There were no statistically significant differences in MSIS-29 scores between the three subsamples and between time 1 and time 2 scores for the test–retest reproducibility subsample. These results indicate that the MSIS-29 satisfied criteria for acceptability and reliability.

Validity. Total scores for physical and psychological scales of the MSIS-29 were correlated 0.62, indicating that the two scales measure related but distinct constructs. Table 3 provides evidence for the convergent and discriminant validity of

Table 3 *Convergent and discriminant construct validity of the MSIS-29*

Instrument	Scale/dimension/variable	MSIS-29 scale* $r^{\dagger}$ (n)	
		Physical	Psychological
SF-36[‡]	Physical functioning	–0.79	–0.41
	Role limitations physical	–0.43	–0.40
	Bodily pain	–0.45	–0.50
	General health perception	–0.48	–0.53
	Vitality	–0.49	–0.55
	Social functioning	–0.64	–0.56
	Role limitations emotional	–0.29	–0.52
	Mental health	–0.41	–0.76
FAMS[§]	Mobility	–0.88	–0.50
	Symptoms	–0.55	–0.64
	Emotional well-being	–0.68	–0.68
	General contentment	–0.64	–0.58
	Thinking and fatigue	–0.56	–0.73
	Family/social well-being	–0.37	–0.50
EQ-5D[¶]	Mobility	0.61	0.23
	Self-care	0.69	0.37
	Usual activities	0.69	0.42
	Pain/discomfort	0.44	0.43
	Anxiety/depression	0.36	0.68
GHQ-12**	Total score	0.46	0.68
Postal Barthel Index[††]	Total score	–0.71	–0.35
Demographic variables	Age (n = 678)	0.22	0.03
	Sex (n = 686)	0.05	–0.05
	Years since diagnosis (n = 629)	0.19	0.03

*Multiple Sclerosis Impact Scale: high scores = worst health; [†]Pearson product–moment correlation coefficients; [‡]Medical Outcomes Study 36-item Short Form Health Survey (n = 263–280): high scores = best health; [§]Functional Assessment of Multiple Sclerosis (n = 233–259): high scores = best health; [¶]EuroQol: high scores = worst health (n = 520–550); **General Health Questionnaire (n = 248 and 249, respectively): high scores = worst health; [††]high scores best health (n = 260 and 243, respectively).

MSIS-29 scales as measures of the physical and psychological impact of multiple sclerosis. The direction, magnitude and pattern of correlations are consistent with predictions. For example, the MSIS-29 physical scale correlates most with the FAMS mobility scale, the SF-36 physical functioning scale and the BI, and least with the EQ-5D anxiety/depression dimension, SF-36 emotional role limitations scale and the FAMS family/social well-being scale. Similarly, the MSIS-29 psychological scale correlates most with the SF-36 mental health scale, the FAMS thinking/fatigue scale and the GHQ-12, and least with EQ-5D mobility and self-care dimensions and the BI. In addition, both MSIS-29 scales have low correlations with age, sex and duration of multiple sclerosis, indicating that they are not biased by these variables. Some correlations, however, are not consistent with predictions. Notably, the MSIS-29 physical scale correlates more highly than expected with the FAMS emotional well-being scale.

The MSIS-29 confirms hypothesized group differences (Table 4). As predicted, mean scores for people who were retired due to multiple sclerosis were significantly higher than for those who were still employed. In contrast, mean scores for men and women, and those with or without a degree or professional qualification were not significantly different. Also as predicted, mean MSIS-29 scores for people with increasing problems in mobility, self-care and anxiety/depression, as defined by the EQ-5D, demonstrate a step-wise increase in magnitude and statistically significant F statistics (ratio of between-groups to within-groups variance). Furthermore, the relative validity calculations (pairwise F statistics) indicate that the MSIS-29 physical scale is more valid for detecting group differences in mobility and self-care, whilst the MSIS-29 psychological scale is more valid for detecting group differences in anxiety/depression.

Responsiveness study

Four people recruited to the responsiveness sample were excluded because of cognitive impairment. Although the responsiveness sample is small (n = 55), its characteristics are similar to those of the larger field test (Table 1). Scores for both the MSIS-29 scales were lower at time 2 than time 1 (Table 2), indicating improvement associated with in-patient rehabilitation and following i.v. steroid treatment. Change scores for both scales were similar in magnitude and statistically significant. Effect sizes were large to moderate.

