Ratings 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, and there is increasing evidence of reliability and validity and preliminary evidence of responsiveness. Despite the clinical importance of responsiveness, 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, it has become common to report scale responsiveness as a "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. 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, 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. 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 mean.
Outcome measures
All patients completed four self-report scales at time 1 and time 2; the MSIS-29; the Short-Form 36 Health Survey (SF-36); the 59-item Functional Assessment of MS (FAMS); and the 12-item version of the General Health Questionnaire (GHQ-12).

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 EWBI); 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) and standardised response means (SRM: mean change divided by SD change) as they can produce different values. They were interpreted using Cohen’s arbitrary criteria (0.2, small; 0.5, moderate; 0.8, large). 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 psychological health, and therefore apparent instrument responsiveness, should be smallest (or none) in the PPMS sample as this was determined by computing differential responsiveness validity. 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. For the reasons discussed above, we computed RP as pair wise F values from the Kruskal-Wallis H test. For each comparison group (physical or psychological scales) the instrument with the largest F 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.

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 (t² scale 1/t² scale 2), 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 this was more (or less) precise one measure is compared with another at detecting group differences. Typically, RE is computed as pair wise squared z values from Wilcoxon’s signed ranks test as this indicates, as a proportion, how much more (or less) precise one measure is compared with another at detecting group differences. Typically, RE is computed as pair wise standardised response means \( \left( \frac{SRM \text{ scale 1}}{SRM \text{ scale 2}} \right) \), as they can produce different values. They were interpreted using Cohen’s arbitrary criteria (0.2, small; 0.5, moderate; 0.8, large). 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.

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 \( \left( \frac{1}{SRM \text{ scale 1/SRM scale 2}} \right) \), 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. For the reasons discussed above, we substituted z values for SRMs in this calculation.

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

<table>
<thead>
<tr>
<th>Sample characteristics</th>
<th>All</th>
<th>PPMS</th>
<th>Rehabilitation</th>
<th>Steroids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>245</td>
<td>104</td>
<td>64</td>
<td>77</td>
</tr>
<tr>
<td>Age*</td>
<td>46.8 (12.0); 17-83</td>
<td>52.0 (10.3); 28-73</td>
<td>45.3 (11.9); 18-69</td>
<td>41.6 (11.4);17-83</td>
</tr>
<tr>
<td>No. female (%)</td>
<td>140 (57.4)</td>
<td>44 (22.7)</td>
<td>41 (64.1)</td>
<td>55 (71.4)</td>
</tr>
<tr>
<td>Type of MS (%)†</td>
<td>PPMs 117 (51.1)</td>
<td>104 (100)</td>
<td>12 (18.8)</td>
<td>1 (1.3)</td>
</tr>
<tr>
<td>SPMS 61 (24.9)</td>
<td>0 (0)</td>
<td>40 (62.5)</td>
<td>21 (27.3)</td>
<td></td>
</tr>
<tr>
<td>RRMS 67 (27.3)</td>
<td>0 (0)</td>
<td>12 (18.8)</td>
<td>55 (71.4)</td>
<td></td>
</tr>
<tr>
<td>Indoor mobility (%)</td>
<td>Unaided 63 (26.0)</td>
<td>19 (18.8)</td>
<td>3 (4.7)</td>
<td>41 (53.2)</td>
</tr>
<tr>
<td>With aid 116 (47.9)</td>
<td>54 (53.9)</td>
<td>29 (45.3)</td>
<td>33 (42.9)</td>
<td></td>
</tr>
<tr>
<td>Wheelchair 63 (26.0)</td>
<td>28 (27.7)</td>
<td>32 (50.0)</td>
<td>3 (5.9)</td>
<td></td>
</tr>
<tr>
<td>Duration of MS since diagnosis‡</td>
<td>Unaided 10.3 (8.3); 0–60</td>
<td>9.1 (4.9); 2–23</td>
<td>12.3 (10.2); 0–35</td>
<td>10.4 (9.7); 1–60</td>
</tr>
<tr>
<td>Marital status (%)</td>
<td>Single 43 (17.7)</td>
<td>17 (16.5)</td>
<td>11 (17.9)</td>
<td>15 (19.9)</td>
</tr>
<tr>
<td>Separated 2 (0.8)</td>
<td>0 (0)</td>
<td>1 (1.6)</td>
<td>1 (1.3)</td>
<td></td>
</tr>
<tr>
<td>Married 152 (62.6)</td>
<td>70 (68.0)</td>
<td>41 (65.1)</td>
<td>41 (53.2)</td>
<td></td>
</tr>
<tr>
<td>Inpatient stay weeks*</td>
<td>N/A</td>
<td>N/A</td>
<td>3.67 (1.2); 2–7</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Mean (SD); †PPMS, SPMS, RRMS: primary progressive, secondary progressive, and relapsing remitting MS, respectively; ‡patient reported.
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 (SF-36-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 2  responsiveness of physical scales**

