Psychometric properties of the Fatigue Severity Scale: Rasch analyses of responses in a Norwegian and a Swedish MS cohort

A Lerdal¹,², S Johansson³,⁴, A Kottorp⁵ and L von Koch³,⁵

Abstract
Background: Rigorous testing of the original Fatigue Severity Scale (FSS–9) with modern psychometric methods is warranted.
Objective: To determine the psychometric properties of the FSS–9 in multiple sclerosis (MS): internal scale validity; person response validity; unidimensionality; uniform differential item functioning; temporal stability of response patterns; and ability to separate people into distinct groups of fatigue.
Methods: Rasch analyses were conducted on data from a Norwegian and a Swedish MS cohort followed for two years.
Results: Item estimations in the FSS–9 did not differ between sex or levels of education but between the cohorts with regard to disability, disease course and time for evaluation, however, items 1 and 2 demonstrated unacceptable high outfit mean-square values in both cohorts. In an FSS–7 item version, items 3 and 4 in the Norwegian and 4 in the Swedish cohort demonstrated unacceptable goodness of fit but high separation indexes. In the FSS–7, the first unidimensional factor explained 87.5% (Norwegian cohort) and 86.4% (Swedish cohort) of the total variation.
Conclusions: In MS, the FSS–7 demonstrates better psychometric properties than the FSS–9; items 1 and 2 neither empirically nor conceptually fit with the other seven items.

Keywords
fatigue, multiple sclerosis, psychometrics, questionnaires, Rasch analysis

Date received: 15th November 2009; revised: 28th January 2010 and 1st April 2010; accepted: 3rd April 2010

Introduction

Fatigue is a common and disabling symptom in neurological diseases for example, in multiple sclerosis (MS). The prevalence of fatigue in people with MS has been reported to be between 55–83%.¹ ²

The causes of fatigue in people with MS may be a result of processes within the central nervous system, and/or of immune and/or neuroendocrine dysregulation.³ ⁴ Fatigue may also arise as a secondary effect of depressive symptoms,⁵ ⁶ impaired sleep,⁷ heat sensitivity,⁸ physical deconditioning and medications.⁹ Fatigue in people with MS can be defined as a sense of exhaustion, lack of perceived energy and tiredness.⁹ Because of the subjective nature of fatigue, several inventories based on self-report have been developed.

A frequently used inventory for measuring fatigue is the Fatigue Severity Scale (FSS), a nine-item unidimensional questionnaire developed by Krupp et al.¹⁰ (Table 1). Each item is scored on a seven-point Likert scale ranging from 1 (‘completely disagree’) to 7 (‘completely agree’). The mean score of the nine items is used as the FSS score. In order to categorize severity of fatigue in people with MS, researchers use different criteria. Originally the cut-off for severe fatigue was set at an FSS score ≥4.¹⁰ This cut-off is still used in studies of post-stroke fatigue.¹¹ However, more recent studies of people with MS mainly use a cut-off of ≥5.⁵ ⁶ ¹² To our knowledge, the different cut-off values have not been validated clinically.

¹Department of Health Sciences, Buskerud University College, Drammen, Norway.
²Research Centre, Oslo University Hospital Aker, Oslo, Norway.
³Division of Neurology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.
⁴Division of Physiotherapy, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.
⁵Division of Occupational Therapy, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.

Corresponding author:
Anners Lerdal, Buskerud University College, Department of Health Sciences, PO Box 7053, 3007 Drammen, Norway
Email: anners.lerdal@idh.no
In two studies of the effects of medical treatment on fatigue in people with MS, the FSS scores have proved to be sensitive to measure change in levels of fatigue. Studies on the effect of multidisciplinary rehabilitation on fatigue have reported no change in fatigue when using the FSS, which might be a result of insensitiveness of the FSS, or an ineffective intervention.

Assessment of the psychometric properties of the FSS based on classical test theory has shown that the FSS scores correlate strongly with the scores of other fatigue scales. Furthermore, studies have documented high internal consistency as analysed with Cronbach’s alpha, and high test–retest reliability. However, when inter-item correlations were analysed in a sample from the general population, items 1 and 2 showed a relatively low correlation with the rest of the items.

