Original Article

The Neurological Sleep Index: A suite of new sleep scales for multiple sclerosis

RJ Mills, A Tennant and CA Young

Abstract

Objective: The objective of this study was to develop patient-reported outcome measures for sleep dysfunction and sleepiness in multiple sclerosis (MS), since there are currently no MS-specific measurement tools for these clinically important entities.

Methods: Items were generated from semi-structured interviews followed by cognitive debrief. A 42-item pool was administered to patients with MS at three neuroscience centres in the UK. Comparator scales were co-administered. Constructs were validated by Rasch analysis, guided by initial exploratory factor analysis.

Results: There were two supraordinate qualitative themes of diurnal sleepiness and non-restorative nocturnal sleep. Rasch analysis on 722 records produced three scales, which corresponded to diurnal sleepiness, non-restorative nocturnal sleep and fragmented nocturnal sleep. All had excellent fit parameters, were unidimensional and were free from differential item functioning. A summed raw score cut-point of 31/48 in the Diurnal Sleepiness Scale equated to the standard cut-point of 10 on the Epworth Sleepiness Scale (ESS).

Conclusion: Three high-quality measurement scales were developed, and together they compose the Neurological Sleep Index for MS (NSI-MS). The Diurnal Sleepiness Scale might provide an alternative to the ESS. The Non-Restorative Nocturnal Sleep Scale and the Fragmented Nocturnal Sleep Scale appear to be the only such measures for use in MS.

Introduction

There is a high prevalence of sleep disorders in multiple sclerosis (MS), much of which is thought to remain clinically undiagnosed.[1,2] Sleep disturbance in MS is associated with reduced quality of life[3,4] and may be related to MS fatigue;[5,6] indeed, treatment of conditions such as sleep-disordered breathing in MS may improve levels of fatigue.[7,8] Qualitatively, it is clear that there is an intimate relationship between sleep and fatigue in MS,[9] and, quantitatively, sleeping for either too little or too long at night is associated with higher levels of fatigue; those with the lowest levels of fatigue have an average nocturnal sleep duration of 7.5 hours.[10] In addition, those with broken nocturnal sleep have greater fatigue.[10]

Despite the importance of sleep disturbance in MS, there is currently no MS-specific scale for measuring sleepiness or quality of nocturnal sleep. Caution needs to be exercised when instruments that are not validated for specific neurologic diseases are used by patients with MS, an example being a high false-positive rate of restless legs syndrome (RLS) when using the International Restless Legs Syndrome Study Group diagnostic questionnaire.[11] A recent review of sleep disorders in MS and their relationship to fatigue similarly recognised that generic scales may be confounded in MS and recommended the creation of instruments that would be specific to sleep disorders in the context of MS-related fatigue.[12]

During the construction of the Neurological Fatigue Index for MS (NFI-MS), two such sleep-related scales were identified — a relief by diurnal sleep or rest scale and an abnormal nocturnal sleep scale — but it

Correspondence to:
RJ Mills
Roger J. Mills, Lancashire
Teaching Hospitals NHS
Foundation Trust, Dept.
Neurology, Sharoe Green
Lane, Fulwood, Preston,
Lancashire PR2 9HT, UK.
Email:
rjm@crazydiamond.co.uk

RJ Mills
Lancashire Teaching
Hospitals NHS Foundation
Trust, Preston, UK
A Tennant
Swiss Paraplegic Research,
Nottwil, Switzerland

CA Young
The Walton NHS Foundation
Trust, Liverpool, UK
University of Liverpool,
Liverpool, UK

was concluded that further development of these was required.[13]

Objective

This paper takes the development of the two sleep-related patient-reported outcome measures of the NFI further, with the objective of achieving high psychometric validity, including fit to the Rasch measurement model.

Methods

The study had approval from the local research ethics committees. All subjects received written information on the study and gave written informed consent prior to participation.

Qualitative phase

Qualitative data from semi-structured interviews previously performed with 40 patients with MS to explore fatigue, details of which can be found elsewhere,[9] were re-analysed in order to generate new sleep-related scale items, with each theme represented by a small number of items. Standard techniques of content analysis were employed.[14]

The new items were combined with the previously derived NFI-MS sleep-related scales[13] to create a pool of 42 items, each with a common four-point, Likert-style response option[15] of 'strongly disagree', 'disagree', 'agree' and 'strongly agree', with each item being scored 0, 1, 2 or 3. There was a single-sentence instruction at the start of the scale asking respondents to consider their experience over the previous 2 weeks.

Items were put to a multidisciplinary panel of professionals experienced in MS and sleep, comprising MS specialist nurses, MS specialist physiotherapists and occupational therapists, consultants in neurology and neurorehabilitation with a specialist interest in either MS or neurological sleep disorders, and a clinical physiologist in sleep medicine, in order to confirm appropriateness and completeness of the pool.

Cognitive debriefing of the draft scale was performed following face-to-face administration to 10 MS patients in the outpatient clinic. This allowed any gross problems with wording or item dysfunction to be identified and remedied.

