Methods in Neuroepidemiology

Neuro
epidemiology

Neuroepidemiology 2010;35:1–11
DOI: 10.1159/000301714

Received: September 23, 2009
Accepted: January 29, 2010
Published online: March 25, 2010

Do 'Sleepy' and 'Tired' Go Together? Rasch Analysis of the Relationships between Sleepiness, Fatigue and Nonrestorative Sleep Complaints in a Nonclinical Population Sample

Daniel Neu[a] Olivier Mairesse[d] Guy Hoffmann[a] Jean-Baptiste Valsamis[b]
Paul Verbanck[a] Paul Linkowski[c] Olivier Le Bon[a, e]

[a]Sleep Laboratory and Unit for Chronobiology U78, Brugmann University Hospital, Université Libre de Bruxelles, [b]Department of Bio Electro and Mechanical Systems (BEAMS), Faculty of Applied Sciences, Université Libre de Bruxelles, [c]Department of Psychiatry, Erasme Hospital, University Clinics of Brussels, Université Libre de Bruxelles, and [d]MOBI research group, Faculty of Economic, Social and Political Sciences, Vrije Universiteit Brussel, Brussels, and [e]Department of Psychiatry, Tivoli University Hospital, Université Libre de Bruxelles, La Louvière, Belgium

Key Words
Fatigue · Sleepiness · Nonrestorative sleep · Rasch analysis

Abstract
Objective: The lack of distinction in the clinical use of terms like fatigue and sleepiness is an important issue. While both fatigue and sleepiness can potentially be associated with nonrestorative sleep (NRS) complaints, their relationships are still poorly described. We propose to use Rasch analysis-based methods to study the interrelations of fatigue, sleepiness and NRS. ***Methods:*** 150 subjects (mean age = 39.3 years, range = 18–65) from a community sample underwent a structured computer-assisted web interview. We assessed demographic data, sleep habits, and subjective fatigue with the Fatigue Severity Scale (FSS), global and situational sleepiness with the Epworth Sleepiness (ESS) and the Stanford Sleepiness Scales, respectively, and affective symptoms with the Hospital Anxiety and Depression Scale. Dimensionality, measurement invariance and common person equating were investigated to study the FSS, ESS and their relations to NRS. ***Results:*** NRS was linked to shorter habitual sleep duration and to higher scores on psychometric scales. Both sleepiness and daytime fatigue were positively correlated to each other and to the intensity of affective symptoms. Rasch analyses showed both the ESS and FSS to measure unidimensional concepts of sleepiness and fatigue, respectively. In contrast to the FSS, the ESS only showed partial invariance to an NRS complaint. Common person equating suggests that, despite similar Rasch-derived agreeability scores, fatigue and sleepiness (as measured by the FSS and ESS) nevertheless designate distinct constructs. ***Conclusion:*** NRS complaints can simultaneously present with higher daytime fatigue and sleepiness levels but the associative relationships between fatigue and sleepiness remain relatively unaffected by NRS. Although participants might not present adequate differentiation, fatigue and sleepiness seem to relate to different underlying concepts.

Introduction

Complaints of fatigue and nonrestorative sleep (NRS) are very frequent in the general population [1, 2] and in primary care [3]. Also, excessive daytime sleepiness is of-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0251–5350/10/0351–0001$26.00/0

Accessible online at:
www.karger.com/ned

Dr. Daniel Neu, MD
Sleep Laboratory and Unit for Chronobiology U78, Department of Psychiatry
Brugmann University Hospital, Université Libre de Bruxelles
Building Hh, Arthur Van Gehuchten Square, BE–1020 Brussels (Belgium)
Tel. +32 2 477 25 54, Fax +32 2 477 21 62, E-Mail daniel.neu@chu-brugmann.be

ten reported in general population samples and both fatigue and sleepiness can present heavy burdens for public health care [4–7].

Sleepiness is generally described as a trigger signal for a spontaneous onset of sleep. It is first of all a physiological phenomenon depending on previous sleep and occurring at regular intervals following a circadian rhythm [8, 9]. But in pathological conditions, excessive daytime sleepiness can be irrepressible when associated with narcolepsy or sleep apnea-hypopnea syndrome for instance. Excessive daytime sleepiness is also linked to other sleep disorders such as periodic limb movement disorder, idiopathic hypersomnia or sleep deprivation [5].

