PAPER

How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with some other self report scales

J C Hobart, A Riazi, D L Lamping, R Fitzpatrick, A J Thompson

J Neurol Neurosurg Psychiatry 2005;76:1539–1543. doi: 10.1136/jnnp.2005.064584

See end of article for authors' affiliations

Correspondence to:
Dr Jeremy Charles Hobart,
Peninsula Medical School,
Derriford Hospital,
Plymouth, Devon PL6 8DH,
UK; Jeremy.Hobart@pms.ac.uk

Received 1 February 2005
Revised version received
1 March 2005
Accepted 23 March 2005

Objectives: To compare the responsiveness of the Multiple Sclerosis Impact Scale (MSIS-29) with other self report scales in three multiple sclerosis (MS) samples using a range of methods. To estimate the impact on clinical trials of differing scale responsiveness.

Methods: We studied three discrete MS samples: consecutive admissions for rehabilitation; consecutive admissions for steroid treatment of relapses; and a cohort with primary progressive MS (PPMS). All patients completed four scales at two time points: MSIS-29; Short Form 36 (SF-36); Functional Assessment of MS (FAMS); and General Health Questionnaire (GHQ-12). We determined: (1) the responsiveness of each scale in each sample (effect sizes): (2) the relative responsiveness of competing scales within each sample (relative efficiency): (3) the differential responsiveness of competing scales across the three samples (relative precision); and (4) the implications for clinical trials (samples size estimates scales to produce the same effect size).

Results: We studied 245 people (64 rehabilitation; 77 steroids; 104 PPMS). The most responsive physical and psychological scales in both rehabilitation and steroids samples were the MSIS-29 physical scale and the GHQ-12. However, the relative ability of different scales to detect change in the two samples was variable. Differing responsiveness implied more than a twofold impact on sample size estimates.

Conclusions: The MSIS-29 was the most responsive physical and second most responsive psychological scale. Scale responsiveness differs notably within and across samples, which affects sample size calculations. Results of clinical trials are scale dependent.

Rating scales are consistently used as outcome measures for clinical trials. As they are the central dependent variables on which treatment decisions are made, they should provide reliable and valid measurements, and detect change. The Multiple Sclerosis Impact Scale (MSIS-29) was developed with these measurement properties in mind,[1] and there is increasing evidence of reliability and validity[1–5] and preliminary evidence of responsiveness.[3 4]

Despite the clinical importance of responsiveness,[6] few studies examine it comprehensively or investigate the implications for clinical trials of differing scale performance. Typically, responsiveness is determined by comparing scores pre-post an intervention expected to produce a change in health. As the interpretation of p values is somewhat binary and sample size dependent,[7] it has become common to report scale responsiveness as an "effect size", or standardised change score, by converting change scores into standard deviation units.

Effect sizes and p values are limited indicators of responsiveness because they are inseparably linked to the magnitude of change.[8] This can be misleading. For example, when change is small the *ability* of a scale to detect change may be mistakenly perceived to be limited. This can be overcome, in part, by comparing rating scales head-to-head in the same sample,[9] which keeps sample and treatment effect constant, and enables investigators to compare the relative responsiveness of competing scales. Even this method only goes part way to determining the *ability* of a scale to detect change because there is no assessment of the extent to which the change detected by a scale is consistent with expectation. Although predicting change is difficult, it can be approximated by examining hypotheses about the differential responsiveness of scales across samples and/or treatments expected to be associated with variable change. We took that approach in this study whose aim was to compare head-to-head the responsiveness of some self report physical and psychological scales for multiple sclerosis (MS), in and across multiple samples, and examine the implications for clinical trials of using different scales.

METHODS

Samples and procedures

Three samples of people with neurologist confirmed MS were invited to participate. All patients were recruited from one clinical centre, the National Hospital for Neurology and Neurosurgery/Institute of Neurology (NHNN/ION), whose ethics committees approved the study. Sample one was consecutive admissions for inpatient multidisciplinary rehabilitation.[10] Sample two was consecutive admissions for intravenous steroid treatment of relapses. Sample three was a natural history cohort of people with primary progressive MS (PPMS).

Data were collected at two time points. Data for sample one were collected within 48 h of admission to, and discharge from, the rehabilitation unit. Data for sample two were collected immediately before, and 6 weeks after, IV steroids; 6 weeks was chosen to represent a time when it was likely that a change would have occurred. These people were invited to attend an outpatient appointment; non-attenders were sent postal questionnaires. Data for sample three were collected via two postal surveys 9 months apart; this was an arbitrary time interval selected to be practical.