Table 4 *MSIS-29 group differences and relative validity*

Variable	MSIS-29 score: mean (SD)	
	Physical	Psychological
Employment status		
Employed (*n* = 107)	30.6 (23.1)	31.1 (22.5)
Retired due to MS (*n* = 390)	64.3 (23.0)	49.9 (24.9)
Mean difference (*P*)	–33.7 (<0.001)	–18.8 (<0.001)
EQ-5D mobility dimension		
No problems in walking about (*n* = 61)	17.5 (17.2)	27.7 (23.1)
Some problems in walking about (*n* = 389)	56.4 (21.7)	46.6 (23.5)
Confined to bed (*n* = 70)	82.6 (16.8)	51.4 (28.1)
F (*P*)*	164.3 (<0.001)	19.3 (<0.001)
Relative validity[†]	1.0	0.12
EQ-5D self-care dimension		
No problems with self-care (*n* = 227)	35.5 (22.1)	34.5 (23.0)
Some problems with self-care (*n* = 235)	66.6 (16.9)	51.5 (22.0)
Unable to wash or dress myself (*n* = 76)	85.2 (15.3)	58.9 (27.3)
F (*P*)	256.7 (<0.001)	46.2 (<0.001)
Relative validity	1.0	0.18
EQ-5D anxiety/depression dimension		
Not anxious or depressed (*n* = 229)	45.8 (27.4)	27.1 (17.8)
Moderately anxious or depressed (*n* = 277)	62.2 (22.7)	55.6 (19.3)
Extremely anxious or depressed (*n* = 38)	75.3 (22.7)	81.6 (16.3)
F (*P*)	39.5 (<0.001)	231.3 (<0.001)
Relative validity	0.17	1.0
Gender		
Female (*n* = 489)	55.0 (26.9)	46.1 (25.8)
Male (*n* = 197)	58.1 (26.1)	43.5 (23.6)
Mean difference (*P*)[‡]	–3.1 (0.165)	2.6 (0.197)
Degree or professional qualification		
Yes (*n* = 183)	53.2 (26.7)	41.6 (25.8)
No (*n* = 491)	56.7 (26.5)	46.8 (24.8)
Mean difference (*P*)	–3.5 (0.131)	–5.3 (0.133)

*One-way ANOVA with Duncan's *post hoc* comparisons; [†]calculated as the ratio of paired *F* values using the largest as the denominator; [‡]independent samples *t* tests, equality of variances not assumed.

Discussion

The aim of this study was to develop a multiple sclerosis-specific outcome measure that combines the patient perspective with a rigorous scientific approach. We tried to achieve these aims by generating items from in-depth patient interviews, using the self-report method of administration, selecting items on the basis of psychometric performance in a large field test and rigorously applying psychometric methods. In the samples we have studied, the MSIS-29 satisfies criteria as a summed rating scale and is acceptable, reliable and valid. Furthermore, there is preliminary evidence that the MSIS-29 detects change. Finally, all items could be referenced back to statements made by patients during the interviews.

The MSIS-29 is a measure of the physical and psychological impact of multiple sclerosis from the patients' perspective. This description has been chosen as it best defines the health constructs that we intended to measure, and because the terms health-related quality of life and disablement, both of which could be used to categorize the MSIS-29, have several different definitions (Fitzpatrick *et al.*, 1998). We feel it is important to be as specific as possible when defining the purpose of a measure, to guide prospective users of any scale.

Stringent criteria for item selection were adopted in an attempt to develop an instrument with strong psychometric properties. In order to create a responsive scale, items were selected that discriminated well between individuals, while items with maximum endorsement frequencies >40% were eliminated. Similarly, in order to reduce overlap between the two MSIS-29 scales, we eliminated items with limited item convergent and discriminant validity. Such a rigorous approach to health measurement is important because the results of studies are dependent on the quality of the measures used for data collection, and the limitations of measures cannot be overcome easily by improvements in study design and powerful statistical methods (Fleiss, 1986). Results concerning the responsiveness of the MSIS-29 must be considered preliminary due to the small sample size, and further evaluations of responsiveness are needed in different samples and settings.