On the other hand, fatigue is generally described as a condition in which maintaining of a motor or mental energy level gets more difficult with duration of exercise. Fatigue needs rest not sleep to recover from. Chronic severe daytime fatigue is, for instance, the core symptom in clinical conditions like chronic fatigue syndrome. Nevertheless, even in presumably exclusive fatigue conditions such as chronic fatigue syndrome significantly overlapping subjective sleepiness has recently been described [10].

Although classical psychometric assessment tools like the Epworth Sleepiness (ESS) and the Fatigue Severity Scales (FSS) are available for the assessment of subjective fatigue and sleepiness complaints, the definition and semiological distinction of both concepts often remain difficult for both patients and clinicians [5, 11]. Moreover, although both fatigue and sleepiness can be associated with complaints of NRS [1], their relationships to each other and to sleep remain insufficiently understood. These are important issues as the etiopathogenesis of related clinical conditions and the implications for therapeutic orientation can be very different [5, 12]. Potentially overlapping descriptive features can also lead to imprecise diagnosis and subsequently to inadequate or insufficient treatment attitudes [5, 6, 12]. Shen et al. [13] pointed towards the coexistence of both phenomena and to the fact that both can potentially be related to sleep deprivation. However, only a few previous studies reported associations between sleepiness and fatigue [11, 14–16]. As a contribution to the disentangling of fatigue and sleepiness, Bailes et al. [11] proposed two empirical scales measuring either fatigue or sleepiness. The authors found significant correlations between the ESS and the FSS ranging from 0.16 to 0.42. Subsequently, they developed the Empirical Fatigue and Sleepiness Scales based on a discriminant validity rationale by retaining items not significantly correlated with items of the opposite construct. The authors reported the ability of those scales to identify 'sleepiness which is not fatigue' [11].

Furthermore, a Rasch methodology-based analysis has recently been applied separately to qualify the measurement properties and stable hierarchical item structure of the ESS and the FSS in Parkinson's disease [17, 18]. However, due to the absence of combined explorations, it is presently unclear how fatigue and sleepiness relate to each other and to NRS.

Hence, in order to assess if individuals perceive subjective sleepiness and fatigue in a similar way and if they are related differently to NRS, we propose to use Rasch-based common person equating [19]. Common person equating is conventionally used to test the equivalence among examinee populations and evaluate different tests that measure the same construct, administered to a common group of persons [20]. In the present study, a somewhat different approach is used regarding common person equating. We hypothesize that examinees confounding sleepiness with fatigue or vice versa will have similar Rasch-derived person estimates on the ESS and the FSS or put differently, examinees will perceive fatigue and sleepiness as undifferentiated concepts. Additionally, by linking persons with or without NRS with similar fatigue and sleepiness levels, different response probabilities to specific ESS and FSS items can be investigated (i.e. differential item functioning or DIF) which eventually allows for a substantive examination of the perceptive associations between NRS, sleepiness and fatigue.

Methods

Subjects

From a privately held representative dataset of a general population sample (Dedicated Research® Inc., Brussels, Belgium), 150 subjects in the Belgian community were selected via mass mailing according to predefined selection criteria and sample size. Inclusion criteria were an available e-mail address and age between 18 and 60 years. Exclusion criteria were the self-reported presence of a known sleep disorder or a former hospitalization in a sleep laboratory. Further exclusion criteria mainly were an earlier participation in an epidemiological survey or a similar Internet-based interview.

A structured computer-assisted web survey was administered. Demographic data, social data, lifestyle components, sleep habits and psychometric scales of fatigue, sleepiness and affective symptoms were recorded. Participation in a lottery, with gift certificates of up to 25 EUR, had been proposed at the beginning of the questionnaire. A total of 10 subjects were randomly chosen and received vouchers of 25 EUR each.