Abbreviations: DR, differential responsiveness; ES, effect size; FAMS, Functional Assessment of MS; GHQ-12, 12-item version of the General Health Questionnaire; MS, multiple sclerosis; PPMS, primary progressive MS; RE, relative measurement efficiency; RP, relative measurement precision; SF-36, Short-Form 36 Health Survey; SRM, standardised response means

Outcome measures

All patients completed four self report scales at time 1 and time 2: the MSIS-29[1]; the Short-Form 36 Health Survey (SF-36)[11]; the 59-item Functional Assessment of MS (FAMS)[12]; and the 12-item version of the General Health Questionnaire (GHQ-12).[13]

Responsiveness testing

Analyses were confined to comparing scales measuring the physical (MSIS-29 physical scale; SF-36 physical functioning dimension (SF-36PH); FAMS mobility scale (FAMS MOB)) and psychological (MSIS-29 psychological scale; SF-36 mental health dimension (SF-36MH); FAMS emotional well-being scale (FAMS EWB); GHQ-12) impact of MS.

Effect sizes and standardised response means

The responsiveness of each scale in each sample (that is, rehabilitation, steroids, PPMS) was determined by computing both effect sizes (ES: mean change divided by SD at time 1)[14] and standardised response means (SRM: mean change divided by SD change)[15] as they can produce different values.[16 17] They were interpreted using Cohen's arbitrary criteria (0.2, small; 0.5, moderate; 0.8, large).[18] Analysing scale scores across three samples enabled us to test the clinical hypothesis that change in both physical and psychological health, and therefore apparent instrument responsiveness, should be smallest (or none) in the PPMS group and largest in the steroid group.

Relative measurement efficiency (RE)

The relative responsiveness of competing scales *within each of the two treatment groups* (rehabilitation and steroids samples) was determined by computing relative measurement efficiency (RE). This was not computed for the PPMS sample as this was a natural history cohort rather than a treatment group. Under these circumstances, where change is likely to be very small, indicators of responsiveness that compare scales in proportional terms can give misleading results. Typically, RE is computed as pair wise squared t values (t^2 scale 1/t^2 scale 2),[19] and indicates, as a proportion, how much more (or less) efficient one scale is compared with another at measuring change in that sample. We computed RE as pair wise squared z values from Wilcoxon's signed ranks test as there are concerns that results generated by parametric statistics confound responsiveness with the effects of non-normality such that scales with more normally distributed outcomes are favoured.[20] In each comparison group (for example, physical scales in the rehabilitation sample), the scale with the largest z value was chosen as the denominator for the pair wise calculation. This scale has a measurement precision of 100% and the others are estimated as a per cent of the most responsiveness scale.

Differential responsiveness

The relative responsiveness of competing scales *across the three samples* was determined by computing differential responsiveness, the ability of a scale to detect different degrees of responsiveness in different samples. This approach applies the statistical logic of examining relative measurement precision (RP) in group differences validity.[21] That is, the most responsive scale is the one that best separates the three samples (here in terms of their change scores) relative to the variance within the samples. The *F* statistic from a one way analysis of variance, determines this as it defines the ratio of between-group to within-group variance. Higher *F* statistics indicate greater relative precision. Typically, RP is computed as pair wise *F* statistics (*F* for one scale divided by *F* for the other) as this indicates, as a proportion, how much more (or less) precise one measure is compared with another at detecting group differences.[22] For the reasons discussed above, we computed RP as pair wise χ^2 values from the Kruskal-Wallis H test. For each comparison group (physical or psychological scales) the instrument with the largest χ^2 value was chosen as the denominator in the pair wise computation. This scale has a measurement precision of 100% and the others are estimated as a per cent of this.

Implications of differing responsiveness on sample size estimates

The potential implications for clinical trials of using scales with differing responsiveness was examined by computing the number of patients required for each scale to detect the same effect size. This is typically computed from the square of pair wise standardised response means $\{(\text{SRM scale 1/SRM scale 2})^2\}$,[23] as the sample size required to demonstrate a specified clinical effect, assuming constant power and type 1 error, is inversely proportional to the square of the SRM.[24] For the reasons discussed above, we substituted z values for SRMs in this calculation.