<table>
<thead>
<tr>
<th>Scale</th>
<th>PPMS (n = 104)</th>
<th>Rehabilitation (n = 64)</th>
<th>Steroids (n = 77)</th>
<th>z*2 (DR†)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSIS-29</td>
<td>n = 85</td>
<td>n = 60</td>
<td>n = 65</td>
<td>39.8 (70%)</td>
</tr>
<tr>
<td>ES (SRM)</td>
<td>0.15 (0.16)</td>
<td>0.44 (0.54)</td>
<td>0.72 (0.90)</td>
<td>96%</td>
</tr>
<tr>
<td>z value (p)</td>
<td>-1.286 (0.199)</td>
<td>3.647 (0.000)</td>
<td>5.826 (0.000)</td>
<td>48%</td>
</tr>
<tr>
<td>RE</td>
<td>54%</td>
<td>96%</td>
<td>1.392 (0.164)</td>
<td>45%</td>
</tr>
<tr>
<td>SF-36PF</td>
<td>n = 81</td>
<td>n = 60</td>
<td>n = 64</td>
<td>22.4 (39%)</td>
</tr>
<tr>
<td>ES (SRM)</td>
<td>-0.13 (0.17)</td>
<td>0.31 (0.48)</td>
<td>0.44 (0.53)</td>
<td>49%</td>
</tr>
<tr>
<td>z value (p)</td>
<td>-1.448 (0.148)</td>
<td>3.441 (0.001)</td>
<td>4.358 (0.000)</td>
<td>49%</td>
</tr>
<tr>
<td>RE</td>
<td>48%</td>
<td>49%</td>
<td>0.90</td>
<td>45%</td>
</tr>
<tr>
<td>FAMS EWB</td>
<td>n = 85</td>
<td>n = 60</td>
<td>n = 65</td>
<td>21.5 (38%)</td>
</tr>
<tr>
<td>ES (SRM)</td>
<td>-0.04 (0.05)</td>
<td>0.52 (0.71)</td>
<td>0.38 (0.50)</td>
<td>45%</td>
</tr>
<tr>
<td>z value (p)</td>
<td>-0.408 (0.663)</td>
<td>-4.710 (0.000)</td>
<td>-3.981 (0.000)</td>
<td>45%</td>
</tr>
<tr>
<td>RE</td>
<td>90%</td>
<td>90%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>GHQ-12</td>
<td>n = 86</td>
<td>n = 60</td>
<td>n = 65</td>
<td>56.9 (100%)</td>
</tr>
<tr>
<td>ES (SRM)</td>
<td>-0.15 (0.17)</td>
<td>0.60 (0.79)</td>
<td>0.87 (0.94)</td>
<td>100%</td>
</tr>
<tr>
<td>z value (p)</td>
<td>-1.392 (0.164)</td>
<td>4.954 (0.000)</td>
<td>-5.939 (0.000)</td>
<td>100%</td>
</tr>
</tbody>
</table>

* z2 values from Kruskal Wallis ANOVA; †DR (differential responsiveness) = pairwise z2 values (for example, for SF-36PF = 100 x 22.6/45.8); zES, effect size; SRM, standardised response mean; z2 values from Wilcoxon’s signed ranks test; *RE (relative efficiency) = pairwise square z values (for example, for SF-36PF in rehabilitation = 100 x (−3.441)2/ (−4.255)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 \( \chi^2 \) 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 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.
How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)

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 expressed 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 Free Hospital, London, UK.

The Multiple Sclerosis Impact Scale (MSIS-29) is a reliable and sensitive measure. J Neurol Neurosurg Psychiatry 2001;72:462–70.

References