Psychometric Assessment

Global Fatigue and Sleepiness

Global fatigue was investigated by means of the FSS and global sleepiness by means of the ESS. The ESS consists of 8 items (described situations) arranged on a 4-point Likert scale ranging from 0 (never doze) to 3 (high chance of dozing during daytime). The summed scores range from 0 to 24 and scores above 10 are commonly interpreted as increased global sleep propensity [21]. The FSS is a self-report instrument to assess levels of fatigue and its effect on daily functioning [22]. The FSS has been used in many studies investigating fatigue in several chronic conditions and in general population samples [23, 24]. It is a 9-item 7-point Likert-type scale ranging from 'completely disagree' (= 1) to 'completely agree' (= 7). Scores are usually reported as 'mean scores' (ranging from 1 to 7) obtained by dividing the total score (ranging from 7 to 63) by 9. The most often proposed cutoff point on mean scores is 4 [25, 26]. Other authors also proposed a cutoff of 5 [23].

Situational Sleepiness

The Stanford Sleepiness Scale (SSS) contains 7 statements describing different levels of current alertness ranging from 1 ('feeling alert and vital') to 7 ('almost in reverie, lost struggle to remain awake') [27]. The patient has to choose the most appropriate description of his usual sleepiness level for several given time points (9 a.m., 1 p.m., 5 p.m., 9 p.m.) during the day. The SSS was answered all at once in the present study.

Affective Symptoms

The Hospital Anxiety and Depression Scale is a self-report rating scale of 14 items on a 4-point Likert scale (with a range from 0 to 3). It is designed to measure the intensity of anxiety and depression (7 items for each subscale) symptoms [28]. The total score is the sum of the 14 items (ranging from 0 to 42), and for each subscale the score is the sum of the respective 7 items (ranging from 0 to 21). The reliability and validity of the Hospital Anxiety and Depression Scale have been tested in a number of international clinical and nonclinical studies [29].

Statistics and Analyses

All variables were compatible with the use of parametric tests. Descriptive between-group comparisons involving continuous data were computed using MANOVA with a single factor (NRS complaint). Exploratory associations between continuous variables were tested using the Pearson product moment correlation coefficient r. Hypothesis tests were two-sided and carried out at a 5% significance level. Trends are marked between brackets at a 10% level. All results are expressed as mean ± standard deviation. All analyses except the Rasch analyses were computed using SPSS 16® (SPSS Inc., Chicago, Ill., USA). The main outcomes from the Rasch analyses were carried out with Winsteps® for Windows (Winsteps 3.68.0; SWREG, USA).

Results

Thirty-eight percent of the total sample reported an NRS complaint. According to usual cutoff points on the FSS (>4) and the ESS (>10), 34% (n = 51) showed a high level of fatigue and 37% (n = 56) showed a high level of sleepiness. None of the explored variables showed sex-related differences.

Table 1. Descriptive comparisons of key variables regarding NRS

Variable	NRS(–) (n = 93)	NRS(+) (n = 57)	F	p	Partial η^2
BMI	24.14 (5.1)	24.11 (5.0)	0.371	n.s.	0.003
HSTw, h	7.11 (1.0)	6.65 (0.9)	6.309	0.013	0.053
HSTwe, h	8.76 (2.3)	8.31 (2.7)	0.895	n.s.	0.008
Work, h/week	36.03 (12.2)	32.0 (13.9)	1.435	n.s.	0.013
ESS score	8.24 (4.1)	9.65 (3.6)	3.632	0.059[a]	0.031
HAD-A score	6.3 (3.1)	8.63 (3.8)	12.602	0.001	0.100
HAD-D score	3.57 (2.8)	5.72 (3.9)	7.161	0.009	0.060
HAD total score	9.87 (5.1)	14.35 (7.1)	11.910	0.001	0.095
FSS score	3.04 (1.4)	4.03 (1.5)	11.553	0.001	0.093
SSS 9 a.m. score	2.07 (1.8)	2.49 (1.5)	0.617	n.s.	0.005
SSS 1 p.m. score	2.24 (1.4)	2.72 (1.3)	4.203	0.043	0.036
SSS 5 p.m. score	2.64 (1.31)	3.07 (1.4)	0.655	n.s.	0.006
SSS 9 p.m. score	3.07 (1.5)	3.74 (1.8)	7.669	0.007	0.064
SSS mean score	2.51 (0.7)	3.00 (0.9)	8.604	0.004	0.071

The values are represented as means with standard deviations in parentheses. HST = Self-reported estimated habitual sleep time in hours during the week (w) or the weekend (we); Work = reported mean working hours per week; HAD = Hospital Anxiety and Depression Scale; HAD-A = total HAD anxiety score; HAD-D = total HAD depression score; SSS mean = average score (from 4 measure points: 9 a.m., 1 p.m., 5 p.m., 9 p.m.) on the SSS.