RESULTS

Samples

A total of 245 patients were studied. Table 1 shows their characteristics. Overall, this was an older group of people

Table 1 Sample characteristics

	Sample			
	All	**PPMS**	**Rehabilitation**	**Steroids**
Sample size	245	104	64	77
Age*	46.8 (12.0); 17–83	52.0 (10.3); 28–73	45.3 (11.9); 18–69	41.6 (11.4);17–83
No. female (%)	140 (57.4)	44 (42.7)	41 (64.1)	55 (71.4)
Type of MS (%)†				
PPMS	117 (51.1)	104 (100)	12 (18.8)	1 (1.3)
SPMS	61 (24.9)	0 (0)	40 (62.5)	21 (27.3)
RRMS	67 (27.3)	0 (0)	12 (18.8)	55 (71.4)
Indoor mobility (%)				
Unaided	63 (26.0)	19 (18.8)	3 (4.7)	41 (53.2)
With aid	116 (47.9)	54 (53.5)	29 (45.3)	33 (42.9)
Wheelchair	63 (26.0)	28 (27.7)	32 (50.0)	3 (3.9)
Duration of MS since diagnosis*‡	10.3 (8.3); 0–60	9.1 (4.9); 2–23	12.3 (10.2); 0–35	10.4 (9.7); 1–60
Marital status (%)				
Single	43 (17.7)	17 (16.5)	11 (17.5)	15 (19.5)
Separated	2 (0.8)	0 (0)	1 (1.6)	1 (1.3)
Married	152 (62.6)	70 (68.0)	41 (65.1)	41 (53.2)
Inpatient stay weeks*	N/A	N/A	3.67 (1.2); 2–7	N/A

*Mean (SD); range; †PPMS, SPMS, RRMS: primary progressive, secondary progressive, and relapsing remitting MS, respectively; ‡patient reported.

Table 2 Responsiveness of physical scales

Scale	PPMS	Rehabilitation	Steroids	χ^{2}* (DR)†
MSIS-29	n = 85	n = 60	n = 65	45.8 (100%)
ES (SRM)‡	0.01 (0.02)	0.64 (0.66)	1.01 (1.11)	
z value (p)§	−0.169 (0.866)	−4.255 (0.000)	−6.405 (0.000)	
RE¶		100%	100%	
SF-36PF	n = 80	n = 60	n = 65	22.6 (49%)
ES (SRM)	−0.06 (−0.10)	0.45 (0.35)	0.48 (0.57)	
z value (p)	−0.718 (0.472)	−2.758 (0.006)	−4.393 (0.000)	
RE		42%	47%	
FAMS MOB	n = 85	n = 59	n = 65	34.1 (75%)
ES (SRM)	−0.01 (−0.01)	0.61 (0.58)	0.68 (0.82)	
z value (p)	−0.367 (0.714)	−4.121 (0.000)	−5.425 (0.000)	
RE		94%	72%	

*χ^2 values from Kruskal Wallis ANOVA; †DR (differential responsiveness) = pairwise χ^2 values (for example, for SF-36PF = 100×22.6/45.8); ‡ES, effect size; SRM, standardised response mean; §z values from Wilcoxon's signed ranks test; ¶relative efficiency = pairwise square z values (for example, for SF-36PF in rehabilitation = $100\times(-2.758)^2/(-4.255)^2$].

with MS (mean age 47 years) with well established disease (mean duration 14 years). The mean duration of rehabilitation for the sample was 3.7 weeks, which is representative of the Unit.[25] Differences between the three samples were consistent with clinical expectation: there were more females in the steroid group, more males in the PPMS group, and the rehabilitation group was the most disabled in terms of indoor mobility.

In the PPMS sample, 119 questionnaires were sent at time 1 and 104 completed questionnaires were returned (response rate of 87%). At time 2, questionnaires were sent to all 104 time 1 responders, 88 were returned completed, and three were returned blank (moved house, address unknown) giving a time 2 response rate of 87%.

In the steroids sample (n = 77), 31 people (40%) did not attend their time 2 hospital appointment despite being offered other appointment. Nineteen returned postal questionnaires. Time 2 data were available for 84% (n = 65).

Responsiveness testing

Effect sizes and standardise response means

Tables 2 and 3 show scale responsiveness in the three samples (PPMS, rehabilitation, steroids) for physical (table 2) and psychological (table 3) scales. All scales detected significant changes in both rehabilitation and steroid samples. Four of the seven scales showed a clear stepwise progression in magnitude of ES across the three samples (PPMS<rehabilitation<steroids). One scale (FAMS EWB) had a smaller ES in the steroids group (0.38) than the rehabilitation group (0.52), and two scales (SF-36PF, FAMS MOB) demonstrated almost identical ES in the rehabilitation and steroid samples.