Although the FSS has been found satisfactory when using classical test theory, a recent study using modern test theory (Rasch analysis) showed inconsistent responses in the FSS. In 416 people with MS, items 1 and 2 had high positive residual statics; item 6 showed uniform differential item functioning (DIF) by age, disability and disease course and, for item 8, by disability. With these four items removed, the remaining five-item model showed satisfactory unidimensionality. The responses in the revised model were not biased with regard to age, sex, disease duration, level of disability and disease course. Conceptually, the five items may well be related to consequences of fatigue, whereas the removed items relate to causes and experiences of fatigue.

Considering the high prevalence of fatigue in people with MS and its impact on functioning and quality of life, it is of utmost importance to establish conceptually sound instruments with good psychometric properties for the assessment of fatigue, both for use in research and in the clinic. Such instruments should provide valid measures of fatigue regardless of sex, educational background and culture. Previous analyses of the FSS using the Rasch model were also applied, based on cross-sectional data from a sample with an attrition rate of almost 50% and with no analysis of DIF in relation to the development of fatigue over time. As the disease course and the severity of MS may influence the level and pattern of fatigue over time, it is important to make sure that comparisons of the FSS scores over time are not biased. In addition, person–response validity has not been addressed in earlier studies. Thus additional in-depth analyses are needed in order to establish the most useful version of the FSS for people with MS.

The aim of this study was to examine the psychometric properties of the FSS in people with MS from Norway and Sweden: 1) the presence of uniform DIF in relation to sex, level of education and country, 2) aspects of internal scale validity, person–response validity and unidimensionality, 3) the stability of the response patterns over time with regard to disability, disease course and time for evaluation, and 4) the ability of the FSS to sort people into distinct groups of fatigue.

### Methods

The data set was composed of data from two longitudinal studies: an MS population cohort in the city of Oslo, Norway and an MS cohort recruited at a specialist outpatient clinic at Karolinska University Hospital in Huddinge, Stockholm, Sweden.

#### The Norwegian sample

Based on the Oslo City MS Registry, 502 people with definite MS according to Poser criteria were invited to participate in the study. Via ordinary mail they received a questionnaire at three time points in May/June, at 1-year intervals, starting in 2000. In addition to the FSS, the questionnaire covered demography, MS-related issues, other health-related factors, psychological distress, coping and quality of life. The characteristics of the sample at baseline have been reported elsewhere, as have the variations and predictors of fatigue over the 2 years. The Norwegian version of the FSS was filled in by 368 respondents at baseline, 313 at 12, and by 267 at 24 months.

#### The Swedish sample

In the Swedish sample, all those with definite MS according to Poser criteria who were scheduled for an outpatient appointment with their senior neurologist were eligible. Data on clinical and contextual
characteristics, functioning and disability were collected at five time points at 6-month intervals and the results have been reported elsewhere.\textsuperscript{25} as have the variations in fatigue and predictors of fatigue over the 2 years.\textsuperscript{3} The Swedish version of the FSS was filled in by 224 respondents at baseline, 219 at 6 months, 212 at 12, 203 at 18, and by 199 at 24 months.

**Variables used for analysis**

Fatigue was assessed with the FSS. Disability in the Norwegian cohort was categorized according to the use of technical mobility aids: 1) never/seldom use of walking aid; 2) sometimes/often use of walking aid; 3) regular use of wheelchair or bed bound. Disability in the Swedish cohort was categorized according to the Expanded Disability Status Scale (EDSS)\textsuperscript{26}: 1) no walking restriction (EDSS 0–3.5); 2) restricted walking capacity (EDSS 4.0–5.5); 3) regular use of wheelchair or bed bound (EDSS 6.0–9.5). Data on age (years) and sex, time since onset (years), and disease course defined as relapsing–remitting including secondary progressive or primary progressive, were collected from the medical records. Self-reported level of formal education was categorized as low (7–10 years), medium (11–13 years) and high (14 years and above). While data from the Norwegian cohort were collected three times at yearly intervals, data from the Swedish cohort were collected five times, at six-monthly intervals. We chose to use this data as a longitudinal variable (here called time for evaluation) (0, 6, 12, 18 and 24 months) in order to evaluate whether the pattern of responses was also stable over time in the sub-samples.