Main data phase

A cross-sectional cohort of patients with clinically definite MS[16] was identified from consecutive individual outpatient attendances in three neuroscience centres in the UK (Liverpool, Preston and Manchester). The recruitment was part of a larger, ongoing project aiming to determine trajectories in neurological outcomes, including quality of life, in MS (the TONiC study). The data were collected over the first 12 months of study recruitment. Those recruited for the larger study then received a questionnaire pack containing the set of potential items for the proposed scale, questions on demographics and basic disease information, self-estimated duration of both nocturnal and any diurnal sleep, together with the Epworth Sleepiness Scale (ESS)[17] and Medical Outcomes Study Sleep Scale (MOS-SS)[18] and the Neurological Fatigue Index for MS (NFI-MS),[13] for comparative analysis (see supplemental material for description of comparator measures).

Data from the comparator measures were converted to interval level by Rasch analysis; data from the NFI-MS and the ESS were expected to fit the Rasch model (the NFI-MS was created with Rasch validation and the ESS subjected to unpublished analysis by the authors on previous data);[19] the MOS required *de novo* analysis.

Participants of any age, disease type and disability level were included (the range of Expanded Disability Status Scale (EDSS) scores[20] was 0—9.0 as rated by neurologists at the time of database enrolment).

Retesting was performed at 2—4 weeks. The scale was re-administered with a global indicator of change question to determine whether symptoms had been worse, unchanged or better.

In order to corroborate the qualitative themes and inform the psychometric analysis, the responses to items representative of each theme were re-scored 0, 0, 1, 1, thus dichotomising the data into disagree/agree. The prevalence of each theme in the sample could then be calculated as the percentage endorsing each item.

Data were transcribed to a computer database (transcription error based on checking a random 10% sample was <0.1%).

Psychometric analysis/item reduction

An exploratory factor analysis was undertaken to determine any multidimensionality within the item pool which might inform item groups taken forward to Rasch analysis. Valid scales were identified by fit of data to the Rasch measurement model. Briefly, the

process of Rasch analysis is concerned with whether or not the data meet the model expectations, and provides an assessment of the suitability of the response scale, the fit of individual items, item bias, local dependency and the dimensionality of the scale as a whole. Details of the exploratory factor analysis technique and the Rasch fit criteria can be found in the supplemental material.

For Rasch analysis, a sample size of 243 will provide accurate estimates of item and person locations irrespective of the scale targeting.[21] Assuming a minimum 50% response rate from the recruited subjects, the expected sample size would allow the data to be split randomly into two equal samples of approximately $n = 350$, one data set for initial evaluation and the second set aside in order to validate the results.

External comparison
For each new scale, linear correlation of the Rasch-derived interval-level person estimates with the comparator measures, which had also been transformed to interval scaling by Rasch analysis, was performed. All correlations were expected to be moderate (rho 0.4–0.7) in size. Correlation with estimates of nocturnal and diurnal sleep duration was also performed, and other relationships with features of sleep dysfunction were investigated.

Test–retest reliability
The test–retest reliability of scales was examined by differential item functioning (DIF) by time on all who completed the retest, and by concordance correlation coefficient (CCC)[22,23] using only those respondents who indicated no change on the global indicator of change question. CCC was rated using the Landis and Koch benchmarks: 0 to 0.20 'slight', 0.21 to 0.4 'fair', 0.41 to 0.60 'moderate', 0.61 to 0.8 'substantial', and above 0.81 'almost perfect' consistency or conformity.[24]

Raw score to interval scale conversion
Given fit to the Rasch model, a straightforward conversion is available between the raw score for each scale and the interval scale estimate provided by the model (the person location), in logits. The logit estimates are converted to the same range as the raw score by a further simple linear transformation. This nomogram can be used to obtain linear estimates from the raw scores of other samples only when their data are complete.

Results

Sample characteristics

Qualitative phase. Forty subjects were interviewed, of whom 32 (80%) were female, mean age was 49.0 years (SD 9.7, range 34–78), and mean disease duration was 16.3 years (SD 9.4, range 1–41); a full range of disease phenotypes and a wide range of EDSS scores (0–9.0) were represented.

Main data sample. From an initial data census for the TONiC study on 31 October 2014, 722 packs were returned; 519 (72.5%) were female. Mean age was 48.9 years (SD 11.6, range 17–82); 71 (9.8%) had primary progressive disease, 40 (5.5%) rapidly evolving MS, 418 (57.9%) relapsing–remitting and 161 (22.3%) secondary progressive disease; 32 (4.4%) had unknown disease type. The mean duration of MS was 11.5 years (SD 9.1, range 0–49). There was a wide range of EDSS scores (0–9.0). Seventy-five subjects completed the retest at 2–4 weeks and were examined for DIF by time; 43 of these reported no change in sleepiness and were used in the CCC analysis.

Qualitative analysis
Two clear supraordinate sleep-related themes emerged, concerning diurnal sleepiness and the unrefreshing or non-restorative nature of nocturnal sleep, which are expanded in the supplemental material; the prevalence of these themes in the main data sample can also be found there (Table S1).