The SRM results were slightly different. Five scales showed the hypothesised stepwise progression, one scale (FAMS EWB) had a larger SRM in the rehabilitation than the steroids sample, and one scale (SF36-MH) had similar values in the two treatment samples. It is notable that the ES and SRM values for each sample varied across scales, as did the extent of the stepwise progression.

In the PPMS sample, all scales detected non-significant changes in physical and psychological health. All three physical scales had near zero ES/SRM implying the scales detected no worsening of self reported physical function over 9 months. All four psychological scales had similar sized negative values suggesting the detection of a small worsening in psychological functioning over this time.

Relative efficiency (RE)

This analysis compares, in proportional terms, the responsiveness of scales *within* each of the two treatment samples (rehabilitation, steroids). There was notable variability in relative efficiency. For example, consider the steroids sample. The MSIS-29 physical scale and GHQ-12 were the most responsive physical and psychological scales because they

Table 3 Responsiveness of psychological scales

Scale	PPMS (n = 104)	Rehabilitation (n = 64)	Steroids (n = 77)	χ^{2}* (DR)†
MSIS-29	n = 85	n = 60	n = 65	39.8 (70%)
ES (SRM)‡	−0.15 (−0.16)	0.44 (0.54)	0.72 (0.90)	
z value (p)§	−1.286 (0.199)	−3.647 (0.000)	−5.826 (0.000)	
RE¶		54%	96%	
SF-36MH	n = 81	n = 60	n = 64	22.4 (39%)
ES (SRM)	−0.13 (−0.17)	0.31 (0.48)	0.44 (0.53)	
z value (p)	−1.448 (0.148)	−3.441 (0.001)	−4.138 (0.000)	
RE		48%	49%	
FAMS EWB	n = 85	n = 60	n = 65	21.5 (38%)
ES (SRM)	−0.04 (−0.05)	0.52 (0.71)	0.38 (0.50)	
z value (p)	−0.408 (0.683)	−4.710(0.000)	−3.981 (0.000)	
RE		90%	45%	
GHQ-12	n = 86	n = 60	n = 65	56.9 (100%)
ES (SRM)	−0.15 (−0.17)	0.60 (0.79)	0.87 (0.94)	
z value (p)	−1.392 (0.164)	−4.954 (0.000)	−5.939 (0.000)	
RE		100%	100%	

*χ^2 values from Kruskal Wallis ANOVA; †DR (differential responsiveness) = pairwise χ^2 values (for example, for SF-36MH = 100×22.4/56.9); ‡ES, effect size; SRM, standardised response mean; §z values from Wilcoxon's signed ranks test; ¶RE (relative efficiency) = pairwise square z values (for example, for SF-36MH in rehabilitation = $100\times(-3.441)^2/(-4.954)^2$).

Table 4 Implications of different responsiveness for sample size calculations

	Sample and sample size*	
Scale	Rehabilitation	Steroids
Physical		
MSIS-29	100	100
SF-36PF	238	213
FAMS MOB	107	139
Psychological		
MSIS-29	185	104
SF-36MH	207	206
FAMS EWB	111	223
GHQ-12	100	100

*Sample size requirements computed as $100\times[(\text{z value scale with largest z value}/\text{z value this scale})^2]$.

had the largest z values. Therefore, they were assigned REs of 100%. Consequently, the SF-36PF was 47% $\{100\times[(-4.393)^2/(-6.405)^2]\}$ as responsive as the MSIS-29 physical scale, and the SF-36MH was 49% $\{100\times[(-4.138)^2/(-5.939)^2]\}$ as responsive as the GHQ-12, in this sample.

The MSIS-29 was the most responsive of the physical scales in both rehabilitation and steroids samples. The SF-36PF detected the greatest negative change (worsening) in the PPMS sample, although the ranges of ES (0.01 to −0.06) and SRM (−0.02 to −0.10) were very small. The GHQ-12 was consistently the most responsive psychological scale. The FAMS EWB scale had similar responsiveness to the GHQ-12 in the rehabilitation sample, and the MSIS-29 psychological scale had similar responsiveness to the GHQ-12 in the steroids group.

Differential responsiveness

This analysis compared scales *across* the three samples, and quantified the relative extent to which scales demonstrated differential responsiveness (DR in tables 2 and 3). Clinically, as a group effect, greater changes would be expected in people admitted for steroid treatment of relapses than in people admitted for rehabilitation. Similarly, we would expect greater change in the group admitted for rehabilitation than in the PPMS sample. The extent to which these differences were manifested is reflected by the magnitude of the χ^2 values.