**Statistical analysis**

Rasch models have been increasingly used for evaluating the psychometric properties of both new and existing instruments.\textsuperscript{27} The Rasch programs convert the raw item scores from a test or questionnaire into equal-interval measures. Furthermore, they can be used to examine whether items from a scale measure a unidimensional construct, which is viewed as crucial in measurement statistics.\textsuperscript{28}

In this study, an application of a Rasch model was chosen for the following reasons: firstly, the items included in the FSS represent different aspects of fatigue interference in everyday life, which may be assumed to vary in level of severity. The Rasch model takes into account the actual challenge of each item scored and adjusts the final person-ability measure based on differences in item challenge. Secondly, Rasch models are suitable for handling datasets, where all items are not scored across all people as compared to more traditional approaches, where comparisons between people are only possible when they have been scored on the same items.\textsuperscript{28–30}

The FSS raw scores were analysed using the WINSTEPS Rasch analysis software program, version 3.63.0.\textsuperscript{30} This program performs a logarithmic transformation, simultaneously resulting in an estimation of a person’s fatigue interference measure and the nine-item difficulty calibrations along a line of the phenomenon to be measured. Rasch models are probabilistic, and states are based on theoretical assertions against which the actual pattern of responses is validated. Although the scales in FSS, as other scales used in medicine, are often viewed and used as rating scales, they may not function in a similar manner across all items.\textsuperscript{31} Therefore a partial credit model, developed for scales where the rating scale may differ across items, was applied in the analysis of FSS in this study.

As an initial step, an evaluation of the psychometric properties of the rating scale was conducted, following Linacre’s guidelines on essential features of a scale.\textsuperscript{32} According to these guidelines, it is important that the average measures for each category on each item should advance monotonically. This means that if the rating score 5 should happen to demonstrate a higher perceived interference of fatigue than the score 4, then the perceived interference of fatigue in people who receive raw scores of 5 should be higher than the perceived interference among people who receive raw scores of 4. The step calibration measures were also required to advance with the scoring category. Finally, a criterion less than 2.0 was recommended in outfit mean-square (MnSq) values for the step calibrations. This criterion will be more described in depth below.\textsuperscript{33}

Different aspects of the evidence of internal-scale validity and person–response validity in the data were further investigated using the item and person goodness-of-fit statistics. The WINSTEPS program generates goodness-of-fit statistics for the items and persons expressed as MnSq residuals and standardized z values. These indicate the degree of match between the actual responses on the FSS and the expected responses of the Rasch rating scale model for each person and item.

Furthermore, the goodness-of-fit statistics are evaluated using infit and outfit statistics. Infit statistics are information-weighted fit statistics and give relatively more weight to the performances of persons who are well targeted to the item, difficulty calibrations. Outfit statistics are not weighted, and therefore are more sensitive to outlying scores, i.e. when a person with high perceived impact of fatigue achieves a low score on a demanding item or when a person with low perceived impact of fatigue achieves a high score on a less demanding item.\textsuperscript{28} According to the literature, there
are different opinions on which of these statistics is more important in evaluating the internal scale or person–response validity. In accordance with the literature, we chose an additional criterion for assessing the quality of a rating scale by the analysis of goodness-of-fit statistics: outfit MnSq values less than 2.0, as described earlier.\textsuperscript{32} As the infit statistics are more sensitive to item performance and more informative when exploring internal scale validity,\textsuperscript{34,35} we chose the infit statistics for evaluation of goodness of fit across individual items and people in this study, also in accordance to other empirical studies.\textsuperscript{36,37}

The MnSq fit statistics have an expected value of 1.0 and \( z \), an expected value of 0.0. One commonly used criterion has been to accept MnSq values less than 1.4 logit with an associated \( z \)-value less than 2 both for individual items and persons.\textsuperscript{36,37} Other studies using simulated and dichotomous data suggest a more conservative and sample-dependent approach when investigating individual item fit.\textsuperscript{36} In this study we have chosen to use a sample-size related criterion for item goodness-of-fit set for MnSq values\textsuperscript{38} with an associated \( z \) < 2. As our criteria, we determined that all items in FSS should demonstrate acceptable goodness of fit to the model.

Further, persons with a MnSq less than 1.4 that were associated with a \( z \) less than 2.0 were considered to meet the criteria for acceptable individual person goodness of fit. It is generally accepted that 5% of the persons may by chance not demonstrate acceptable goodness of fit without being a serious threat to person–response validity.\textsuperscript{36,37}

In order to minimize the risk of additional explanatory factors in the dataset, a principal component analysis of the residuals was also performed.\textsuperscript{39} The criteria set for this analysis were that at least 60% of the total variation in the dataset should be explained by the first latent variable (perceived impact of fatigue) and that any additional factor should not explain more than 5% of the remaining variation of the residuals, associated with an eigenvalue equal to or less than 1.4, after the removal of the first latent variable,\textsuperscript{40,41} in order to further support unidimensionality.