There are potential limitations in using the Multiple Sclerosis Society membership database to define our sampling frame. It is known that not all members have multiple sclerosis (calculations based on our postal surveys estimate this to be a minimum of 56%), and that many members of the Multiple Sclerosis Society are partners, friends or relatives of people with multiple sclerosis. Therefore, we specifically asked people who did not have multiple sclerosis to tick a box on the front of the questionnaire and return it blank. However, the percentage of people in the database with a neurologist-confirmed diagnosis of clinically definite multiple sclerosis, the disease type of those with multiple sclerosis and the representativeness of people who join charitable groups is unknown. Our estimates indicate, however, that we have randomly sampled from ~35% (28 000) of the total UK multiple sclerosis population.

As the psychometric properties of health measurement instruments are sample dependent and cannot be established in a single study (Stewart *et al.*, 1988), further evaluations of the MSIS-29 are needed. Critical evaluations in different settings will define the strengths and weaknesses of the MSIS-29, further define its role in clinical practice and research, and help to determine whether the development process may have been biased by people without multiple sclerosis completing the questionnaire. As traditional psychometric methods were used to develop and evaluate the MSIS-29, it is also important that newer psychometric methods such as Rasch (Rasch, 1960) and Item Response Theory (Lord and Novick, 1968) models are used to evaluate the MSIS-29. Finally, comparisons between the MSIS-29 and widely used measures for multiple sclerosis, such as the Multiple Sclerosis Functional Composite (Cutter *et al.*, 1999) and GNDS (Sharrack and Hughes, 1999), should be undertaken. These studies will determine the advantages and disadvantages of different instruments, how they complement each other, and provide an evidence-based framework to guide the selection of outcome measures for research and audit. Over time, the accumulation of such data will also establish normative values and content-based interpretation of scores and score changes (McDowell and Jenkinson, 1996).

This study produced some unexpected results. Only two distinct dimensions of health, physical and psychological impact, appear to underlie the diverse 129-item pool. Although many other dimensions of health such as symptoms were included in the initial version of the questionnaire, psychometric analyses did not support multiple dimensions. These results support previous findings (Ware *et al.*, 1994; Pfennings *et al.*, 1999) that have observed that a two-dimensional model, consisting of physical and psychological health, appears to underpin the construct of subjective health status. Another unexpected result is that the MSIS-29 physical and psychological scales have the same correlation with the FAMS emotional well-being scale. However, it is encouraging to note that correlations between the MSIS-29 physical scale and other measures of psychological distress (GHQ, SF-36 mental health dimension and EQ-5D anxiety/depression dimension) are low to moderate.

This study has important implications for clinical trials and epidemiological studies. The MSIS-29 can be used in cross-sectional studies to describe the impact of multiple sclerosis, in longitudinal studies to monitor the natural history of the disorder and, most importantly, in clinical trials to evaluate therapeutic effectiveness from the patients' perspective. Furthermore, the availability of reliable, valid and responsive patient-based outcome measures is central to an improved understanding of the impact of multiple sclerosis and its relationships with other indicators of disease activity, such as neuroimaging and neurophysiology.

In addition to physical and psychological impact scores, an overall impact score could be reported as the total scale satisfies criteria as a summed rating scale. Although a single summary score would simplify data analysis, we do not recommend use of an overall summary score for clinical trials or epidemiological studies. This is because evidence indicates that the two scales are measuring related but distinct constructs (intercorrelation between scales = 0.62; factor analysis supports two dimensions). Combining these distinct aspects of outcome into an overall score could mask important (and possibly opposite) differential effects of treatment on physical and psychological health.

Acknowledgements

We wish to thank the people with multiple sclerosis who freely participated in this study, the Multiple Sclerosis Society of Great Britain and Northern Ireland for their collaboration, Mr Peter Cardy and Dr Iain Smith (Advisory Committee), Ms Irene Richardson for undertaking the patient interviews and research assistance during the first field test, Dr Sarah Smith for assistance with the psychometric analyses during the item reduction process, Ms Sara Schroter and Mr Stefan Cano for their contribution to the item reduction process, and Ms Laura Camfield for research assistance during the second field test. The study was funded by a grant from the NHS Health Technology Assessment Programme, but the views and opinions expressed do not necessarily reflect those of the NHS Executive.

References

Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966; 19: 3–11.

Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, editors. Handbook of survey research. New York: Academic Press; 1983. p. 69–121.