Scale items. Items were generated to represent the qualitative themes; wording was selected to be simple and concise, reflecting as much as possible phraseology from the patients' narratives. None were worded for reversed scoring. Some example items were: 'I sometimes simply cannot control the urge to sleep in the day', 'My energy levels drop and I find my eyelids closing', 'Tiredness suddenly hits me like a wave', 'I get a feeling as if I've had too many late nights.'

Review panel and cognitive debriefing. All items were confirmed as being reasonable by the review panel and appeared to be easily understood at cognitive debrief.

Psychometric analyses

Factor analysis. Bartlett's Test of Sphericity was highly significant ($p < .001$) and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy value was 0.962, both supporting the factorability of

Table 1. Summary fit statistics for Rasch analyses.

Analysis	sample	number of items	Item Residual Mean	SD	Person Residual Mean	SD	Chi-Square Value	p	PSI	CA	Unidimensional t-test (lower CI)	extreme scores
1) Factor 1 setup	exp (n = 357)	17	−0.500	3.330	−0.537	1.763	286.9	<0.001	0.923	0.934	8.71 (5.78)%	1.4%
2) Diurnal first scale	exp (n = 356)	9	−0.229	1.124	−0.409	1.149	53.3	0.024	0.900	0.895	6.41 (3.87)%	2.8%
3) Diurnal discard set	exp (n = 356)	7	0.025	1.737	−0.458	1.309	43.4	0.156	0.748	0.758	7.30 (4.60)%	1.4%
4) Diurnal final	exp (n = 349)	16	−0.189	2.137	−0.726	1.021	2.1	0.995	0.913	0.837	7.30 (4.30)%	1.1%
5) Factors 2&3 setup	exp (n = 357)	20	0.040	2.421	−0.565	2.039	212.7	<0.001	0.930	0.938	5.27 (2.96)%	0.8%
6) Nocturnal first scale	exp (n = 355)	8	−0.138	0.927	−0.623	1.459	21.9	0.857	0.860	0.879	4.99%	4.8%
7) Nocturnal discard set	exp (n = 357)	7	−0.202	1.593	−0.461	1.224	28.8	0.759	0.835	0.851	6.70 (4.11)%	2.5%
8) Nocturnal final	exp (n = 348)	15	0.028	0.968	−0.739	1.087	1.5	0.999	0.923	0.929	5.86 (3.43)%	1.4%
9) Fragmented setup	exp (n = 356)	5	0.017	2.891	−0.477	1.224	70.6	<0.001	0.807	0.850	5.04 (2.76)%	6.2%
10) Fragmented final	exp (n = 356)	4	0.069	0.190	−0.519	1.151	20.8	0.410	0.805	0.867	3.06%	12.1%
11) Factor 4 setup	exp (n = 357)	5	0.174	1.534	−0.558	1.310	48.1	0.003	0.598	0.621	4.74%	2.5%
12) Diurnal final	val (n = 350)	16	0.018	1.646	−0.682	1.001	2.6	0.990	0.923	0.872	5.88 (3.44)%	2.6%
13) Nocturnal final	val (n = 347)	15	0.172	0.728	−0.774	1.212	1.5	0.999	0.937	9.422	4.12%	3.4%
14) Fragmented final	val (n = 352)	4	−0.130	0.406	−0.515	1.129	24.4	0.225	0.817	0.871	3.08%	9.9%
15) ESS	val (n = 350)	8	−0.487	1.421	−0.334	0.914	43.2	0.337	0.841	0.869	4.20%	5.4%
16) NFI Summary	val (n = 351)	10	−0.426	1.349	−0.444	1.304	59.5	0.169	0.909	0.939	7.30 (4.60)%	6.1%
17) NFI Physical	val (n = 351)	8	−0.671	1.432	−0.510	1.343	55.6	0.052	0.884	0.934	3.42%	12.8%
18) NFI Cognitive	val (n = 351)	4	0.236	1.108	−0.463	1.114	27.3	0.126	0.834	0.895	5.90 (3.40)%	16.0%
19) MOS final	exp (n = 353)	4	−0.063	1.167	−0.403	1.040	22.1	0.334	0.730	0.762	5.00 (2.80)%	2.3%
20) MOS final	val (n = 350)	4	0.394	0.600	−0.401	1.089	22.0	0.338	0.763	0.801	3.40%	4.3%
Acceptable Values			**0**	**<1.4**	**0**	**<1.4**		**>0.05** bonferroni corrected	**>0.85**	**>0.85**	**<5.0%**	

Exp: exploratory sample, val: validating sample, PSI: person separation index, CA: Cronbach's alpha.

the matrix. Principal component analysis (PCA) with Promax rotation for related factors revealed four potential subscales from the 42-item set, which was also supported by parallel analysis.

The pattern matrix can be found in the supplemental material (Table S2). Seventeen items loaded to the first factor, interpreted as diurnal sleepiness; 15 items to the second factor, non-restorative nocturnal sleep; and five items loaded to the third factor, relating to fragmented nocturnal sleep. The final factor comprised five items with negative loadings ('I have spells of unsatisfying yawning', 'I yawn a lot', 'I have to go to bed really early', 'If I sleep in the day, I don't sleep well at night', 'I'm afraid to fall asleep in the day',) and for this reason was determined to be separate from the underlying latent trait.