To further determine whether the FSS could distinguish people with different levels of fatigue, person-separation reliability was investigated, with regard to the individual and group standard error (SE) of measures. For a scale to distinguish between three or more groups, reliability of at least 0.80 is recommended,\textsuperscript{42} with an associated person-separation index of 2.0.

A stepwise analytical procedure was used to assess the psychometric properties of the FSS. The first step included analysis for Norway and Sweden respectively of all the sets of complete FSS and all nine items included. First, the overall goodness of fit of all the included data was evaluated and the psychometric properties described above and based on the Rasch measurement analysis were investigated for each country. If one item or more did not demonstrate acceptable goodness of fit to the model according to the criteria, one item at a time was removed and the psychometric properties were analysed again, with the remaining items. The procedure was repeated until all items but one demonstrated acceptable goodness of fit to the model. We finally analysed our data using the five-item model suggested by Mills et al.\textsuperscript{21} DIF was further investigated to determine whether the cohorts in the different countries were stable across items for different subgroups of people with MS with regard to country, sex, and level of education. Depending on the number of items in a scale, it is generally accepted that no more than one item in a scale, or 5% of the items should demonstrate DIF in order to further support unidimensionality.\textsuperscript{36,37} For the FSS scale to be considered stable across external variables, we expected to find no item demonstrating DIF, as this short scale only includes nine items.

Finally, additional DIF analyses were performed in order to evaluate the stability of the response patterns over time, disability and disease course. We chose to evaluate the magnitude of DIF in the dataset in all analyses using the Mantel–Haenszel statistics for polytomous scales using log-odds estimators,\textsuperscript{43,44} as reported from the WINSTEPS program, (1% alpha with Bonferroni correction).

**Ethics**

The study was approved by the Regional Medical Research Ethics Committee of Health East of Norway and the Norwegian Data Inspectorate for the Norwegian sample and by the Ethical Committee at Karolinska Institutet in Stockholm, for the Swedish sample. Informed consent was obtained from all respondents.

**Results**

**Characteristics of the sample**

Age, sex, level of education, time since diagnosis, disability, and disease of the sample at baseline are described in Table 2.

**FSS–9 items**

When evaluating rating-scale functioning in the FSS nine-item scale using non-repeated measures for the Norwegian and Swedish cohorts, respectively, two items (1 and 2) did not meet the essential criteria.
The average measures for each category and step calibration measures did not advance monotonically, and furthermore they were associated with higher than acceptable outfit MnSq values (Tables 3 and 4). All other items in the scale demonstrated acceptable profiles on the essential criteria for a scale.

Secondly, we analysed the presence of DIF in relation to country (Norway/Sweden). If no presence of DIF between countries was discovered, continued analyses could be performed with pooled data. The DIF analysis did, however, indicate relative differences between the Norwegian and Swedish cohorts in items 2, 5, and 6 (Table 5). We chose to continue our analyses with separate analyses.

In the analysis of item fit to the model, using all nine items of the FSS, item 1 and item 2 did not demonstrate acceptable goodness of fit to the model both in the Norwegian and the Swedish cohort. Since the number of items not demonstrating acceptable goodness of fit, exceeded our criterion for an acceptable solution, we continued our analysis by removing the item that least demonstrated acceptable goodness of fit, and also did not demonstrate acceptable measures on rating-scale functioning, i.e., item 1. Additional information about unidimensionality, person–response validity and person-separation reliability for the different cohorts of the nine-item FSS is given in Tables 3 and 4.

In the second analysis, including the remaining eight items, two items (2 and 4) did not demonstrate acceptable goodness of fit in the Norwegian data. Item 2 did not meet the criterion in the Swedish data, either. Additional information about unidimensionality, person–response validity and person-separation reliability for the FSS eight-item and the following solutions are given in Tables 3 and 4.

We preceded our step-by-step removal process of FSS items until we reached the criterion set that all items should demonstrate acceptable goodness of fit. This was reached when four items, respectively, had been removed from the two datasets: items 1, 2, 3 and 4 from the Norwegian data and items 1, 2, 4 and 6 from the Swedish data.