Cattell RB. The scree test for the number of factors. Multivar Behav Res 1966; 1: 245–76.

Cella DF, Dineen K, Arnason B, Reder A, Webster KA, Karabatsos G, et al. Validation of the functional assessment of multiple sclerosis quality of life instrument. Neurology 1996; 47: 129–39.

Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969.

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297–334.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955; 52: 281–302.

Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122: 871–82.

Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991; 12: (4 Suppl) 142s–58s.

Dillman DA. Mail and telephone surveys: the total design method. New York: Wiley; 1978.

Duruoz MT, Poiraudeau S, Fermanian J, Menkes C-J, Amor B, Dougados M, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 1996; 23: 1167–72.

Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. Med Care 1979; 17: 902–21.

EuroQoL Group. EuroQoL: a new facility for the measurement of health-related quality of life. Health Policy 1990; 16: 199–208.

Fayers PM, Machin D. Factor analysis. In: Staquet MJ, Hays RD, Fayers PM, editors. Quality of life assessment in clinical trials: methods and practice. Oxford: Oxford University Press; 1998. p. 191–223.

Ferguson E, Cox T. Exploratory factor analysis: a user's guide. Int J Select Assess 1993; 1: 84–94.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. [Review]. Health Technol Assess 1998; 2: i–iv, 1–74.

Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986.

Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Improving measurement scales: is adding items the answer? [abstract]. Ann Neurol 1999; 46: 507.

Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection. The 36-item Short Form Health Survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68: 150–6.

Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. Psychol Med 1979; 9: 139–45.

Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. Clin Rehabil 1994; 8: 233–9.

Guertin WH, Bailey JP Jr. Introduction to modern factor analysis. Ann Arbor (MI): Edwards Brothers; 1970.

Guttman LA. Some necessary conditions for common-factor analysis. Psychometrika 1954; 19: 149–61.

Hobart J, Freeman J, Lamping D, Fitzpatrick R, Thompson AJ. The medical outcomes study 36-item Short-Form Health Survey in multiple sclerosis: why assumptions must be tested [abstract]. Ann Neurol 2000a; 48: 495.

Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000b; 123: 1027–40.

Hobart JC, Lamping DL, Fitzpatrick R, Thompson AJ. Patient based outcome measures for multiple sclerosis are needed and can be developed [abstract]. J Neurol Neurosurg Psychiatry 2000c; 69: 420.

Holmes W, Bix B, Shea J. SF-20 score and item distributions in a human immunodeficiency virus-seropositive sample. Med Care 1996; 34: 562–9.

Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. J Clin Epidemiol 1997; 50: 233–8.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989; 27 (3 Suppl): S178–89.

Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33: 1444–52.

LaRocca NG, Ritvo PG, Miller DM, Fischer JS, Andrews H, Paty DW. 'Quality of life' assessment in multiple sclerosis clinical trials: current status and strategies for improving multiple sclerosis clinical trial design. In: Goodkin DE, Rudick RA, editors. Multiple sclerosis: advances in clinical trial design, treatment and future perspectives. London: Springer-Verlag; 1996. p. 145–60.

Likert RA. A technique for the development of attitudes. Arch Psychol 1932; 140: 5–55.

Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996; 18: 979–92.

Lord FM, Novick MR. Statistical theories of mental test scores. Reading (MA): Addison-Wesley; 1968.

McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy 1996; 1: 238–46.

McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995; 4: 293–307.

McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994; 32: 40–66.

Nortvedt MW, Riise T, Myer K-M, Nyland HI. Performance of the SF-36, SF-12, and RAND-36 summary scales in a multiple sclerosis population. Med Care 2000; 38: 1022–8.

Nunnally JC, Bernstein IH. Psychometric theory. 3rd edn. New York: McGraw-Hill; 1994.

Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. [Review]. Med Care 1989; 27 (3 Suppl): S217–32.

Peto V, Jenkinson C, Fitzpatrick R, Greenhall R. The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease. Qual Life Res 1995; 4: 241–8.

Pfennings LE, Van der Ploeg HM, Cohen L, Bramsen I, Polman CH, Lankhorst GJ, et al. A health-related quality of life questionnaire for multiple sclerosis patients. Acta Neurol Scand 1999; 100: 148–55.

Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.