Table 2 and 3 show that in these samples the MSIS-29 physical scale and GHQ-12 show the greatest differential responsiveness of the physical and psychological scales examined. Compared with the MSIS-29 physical scale, the differential responsiveness of the other two physical scales were 49% (SF-36PF) and 75% (FAMS MOB). Compared with the GHQ-12, the differential responsiveness of the other three psychological scales ranged from 38% (FAMS EWB) to 70% (MSIS-29 psychological scale).

Implications of differing responsiveness on sample size estimates

Table 4 represents different scale responsiveness as sample size estimates required for each scale to achieve the same effect. Values are computed relative to 100 patients using the most responsive scale. For example, for every 100 patients required to demonstrate the effect on physical impact detected by the MSIS-29 physical scale in the steroid sample, it was estimated that the number of patients required to demonstrate the same effect using the other scales ranged from 139 (FAMS MOB) to 213 (SF-36PF). Similarly, for every 100 patients required to demonstrate the effect on psychological health detected by the GHQ-12 in the steroid sample, it was estimated that the number of patients required to demonstrate the same effect using the other scales ranged from 104 (MSIS-29) to 223 (FAMS EWB).

DISCUSSION

The aim of this study was to compare the responsiveness of the MSIS-29 with some other patient report scales that might be used in MS clinical trials. We used multiple techniques to compare multiple scales within and across multiple samples in which different degrees of change were expected. In doing so, we used the fact that responsiveness and the treatment effect are inseparably linked to study differential responsiveness. Also, we have taken the next step of examining the potential implications for clinical trials of using scales with different responsiveness.

The MSIS-29 performed generally well relative to the other scales. Its physical scale had the largest effect sizes and best relative efficiency to detect change in both rehabilitation and steroids samples, and demonstrated the greatest differential responsiveness across the three study samples. The MSIS-29 psychological scale was less successful. It demonstrated less differential responsiveness than the GHQ-12 overall, but a similar ability to detect change in the steroids sample. The GHQ-12 was the most responsive measure of psychological impact in both rehabilitation and steroid samples.

Responsiveness is sample size dependent. It may also depend on where the patients are ''located'' on a scale. This cannot be determined from the comparisons presented as the steroid and rehabilitation samples had heterogeneous mobility. Consequently, we examined responsiveness of physical scales in subsamples defined by self reported mobility level at time 1 (unaided, with aid, wheelchair). This did not impact on the rank ordering of responsiveness. The impact of location on responsiveness of psychological scales could not be studied adequately as we did not have an external indicator of psychological health at time 1.

Our findings have potential implications for clinical trials. First, although all scales demonstrated significant physical and psychological changes in both treatment samples, responsiveness varied markedly in terms of effect sizes, relative efficiency, and differential responsiveness. The clinical implication of this finding is that there will be studies where the results are scale dependent. The difficulty will be to determine in which trials, and using which scales, this is likely to matter. Second, the relative responsiveness of individual scales was sample dependent. This finding further complicates the choice of scales for studies, which often involves extrapolating findings from studies in different samples. Third, variable scale responsiveness had substantial implications for sample size estimation. Typically, power calculations do not account for these differences.

There are, however, issues that render uncertain the direct applicability of our results to clinical trials in MS. First, the data were not collected within the context of a randomised controlled trial. Second, the steroids and rehabilitation groups were heterogeneous in terms of MS type and mobility level. Third, we compared a limited number of scales in small samples from one clinical site. Another limitation is that we have only compared change in scale scores associated with clinician expected change. An equally important, but independent question[26] concerns the relationship between change in scale scores and patient reported change. Nevertheless, this study is one of the larger and more comprehensive evaluations of responsiveness, and has outlined an approach enabling clinicians to test hypotheses of how instruments should perform if they have the *ability* to detect change.

ACKNOWLEDGEMENTS

We thank the patients who participated in this study.

Authors' affiliations

J C Hobart, Murdoch University, Perth, Western Australia

A Riazi, A J Thompson, Neurological Outcome Measures Unit, University College London, London, UK

D L Lamping, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London, UK

R Fitzpatrick, Department of Public Health and Primary Care, University of Oxford, Oxford, UK

This study was funded by grants from the NHS Health Technology Assessment Programme (but the views and opinions expression are not necessarily those of the NHS Executive) and the MS Society of Great Britain and Northern Ireland. Dr Hobart received support from the Royal Society of Medicine (in the form of an Ellison-Cliffe Travelling Fellowship) and the MS Society of Great Britain and Northern Ireland for a recent sabbatical at Murdoch University, Perth, Western Australia where these data were analysed and this paper written.