The final five-item solutions both demonstrated acceptable person–response validity, but the separation index was lower than expected in the Norwegian cohort. The criteria set for unidimensionality were also met in the five-item solution for the Swedish, but not for the Norwegian cohort.

**FSS–5 items**

In the final Rasch analysis, we evaluated our empirical data using the suggested five-item solution by removing items 1, 2, 6 and 8, as presented earlier in the literature. All items but one (item 4) demonstrated acceptable goodness of fit to the model in both the Norwegian and the Swedish data. The principal component analysis partly supported a unidimensional construct, where
83.8% of the total data variation was explained by the first unidimensional factor, and 4.2% of the residuals could be explained by the second factor. But the eigenvalue exceeded the criteria set in both cohorts which may indicate a second dimension in the data.40,41 The FSS five-item solution also met the set criterion for acceptable person–response validity in the Swedish dataset, but not in the Norwegian data. Finally, the suggested FSS five-item solution demonstrated a person-separation measure similar or somewhat lower than the empirically suggested country-specific five-item solutions.

Tables 3 and 4 also present the number of participants receiving maximum and minimum scores (ceiling

### Table 3. Psychometric properties of the different Fatigue Severity Scale (FSS) item solutions in the Norwegian cohort

<table>
<thead>
<tr>
<th>Item solution</th>
<th>FSS 9-items (n = 362)</th>
<th>FSS 8-items Item 1 removed (n = 361)</th>
<th>FSS 7-items Items 1 and 2 removed (n = 361)</th>
<th>FSS 6-items Items 1, 2, and 4 removed (n = 361)</th>
<th>FSS 5-items Items 1, 2, 3 and 4 removed (n = 361)</th>
<th>FSS 5-items Items 1, 2, 6 and 8 removed (n = 361)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items not meeting all three criteria for rating scale</td>
<td>1</td>
<td>2</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Item misfit</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>None</td>
<td>4</td>
</tr>
<tr>
<td>First latent variable, %</td>
<td>86.3</td>
<td>86.9</td>
<td>87.5</td>
<td>85.6</td>
<td>84.3</td>
<td>83.3</td>
</tr>
<tr>
<td>2nd dimension, %</td>
<td>2.8</td>
<td>3.0</td>
<td>3.0</td>
<td>4.1</td>
<td>5.6</td>
<td>4.9</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>1.8</td>
<td>1.8</td>
<td>1.7</td>
<td>1.7</td>
<td>1.8</td>
<td>1.5</td>
</tr>
<tr>
<td>Person misfit, n (%)</td>
<td>19 (5.2)</td>
<td>22 (6.1)</td>
<td>22 (6.1)</td>
<td>15 (4.2)</td>
<td>14 (3.9)</td>
<td>24 (6.6)</td>
</tr>
<tr>
<td>Maximum score, n (%)</td>
<td>25 (6.9)</td>
<td>26 (7.2)</td>
<td>33 (9.1)</td>
<td>38 (10.5)</td>
<td>43 (11.9)</td>
<td>39 (10.8)</td>
</tr>
<tr>
<td>Minimum score, n (%)</td>
<td>3 (0.8)</td>
<td>7 (1.9)</td>
<td>10 (2.8)</td>
<td>19 (5.3)</td>
<td>25 (5.3)</td>
<td>10 (2.8)</td>
</tr>
<tr>
<td>Person-separation index (without extremes)</td>
<td>2.18</td>
<td>2.17</td>
<td>2.09</td>
<td>1.92</td>
<td>1.72</td>
<td>1.76</td>
</tr>
<tr>
<td>Person-separation reliability</td>
<td>0.83</td>
<td>0.83</td>
<td>0.81</td>
<td>0.79</td>
<td>0.75</td>
<td>0.76</td>
</tr>
</tbody>
</table>