Rasch analysis. The main sample was split randomly into two, making an 'exploratory' and a 'validation' sample. Comparison of these samples by *t*-test or chi-square test revealed no significant differences between age, disease duration, disease type, EDSS level, or marital and employment status.

Data in the evaluation sample, in item groups based on the pattern matrix, were then fitted to the Rasch measurement model. An iterative process of item reduction involved identifying disordered thresholds, DIF, item misfit and breaches of local dependency, including multidimensionality. The summary statistical findings related to the salient analyses are given in Table 1.

Diurnal sleepiness. Rasch analysis of the 17 diurnal sleepiness items (factor 1) indicated that all item thresholds were ordered, suggesting that respondents could properly discriminate between response options. However, overall chi-square was highly significant, and several items displayed misfit; there was multidimensionality, with 8.7% (confidence interval (CI) 5.8−11.6%) of *t*-tests indicating significantly different person estimates derived from different subsets of items (Table 1, analysis 1). An iterative process led to a scale reduction to nine items. This scale had satisfactory fit with absence of local dependency and DIF by age, sex, disability level and disease type, but failed the post hoc *t*-test for unidimensionality. Testlets were created for item pairs 35 and 37, and 51 and 53 (respectively,

the most negative and positive items loading to the first factor in a PCA of the item residuals); only 1% of the total scale variance was shed, but this was enough to render the scale unidimensional (Table 1, analysis 2). The discarded eight items were then analysed; deleting one of these items because of high negative fit residual, indicating redundancy (item 14, 'I need to sleep in the day'), resulted in another seven-item scale with satisfactory fit (Table 1, analysis 3). The 16 items were then re-analysed; there were low levels of local dependency with rho of 0.2–0.4 between eight item pairs which spanned the two scale groups. Testlets were therefore made of the two scale groups, and this resulted in a unidimensional scale with excellent fit, having shed 3% of the item variance in the testlet structure (Table 1, analysis 4).

Non-restorative nocturnal sleep. The 20 items from factors 2 and 3 were initially analysed together. All thresholds were ordered, but the scale failed to meet the model expectations, with highly significant overall chi-square and misfitting items, although the post hoc *t*-test revealed unidimensionality at 5.3% (CI 3.0–7.6%) (Table 1, analysis 5). It became clear that the factor 3 items were being consistently excluded, and so factor 2 items were analysed in isolation. Misfitting items were deleted until an eight-item scale was found which had satisfactory fit, was free of local dependency and DIF, and remained unidimensional (Table 1, analysis 6). The seven discarded items were analysed, and these also formed a scale with satisfactory fit (Table 1, analysis 7). The 15 items were then re-analysed; there were low levels of local dependency, with rho of 0.2–0.3 between item pairs, which spanned the two scale groups. Testlets were therefore made of the two scale groups, and this resulted in a unidimensional scale with excellent fit, without losing any item variance in the testlet structure (Table 1, analysis 8).

Fragmented nocturnal sleep. The five items of factor 3 were analysed in isolation. There was one grossly misfitting item (item 31, 'my legs keep me awake at night') causing overall scale-misfit (Table 1, analysis 9). Once this item was removed, the remaining four items produced a unidimensional scale free from local dependency and DIF (Table 1, analysis 10).

Factor 4. Analysis of the five items of factor 4 revealed extensive misfit with very low reliability indices (Table 1, analysis 11), which could not be recovered, and hence the scale was abandoned.

Validation data. Data from the validation sample for the derived scales were then applied to the Rasch model. The Diurnal Sleepiness Scale (DSS) and Non-Restorative Nocturnal Sleep Scale (NRNSS) satisfied all the fit criteria in their two-testlet structures (Table 1, analyses 12 and 13). The Fragmented Nocturnal Sleep Scale (FNSS) also demonstrated good fit without requiring any further modification (Table 1, analysis 14).

Test–retest reliability. Retesting was performed between 2 and 4 weeks. The invariance of the item difficulty over time was confirmed by the absence of DIF by time. CCC for the DSS was 0.813 (95% CI 0.688–0.891); for the NRNSS it was 0.821 (95% CI 0.694–0.898), implying 'almost perfect' concordance; and for the FNSS it was 0.786 (95% CI 0.638–0.878), implying 'substantial' concordance for each scale.

External construct validity. Unmodified data from the ESS and NFI-MS satisfied the Rasch fit criteria (Table 1, analyses 15–18). In order to achieve satisfactory fit for data from the MOS, two items (trouble staying awake during the day *and* awaking with breathlessness/headache) required removal. This resulted in a scale with face validity for a latent trait of nocturnal sleep quality (original factors of sleep adequacy and sleep disturbance); the final four-item scale had ordered response thresholds, was unidimensional and was free from DIF (Table 1, analyses 19 and 20).

The linear correlations between the sleep scales and comparator measures are shown in Table 2. There was moderate correlation between the DSS, the ESS and hours of day sleep. The NRNSS and the FNSS correlated moderately with the MOS. NRNSS scores increased directly with sleep latency (Figure 1). There was no linear relationship between duration of nocturnal sleep and the NRNSS, but a line plot revealed a U-shaped relationship, with a nadir in non-restorative sleep (NRS) at a sleep duration of 7.5 hours (Figure 2). There was no correlation between any of the scales and either subject age or disease duration.