Competing interests: none declared

REFERENCES

1 **Hobart JC**, Lamping DL, Fitzpatrick R, *et al.* The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain* 2001;**124**:962–73.

2 **Riazi A**, Hobart J, Lamping D, *et al.* Multiple Sclerosis Impact Scale (MSIS-29): reliability and validity in hospital based samples. *J Neurol Neurosurg Psychiatry* 2002;**73**:701–4.

3 **Riazi A**, Hobart J, Lamping D, *et al.* Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. *Mult Scler* 2003;**9**:411–9.

4 **McGuigan C**, Hutchinson M. The Multiple Sclerosis Impact Scale (MSIS-29) is a reliable and sensitive measure. *J Neurol Neurosurg Psychiatry* 2004;**75**:266–9.

5 **Hoogervorst EL**, Zwemmer JN, Jelles B, *et al.* Multiple Sclerosis Impact Scale (MSIS-29): relation to established measures of impairment and disability. *Mult Scler* 2004;**10**:569–74.

6 **Cudkowicz ME**, Schoenfeld D, Williams L. Improving the responsiveness of rating scales: the challenge of stepping twice into the same river. *Neurology* 2004;**62**:1666–7.

7 **Cohen J**. The earth is round (*p*<.05). *Am Psychol* 1994;**49**:997–1003.

8 **O'Connor RJ**, Cano SJ, Thompson AJ, *et al.* Exploring rating scale responsiveness: does the total score reflect the sum of its parts? *Neurology* 2004;**62**:1842–4.

9 **Hobart JC**, Lamping DL, Freeman JA, *et al.* Evidence-based measurement: which disability scale for neurological rehabilitation? *Neurology* 2001;**57**:639–44.

10 **Freeman JA**, Langdon DW, Hobart JC, *et al.* The impact of inpatient rehabilitation on progressive multiple sclerosis. *Ann Neurol* 1997;**42**:236–44.

11 **Ware JE Jr**, Snow KK, Kosinski M, *et al. SF-36 Health Survey manual and interpretation guide*. Boston, MA: Nimrod Press, 1993.

12 **Cella DF**, Dineen K, Arnason B, *et al.* Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. *Neurology* 1996;**47**:129–39.

13 **Goldberg DP**. *Manual of the General Health Questionnaire*. Windsor: NFER-Nelson, 1978.

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

15 **Katz JN**, Larson MG, Phillips CB, *et al.* Comparative measurement sensitivity of short and longer health status instruments. *Med Care* 1992;**30**:917–25.

16 **Liang MH**. Evaluating instrument responsiveness. *J Rheumatol* 1995;**22**:1191–2.

17 **Hobart JC**, Freeman JA, Greenwood RJ, *et al.* Responsiveness of outcome measures: beware the effect of different effect sizes. *Ann Neurol* 1998;**44**:519A.

18 **Cohen J**. *Statistical power analysis for the behavioural sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

19 **Liang MH**, Larson MG, Cullen KE, *et al.* Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;**28**:542–7.

20 **Hobart JC**, Riazi A, Lamping DL, *et al.* Measuring the impact of MS on walking ability: the 12-item MS Walking Scale (MSWS-12). *Neurology* 2003;**60**:31–6.

21 **McHorney CA**, Ware JE Jr, Rogers W, *et al.* The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. *Med Care* 1992;**30**:MS253–65.

22 **Hobart JC**, Kalkers N, Barkhof F, *et al.* Outcome measures for multiple sclerosis clinical trials: relative measurement precision of the Expanded Disability Status Scale and Multiple Sclerosis Functional Composite. *Mult Scler* 2004;**10**:41–6.

23 **Katz JN**, Phillips CB, Fossel AH, *et al.* Stability and responsiveness of utility measures. *Med Care* 1994;**32**:183–8.

24 **Snedecor G**, Cochran W. *Statistical methods*, 8th ed. Ames, IA: Iowa State University Press, 1967.

25 **Freeman JA**, Playford ED, Nicholas RS, *et al.* A neurological rehabilitation unit: audit of activity and outcome. *J R Coll Physicians Lond* 1996;**30**:21–6.

26 **Norman GR**, Stratford, P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997;**50**:869–79.