### Table 4. Psychometric properties of the different Fatigue Severity Scale (FSS) item solutions in the Swedish cohort

<table>
<thead>
<tr>
<th>Item solution</th>
<th>FSS 9-items (n = 224)</th>
<th>FSS 8-items Item 1 removed (n = 224)</th>
<th>FSS 7-items Items 1 and 2 removed (n = 224)</th>
<th>FSS 6-items Items 1, 2, and 4 removed (n = 224)</th>
<th>FSS 5-items Items 1, 2, 3 and 4 removed (n = 224)</th>
<th>FSS 5-items Items 1, 2, 6 and 8 removed (n = 224)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items not meeting all three criteria for rating scale</td>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Item misfit</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>None</td>
<td>4</td>
</tr>
<tr>
<td>First latent variable, %</td>
<td>81.0</td>
<td>83.2</td>
<td>86.4</td>
<td>87.5</td>
<td>87.1</td>
<td>82.7</td>
</tr>
<tr>
<td>2nd dimension, %</td>
<td>4.3</td>
<td>4.4</td>
<td>3.8</td>
<td>3.6</td>
<td>3.6</td>
<td>5.5</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>2.1</td>
<td>2.1</td>
<td>2.0</td>
<td>1.7</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Person misfit, n (%)</td>
<td>12 (5.4)</td>
<td>12 (5.4)</td>
<td>11 (5.3)</td>
<td>10 (4.5)</td>
<td>9 (4.0)</td>
<td>8 (4.0)</td>
</tr>
<tr>
<td>Maximum score, n (%)</td>
<td>6 (2.7)</td>
<td>8 (3.6)</td>
<td>15 (6.7)</td>
<td>15 (6.7)</td>
<td>17 (7.6)</td>
<td>18 (8.0)</td>
</tr>
<tr>
<td>Minimum score, n (%)</td>
<td>1 (0.4)</td>
<td>3 (1.3)</td>
<td>3 (1.3)</td>
<td>4 (1.8)</td>
<td>6 (2.7)</td>
<td>4 (1.8)</td>
</tr>
<tr>
<td>Person-separation index (without extremes)</td>
<td>2.22</td>
<td>2.25</td>
<td>2.25</td>
<td>2.17</td>
<td>2.10</td>
<td>1.91</td>
</tr>
<tr>
<td>Person-separation reliability</td>
<td>0.83</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td>0.82</td>
<td>0.78</td>
</tr>
</tbody>
</table>
People living with fatigue have limited ability to participate in research. The development of valid, reliable and time-effective instruments is, therefore, of utmost importance. Our study showed that, even though the original FSS consists of only nine items, items can be removed and thereby reduce the workload for participants. Our conclusion is that the nine-item solution should be used with caution in research where a summative score for people with MS.

Discussion

The relatively high person-separation index and person-separation reliability in the FSS–7 in both cohorts indicate that this version is a more sensitive instrument for measuring changes in fatigue compared to the FSS–5. The FSS–7, therefore, has a better potential for measuring changes over time, e.g., in intervention studies. On the other hand, two items (5 and 6) demonstrated the presence of DIF in relation to differences between the countries. The higher proportion of misfit and the stronger ceiling effect in the Norwegian data set can probably at least partially be explained by a higher proportion of persons with severe disability than in the Swedish cohort (Table 2). This may suggest that the DIF by country is partly explained by the DIF by disability later explored in the analysis. The results indicate that the item hierarchy in FSS also differs in relation to disease course and time for evaluation. For example, disability was related to differences in responses in four of the seven items. It is not known whether ongoing pathological processes confound the relationships between disability and the item hierarchy in the FSS–7, or whether the differences in item hierarchies logically reflect disease-related characteristics of fatigue. If the differences are related to pathological processes related to fatigue, investigations into the difference in response on these items, might lead to enhanced knowledge concerning fatigue and its aetiology. Our findings on the relationship between DIF and clinical variables should be interpreted with caution. There is also a potential risk that dependency caused by repeated measuring may have influenced the independence of the responses negatively and affected the estimates. Thus, further investigation is necessary. Our results indicate that direct comparison of sums from the FSS in samples where people vary in relation to these socio-demographic and clinical variables should be undertaken with caution. The use of Rasch analysis may support solving some of these issues using item splitting techniques, where the analysis allows some items to have unique specific measures by country.

On the other hand, it is important to carefully consider the potential underlying reasons for DIF, before choosing such solutions. We chose a conservative approach in this explorative study not to split item directly by country based on the initial analyses (Table 5). The latter analyses did actually support that the DIF found by country could rather be a DIF by severity of disability and/or time, and may then not indicate a split item analysis by country as an appropriate solution. The present DIF in relation to rather high eigenvalues in the seven-item solution may also indicate a second dimension in the data. These findings indicate a need for further studies exploring the relationships between country and sample characteristics and also

### Table 5. Differential item functioning (DIF) analysis of the Fatigue Severity Scale (FSS) in a Norwegian/Swedish sample

<table>
<thead>
<tr>
<th></th>
<th>FSS 9-items (n = 595)</th>
<th>FSS 9-items (n = 2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Gender</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Level of formal education</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Disability (EDSS)</td>
<td>1, 2, 4, 6, 8, 9</td>
<td>1, 2, 4, 6, 8, 9, 9</td>
</tr>
<tr>
<td>Disease course</td>
<td>1, 2, 8</td>
<td>1, 2, 8</td>
</tr>
<tr>
<td>Time for evaluation</td>
<td>1, 2, 6</td>
<td>1, 2, 6</td>
</tr>
</tbody>
</table>

EDSS, expanded disability status scale.