The ESS was equated to the DSS; the standard summed raw score cut-point for pathological sleepiness of 10 on the ESS equated to a summed raw score of 31 on the DSS (Figure 3). This meant that 26.6% of respondents in the study sample had pathological diurnal sleepiness.

Table 2. External comparison by Pearson rho (Spearman's rho for duration of sleep) ($n = 708$).

	Diurnal sleepiness	Non-restorative nocturnal sleep	Fragmented nocturnal sleep	ESS	MOS
Diurnal sleepiness	—	0.716	0.459	0.619	0.405
Non-restorative nocturnal sleep	0.716	—	0.581	0.327	0.605
Nocturnal sleep duration	−0.063	−0.109	−0.398	−0.084	−0.477
Diurnal sleep duration	0.587	0.327	0.192	0.562	0.160
ESS	0.619	0.398	0.283	—	0.270
MOS	0.405	0.605	0.682	0.270	—
NFI summary	0.769	0.668	0.434	0.427	0.395
NFI physical	0.740	0.629	0.411	0.404	0.374
NFI cognitive	0.704	0.644	0.421	0.395	0.389

ESS: Epworth Sleepiness Scale; MOS: Medical Outcomes Scale Sleep Scale; NFI: Neurological Fatigue Index.
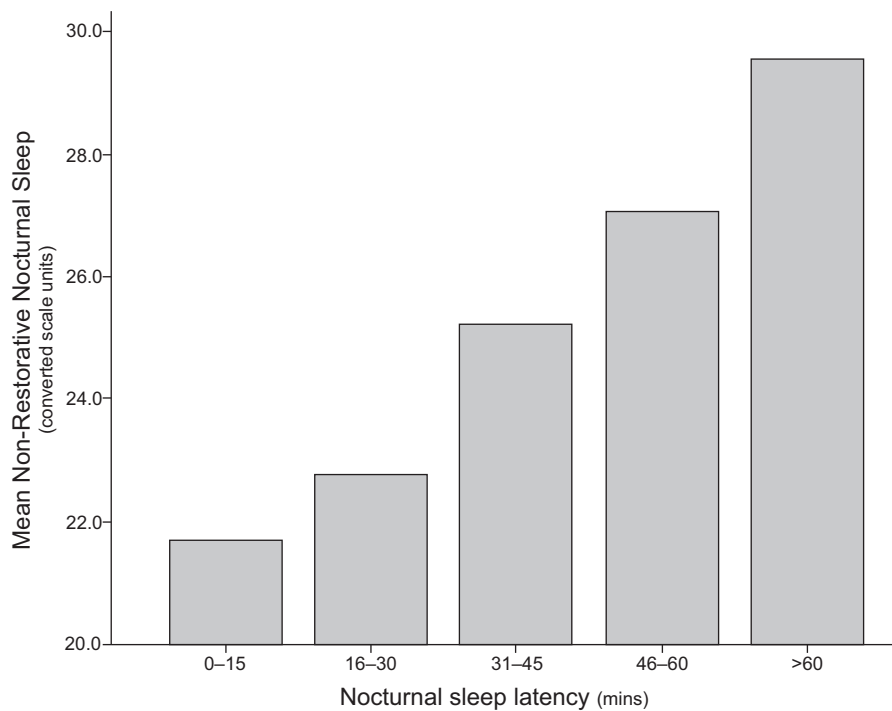
Figure 1. Mean NRNSS vs. nocturnal sleep latency.

Raw score to interval scale conversion. Given fit to the Rasch model, a simple conversion of the raw score for each scale to its interval scale equivalent is provided in Table 3.