Rudick R, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J, et al. Recommendations from the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force. Ann Neurol 1997; 42: 379–82.

Sharrack B, Hughes RA. The Guy's Neurological Disability Scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler 1999; 5: 223–33.

Sharrack B, Hughes RA, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999; 122: 141–59.

Stewart AL, Ware JE Jr, editors. Measuring functioning and well-being: the medical outcomes study approach. Durham (NC): Duke University Press; 1992.

Stewart AL, Hays RD, Ware JE Jr. The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care 1988; 26: 724–35.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd edn. Oxford: Oxford University Press; 1995.

Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. Qual Life Res 1995; 4: 187–206.

Ware JE Jr, Brook RH, Davies-Avery A, Williams KN, Stewart AL, Rogers WH, et al. Conceptualization and measurement of health for adults in the health insurance study: Vol. I: model of health and methodology. Santa Monica (CA): Rand Corporation; 1980.

Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey manual and interpretation guide. Boston (MA): Nimrod Press; 1993.

Ware JE Jr, Kosinski MA, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston (MA): Health Institute, New England Medical Center; 1994.

Ware JE Jr, Harris WJ, Gandek B, Rogers BW, Reese PR. MAP-R for windows: multitrait/multi-item analysis program—revised user's guide. Boston (MA): Health Assessment Laboratory; 1997.

WHOQOL Group. World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998; 46: 1569–85.

Received November 21, 2000. Revised January 17, 2001.
Accepted January 23, 2001

Appendix I

Multiple Sclerosis Impact Scale (MSIS-29)

- The following questions ask for your views about the impact of MS on your day-to-day life **during the past two weeks**
- For each statement, please **circle** the **one** number that **best** describes your situation
- Please answer **all** questions

In the past two weeks, how much has your MS limited your ability to ...		Not at all	A little	Moderately	Quite a bit	Extremely
1.	**Do physically demanding tasks?**	1	2	3	4	5
2.	**Grip things tightly (e.g. turning on taps)?**	1	2	3	4	5
3.	**Carry things?**	1	2	3	4	5

In the past two weeks, how much have you been bothered by...		Not at all	A little	Moderately	Quite a bit	Extremely
4.	**Problems with your balance?**	1	2	3	4	5
5.	**Difficulties moving about indoors?**	1	2	3	4	5
6.	**Being clumsy?**	1	2	3	4	5
7.	**Stiffness?**	1	2	3	4	5
8.	**Heavy arms and/or legs?**	1	2	3	4	5
9.	**Tremor of your arms or legs?**	1	2	3	4	5
10.	**Spasms in your limbs?**	1	2	3	4	5
11.	**Your body not doing what you want it to do?**	1	2	3	4	5
12.	**Having to depend on others to do things for you?**	1	2	3	4	5

Please check that you have answered all the questions before going on to the next page

In the past two weeks, how much have you been bothered by ...		Not at all	A little	Moderately	Quite a bit	Extremely
13.	**Limitations in your social and leisure activities at home?**	1	2	3	4	5
14.	**Being stuck at home more than you would like to be?**	1	2	3	4	5
15.	**Difficulties using your hands in everyday tasks?**	1	2	3	4	5
16.	**Having to cut down the amount of time you spent on work or other daily activities?**	1	2	3	4	5
17.	**Problems using transport (e.g. car, bus, train, taxi, etc.)?**	1	2	3	4	5
18.	**Taking longer to do things?**	1	2	3	4	5
19.	**Difficulty doing things spontaneously (e.g. going out on the spur of the moment)?**	1	2	3	4	5
20.	**Needing to go to the toilet urgently?**	1	2	3	4	5
21.	**Feeling unwell?**	1	2	3	4	5
22.	**Problems sleeping?**	1	2	3	4	5
23.	**Feeling mentally fatigued?**	1	2	3	4	5
24.	**Worries related to your MS?**	1	2	3	4	5
25.	**Feeling anxious or tense?**	1	2	3	4	5
26.	**Feeling irritable, impatient, or short tempered?**	1	2	3	4	5
27.	**Problems concentrating?**	1	2	3	4	5
28	**Lack of confidence?**	1	2	3	4	5
29.	**Feeling depressed?**	1	2	3	4	5

Please check that you have circled ONE number for EACH question

Copies of the scale can be obtained from the corresponding author.