### Strengths and limitations of the FSS–7

and floor effect) across the different solutions. This presentation indicates that the five-item solutions increase the proportion of participants demonstrating maximum or minimum scores on the FSS.

**DIF in relation to disability, disease course, and time for evaluation**

We finally evaluated the presence of DIF in relation to disability, disease course, and time for evaluation. As presented in Table 5, the items demonstrating uniform DIF in relation to disability were 1, 2, 4, 6, 8, and 9. Uniform DIF was also present in relation to disease course: items 1, 2, and 8. Finally, items 1, 2 and 6 also demonstrated uniform DIF in relation to time for evaluation.
the impact of DIF and multidimensionality on the generated measures of FSS in this sample, in order to support future valid comparisons between groups.

**Strengths and limitations of the FSS–5**

The psychometric properties of the FSS–5 shown in our study are comparable with findings in the study of Mills et al. However, in both the Norwegian and the Swedish data, item 4 showed misfit. The Rasch analysis in our study showed to some degree contradictory findings when the Rasch factor was extracted in two different FSS–5 versions. In this analysis, item 6, ‘My fatigue prevents sustained physical functioning’, was only included in the Norwegian version of the FSS–5 version and item 3, ‘I am easily fatigued’, was only included in the Swedish FSS–5 version. The FSS–5 also demonstrated uniform DIF to a lesser extent than the FSS–7. The proportion of participants demonstrating maximum or minimum scores on the FSS was, however, larger in the FSS–5 compared with the FSS–7 in both the Norwegian and Swedish data. This fact, in addition to the lower separation index, suggests that the FSS–5 version may be a less sensitive instrument to detect differences/changes between groups and over time, as compared to the FSS–7 version.

**The FSS**

The FSS is probably one of the most frequently used instruments for measuring fatigue in people with MS and other neurological diseases. While several studies have documented satisfactory psychometric properties according to classical test theory, to our knowledge, this is the second study that has studied the psychometric properties using modern test theory on item level and the first to study to assess differences in response pattern over time. Our study contributes confirmatory material but also new information regarding person–response validity and separation.

As this study reveals unique information about changes in fatigue over time, and the results also indicate that the item hierarchy changes over time (item 6), future studies should analyse the response patterns both with regard to changes in individual measures over time, and in relation to changes in relative challenge across items over time. Such studies may generate important clinical information to detect aspects of fatigue impact that could be the target for specified intervention.

In conclusion, the FSS–7 has better psychometric properties than the original FSS–9 in people with MS and is preferable from an internal scale-validity perspective. If the FSS–9 is used, we recommend excluding items 1 and 2 in a raw sum- or mean-score. The FSS–7 might be more sensitive for measuring change in level of fatigue than other versions of the FSS with fewer items. Future studies are needed in order to examine item 3 in relation to item misfit and to ascertain whether the differences in relation to disease course and time for evaluation in relation to items 6 and 8 and level of disability in relation to items 4, 6, 8 and 9 represent substantial characteristics of fatigue in MS and its development over time. Further analysis of country-specific DIF in FSS is also needed, and treated based upon in-depth analysis of causes to those DIF. Since the FSS is frequently used not only on people with MS, there is a need for assessment of its psychometric properties also in other populations.

**Acknowledgements**

The study was funded by grants from the Health Care Sciences Postgraduate School, Karolinska Institutet and from the Swedish Research Council, Buskerud University College, Drammen, Norway and from Odd Fellow Research Fund, Oslo, Norway. Thanks to Elisabeth Gulowsen Celius for giving access to the Oslo MS-Registry.

**References**

11. van de Port I, Kwakkel G, Schepers VP, Heinemans CT, Lindeman E. Is fatigue an independent factor associated with activities of daily living, instrumental activities of...