Discussion

A suite of short, simple, patient-based, MS-specific self-report scales was developed which measured diurnal sleepiness, non-restorative nocturnal sleep and fragmented nocturnal sleep. The scales had a clear conceptual and qualitative basis and satisfied the rigorous psychometric standards of Rasch analysis in accordance with the US Food and Drug Administration's guidelines for the development of patient-reported outcome measures.[25] The scales provided valid measurement for patients of any age, sex, disease type, and disability level. Collectively, they form the Neurological Sleep Index for MS (NSI-MS).
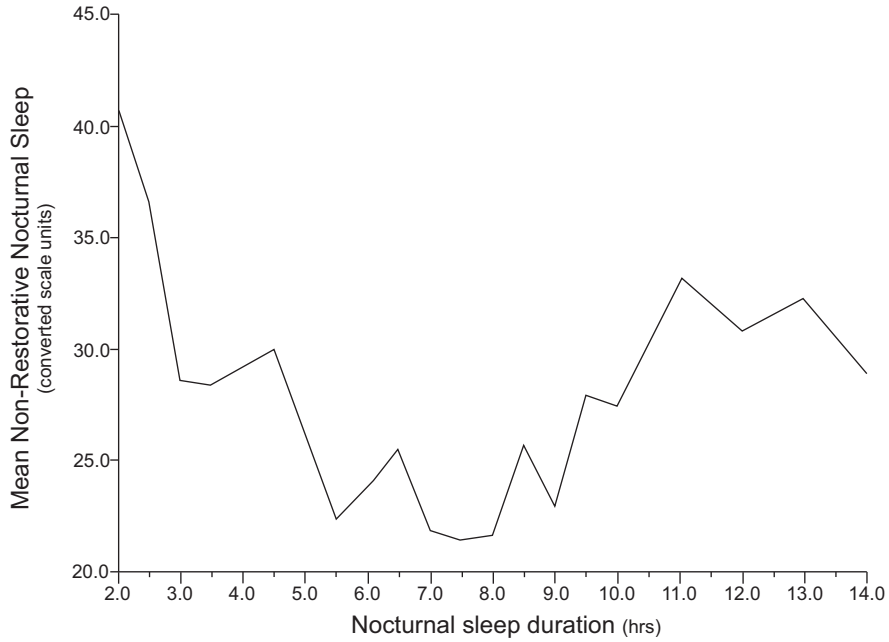
Figure 2. Mean NRNSS vs. nocturnal sleep duration.
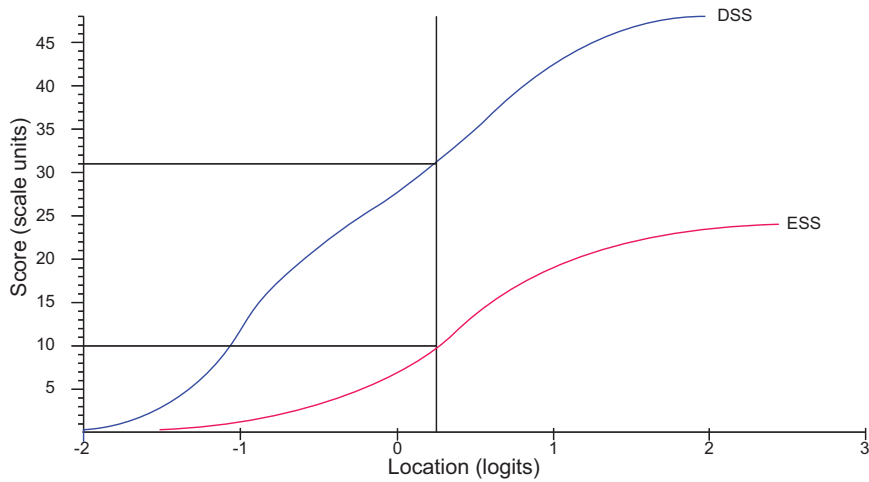
Figure 3. Raw score vs. person location for the Epworth Sleepiness Scale (ESS) and Diurnal Sleepiness Scale (DSS). An ESS summed raw score of 10 equates to a summed raw score of 31 (or 28.67 using the transformed interval score) on the DSS.

The first scale concerned diurnal sleepiness. It might be seen as a direct replacement for the ESS, and has the potential advantage of not being tied to particular situations which may, or may not, be relevant to patients with MS. There is a degree of convergence between diurnal sleepiness and fatigue, but the two entities are considered distinct,[26,27] and it is important to be able to measure this separate variable as a potentially modifying, and modifiable, factor.

The second scale measured non-restorative nocturnal sleep. NRS is an established concept that concerns the sense of being un-refreshed upon waking and a wider sense that nocturnal sleep has been inadequate despite opportunity to sleep, although it is acknowledged that there is no universally accepted definition of NRS in the literature.[28] There is also debate as to whether NRS should be regarded as integral to the concept of insomnia, in addition to the problems of

Table 3. Nomogram of summed raw scores to interval level conversion. The conversions remain valid provided there are no missing data.

summed raw score	Diurnal Sleepiness Scale	Non-restorative Nocturnal Sleep Scale	Fragmented nocturnal sleep scale
0	0.00	0.00	0.00
1	2.84	2.25	1.18
2	4.74	3.80	2.19
3	6.01	4.87	3.06
4	7.00	5.74	3.92
5	7.84	6.51	4.82
6	8.57	7.22	5.77
7	9.26	7.91	6.74
8	9.89	8.60	7.71
9	10.50	9.31	8.66
10	11.11	10.03	9.62
11	11.72	10.78	10.73
12	12.34	11.55	12.00
13	12.98	12.35	
14	13.66	13.18	
15	14.37	14.02	
16	15.13	14.88	
17	15.92	15.76	
18	16.74	16.65	
19	17.59	17.57	
20	18.47	18.49	
21	19.36	19.42	
22	20.28	20.36	
23	21.20	21.31	
24	22.13	22.26	
25	23.07	23.22	
26	24.01	24.19	
27	24.95	25.16	
28	25.89	26.12	
29	26.83	27.09	
30	27.75	28.06	
31	28.67	29.01	
32	29.57	29.97	
33	30.44	30.92	
34	31.30	31.86	
35	32.13	32.78	
36	32.93	33.70	
37	33.70	34.59	
38	34.46	35.47	
39	35.21	36.35	
40	35.98	37.25	
41	36.77	38.20	
42	37.61	39.26	
43	38.53	40.55	
44	39.57	42.38	
45	40.79	45.00	
46	42.33		
47	44.62		
48	48.00		

initiating and maintaining sleep.[28–30] Whatever the nosological dilemma, the current results (see Figure 1) reveal a direct relationship between NRS and sleep latency, as might be expected.[31] There was also a non-linear relationship between NRS and sleep duration, and these well-formed relationships with basic sleep parameters suggest that measurement using the NRNSS can be meaningful.

The final scale provided a measure of sleep fragmentation. There do not appear to be any other existing scales dedicated to measuring this phenomenon in neurologic disease. Fragmented sleep had an indirect, linear correlation with nocturnal sleep duration, which was different from the relationship between NRS and nocturnal sleep duration, suggesting that the FNSS and NRNSS were measuring different entities; however, there is thought to be causation and effect between sleep fragmentation and NRS.[32] The other notable feature of the scale was the necessary omission of the item relating to lower limb symptoms, which would encompass RLS, perhaps suggesting that RLS is separate again from any of the sleep-related traits currently presented.

The exploratory factor analysis revealed one group of items (factor 4) which could not be reconciled. These items probably represented consequences of, or adaptive behaviour to, the other supraordinate themes, and were unlikely to be of importance for clinical measurement in the current context.

Methodology

There was an attempt to make the development of the scales, from qualitative phase to decisions for item selection, as transparent as possible, since the validity of any scale rests not only simply on a series of fit statistics, but also on the scale's ability to measure the desired latent trait. There has been recent debate that Rasch analysis is too restrictive and drives scale construction along lines of fit criteria at the expense of conceptual utility of items,[33] and, for example, that statistical criteria are used alone in order to determine unidimensionality.[34,35] We would argue that any scale has to be underpinned by a clearly defined latent trait of clinical relevance, but measurement has fundamental mathematical requirements, and the *only* way that such requirements can be realised is by application of the Rasch model. A scale which does not meet the Rasch model's requirements is simply a symptom inventory, which may yield qualitative information but cannot be used for measurement. In the same way that statistical criteria are not a substitute or surrogate for conceptual unidimensionality, a conceptually unidimensional scale is meaningless for measurement unless the mathematical measurement requirements are also satisfied.

The point that both conceptual and mathematical unidimensionality are necessary and inextricable is perhaps illustrated in this study with the analysis of the items comprised in factors 2 and 3. The initial qualitative interpretation was that themes of sleep fragmentation would be part of NRS, and hence they were analysed together, overriding the exploratory factor analysis. However, the Rasch analysis could not be resolved in this configuration, and it was clear that NRS was distinct from fragmented sleep, which, in retrospect, was felt to be entirely reasonable as a clinical concept. It should be remembered that Rasch analysis is based on probabilities provided by respondents, so that the stochastic realisation of measurement is driven by the patients with MS themselves, and measurement instruments derived in this way should not be easily dismissed.

Validation of any new scale is an ongoing process involving reproduction of construct validity in samples larger than the relatively small sample used in this study, the capture of longitudinal and clinical trial data, and the determination of the minimal clinically important difference (MCID). This process may also establish the clinical usefulness of the instruments. The initial validation presented here was also perhaps limited by the absence of sleep laboratory investigations such as multiple sleep latency tests (MSLT), maintenance of wakefulness tests (MWT) and polysomnography for external comparison or the corroboration of cut-points for sleepiness and the demonstration of sleep pathology. Such comparison may be the subject of future work. However, the MSLT and MWT should not necessarily be seen as superior or more objective measures of sleepiness or the non-restorative nature of nocturnal sleep, and may have their own limitations.[36–38]

The NSI-MS is free for use by all state-funded health-care organisations and not-for-profit agencies, and can be obtained, after appropriate registration, from the psychometric laboratory at Leeds University (http://medhealth.leeds.ac.uk/info/732/psychometric_laboratory/1493/scales) or by contacting the authors.

Conclusion

Development of the NSI-MS corroborated previous findings that there are latent traits relating to diurnal sleepiness and non-restorative quality of nocturnal sleep that are meaningful to patients with MS, as well as introducing an instrument to measure nocturnal sleep fragmentation. The NSI-MS was shown to fit the Rasch model and therefore measures

unidimensional constructs and generates interval-level data. The DSS does not contain situation-specific items and might therefore provide an alternative to the ESS. The NRNSS and the FNSS appear to be the only such measures for use in MS. It is intended that the suite of scales will allow sophisticated interrogation of the relationships between sleep dysfunction and other clinical features of MS at both an individual and a population-based level.

References

1. Bamer AM, Johnson KL, Amtmann D, et al. Prevalence of sleep problems in individuals with multiple sclerosis. *Mult Scler* 2008; 14: 1127–1130.

2. Brass SD, Li C-S and Auerbach S. The underdiagnosis of sleep disorders in patients with multiple sclerosis. *J Clin Sleep Med* 2014; 10: 1025–1031.

3. Leonavicius R and Adomaitiene V. Features of sleep disturbances in multiple sclerosis patients. *Psychiatr Danub* 2014; 26: 249–255.

4. Sarraf P, Azizi S, Moghaddasi AN, et al. Relationship between sleep quality and quality of life in patients with multiple sclerosis. *Int J Prev Med* 2014; 5: 1582–1586.

5. Kaminska M, Kimoff RJ, Benedetti A, et al. Obstructive sleep apnea is associated with fatigue in multiple sclerosis. *Mult Scler* 2012; 18: 1159–1169.

6. Cameron MH, Peterson V, Boudreau EA, et al. Fatigue is associated with poor sleep in people with multiple sclerosis and cognitive impairment. *Mult Scler Int* 2014; 2014: 872732.

7. Kallweit U, Baumann CR, Harzheim M, et al. Fatigue and sleep-disordered breathing in multiple sclerosis: A clinically relevant association? *Mult Scler Int* 2013; 2013: 286581.

8. Veauthier C, Gaede G, Radbruch H, et al. Treatment of sleep disorders may improve fatigue in multiple sclerosis. *Clin Neurol Neurosurg* 2013; 115: 1826–1830.

9. Mills RJ and Young CA. A medical definition of fatigue in multiple sclerosis. *QJM* 2007; 101: 49–60.

10. Mills RJ and Young CA. The relationship between fatigue and other clinical features of multiple sclerosis. *Mult Scler* 2011; 17: 604–612.

11. Mery V, Kimoff RJ, Suarez I, et al. High false-positive rate of questionnaire-based restless legs syndrome diagnosis in multiple sclerosis. *Sleep Med* 2015; 16: 877–882.

12. Veauthier C and Paul F. Sleep disorders in multiple sclerosis and their relationship to fatigue. *Sleep Med* 2014; 15: 5–14.

13. Mills RJ, Young CA, Pallant JF, et al. Development of a patient reported outcome scale for fatigue in multiple sclerosis: The Neurological Fatigue Index (NFI-MS). *Health Qual Life Outcomes* 2010; 8: 22–28.

14. Pope C, Ziebland S and Mays N. Qualitative research in health care. Analysing qualitative data. *BMJ* 2000; 320: 114–116.

15. Likert R. A technique for the measurement of attitudes. *Archives of Psychology* 1932; 22: 1–55.

16. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the 'McDonald Criteria'. *Ann Neurol* 2005; 58: 840–846.

17. Johns M. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep* 1991; 14: 540–545.

18. Stewart AL and Ware JE (eds). *Measuring functioning and well-being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press, 1992, pp. 235–259.

19. Mills RJ and Young CA. Is the Epworth Sleepiness Scale valid in multiple sclerosis? *Mult Scler J* 2012; 18(S4): 355.

20. Kurtzke J. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology* 1983; 33: 1444–1452.

21. Linacre J. Sample size and item calibration stability. *Rasch Measurement Transactions* 1994; 7: 28.

22. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; 45: 255.

23. Lin LI-K. Corrections. *Biometrics* 2000; 56: 324–325.

24. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

25. Food and Drug Administration. Guidance for industry patient-reported outcome measures: Use in medical product development to support labeling claims. *FDA* 2009; 1–39.

26. Neu D, Mairesse O, Hoffmann G, et al. Do 'sleepy' and 'tired' go together? Rasch analysis of the relationships between sleepiness, fatigue and nonrestorative sleep complaints in a nonclinical population sample. *Neuroepidemiology* 2010; 35: 1–11.

27. Hossain JL, Ahmad P, Reinish LW, et al. Subjective fatigue and subjective sleepiness: Two independent consequences of sleep disorders? *J Sleep Res* 2005; 14: 245–253.

28. Stone KC, Taylor DJ, McCrae CS, et al. Nonrestorative sleep. *Sleep Med Rev* 2008; 12: 275–288.

29. Wilkinson K and Shapiro C. Nonrestorative sleep: Symptom or unique diagnostic entity? *Sleep Med* 2012; 13: 561–569.

30. Zhang J, Lamers F, Hickie IB, et al. Differentiating nonrestorative sleep from nocturnal insomnia symptoms: Demographic, clinical, inflammatory, and functional correlates. *Sleep* 2013; 36: 671–679.

31. Ohayon MM and Hong S-C. Prevalence of insomnia and associated factors in South Korea. *J Psychosom Res* 2002; 53: 593–600.

32. Stepanski EJ. The effect of sleep fragmentation on daytime function. *Sleep* 2002; 25: 268–276.

33. Stefan J and Cano JCH. The problem with health measurement. *Patient Prefer Adherence* 2011; 5: 279–290.

34. Hagell P. Testing unidimensionality using the PCA/t-test protocol with the Rasch model: A cautionary note. *Rasch Measurement Transactions* 2015; 28: 1487–1489.

35. Cano SJ, Barrett LE, Zajicek JP, et al. Dimensionality is a relative concept. *Mult Scler* 2011; 17: 893–894.

36. Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: Failure of the MSLT as a gold standard. *J Sleep Res* 2000; 9: 5–11.

37. Mignot E, Lin L, Finn L, et al. Correlates of sleep-onset REM periods during the Multiple Sleep Latency Test in community adults. *Brain* 2006; 129(Pt 6): 1609–1623.

38. Bonnet MH. ACNS clinical controversy: MSLT and MWT have limited clinical utility. *J Clin Neurophysiol* 2006; 23: 50–58.