[a] p value marking a trend.

Descriptive Comparisons of Key Variables Regarding NRS

MANOVA with a single factor (NRS) was performed to compare reported sleep durations, BMI, weekly working hours, fatigue, sleepiness and affective symptoms between two groups regarding the presence or absence of an NRS complaint. It was found highly significant (Pillai's trace = 0.200; F = 3.832; p = 0.001 for subject groups). Table 1 shows the habitual estimated self-reported mean sleep time during the week, but not during the weekend, being significantly shorter in subjects with an NRS complaint [NRS(+), table 1]. Measures of fatigue, anxiety and depression showed significantly higher scores in the NRS(+) group (table 1). Sleepiness also showed higher levels in the NRS(+) group on the SSS (mean score and at 1 p.m. and 9 p.m.) and a trend on the ESS (table 1). Interestingly, reported working hours during the week and BMI did not differ between groups in

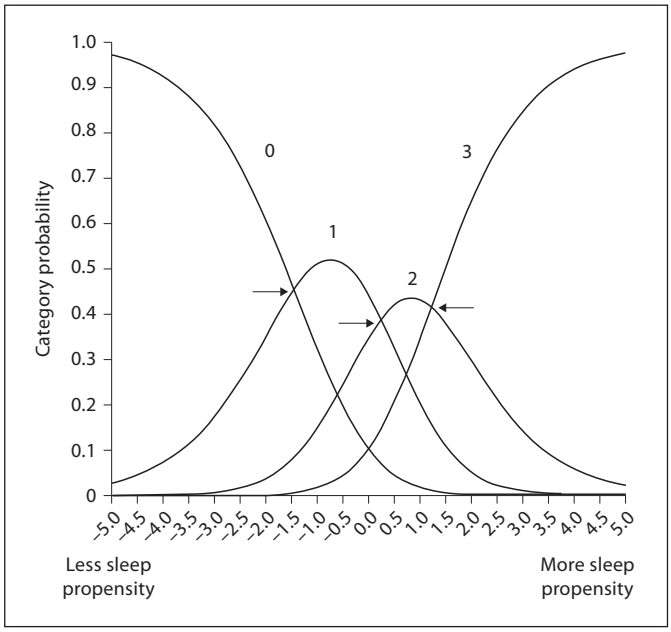


Fig. 1. Rating scale functioning of the ESS. The curves show the probability of observing a category (y-axis, category probability) relative to the items' endorsement rate (x-axis, sleep propensity). The arrows indicate the step calibrations (threshold points where a 50% probability exists of endorsing the adjoining category). Response categories should be ordered as expected and emerge as more and more probable when moving along the sleep propensity continuum, visually represented by an ordered, even succession of hills.

our sample (table 1). Despite statistical significance, comparisons mainly show small and medium effect sizes and except for the FSS, most results display low 'clinical significance' regarding the numerical differences in the absolute scores on psychometric scales between groups (table 1).

Exploratory pairwise correlations (Pearson's product moment r) showed significant relationships not only between affective symptoms (HAD) and fatigue (FSS, $r = 0.555$, $p < 0.001$) but also with sleepiness on the ESS ($r = 0.238$, $p < 0.001$) and the SSS (mean score, $r = 0.499$, $p < 0.001$). Not surprisingly, fatigue and sleepiness were also strongly related to each other (FSS/ESS: $r = 0.488$, $p < 0.001$; FSS/SSS mean: $r = 0.634$, $p < 0.001$). Age was neither correlated to the FSS ($r = -0.98$, $p = 0.141$) nor to the ESS ($r = -0.36$, $p = 0.597$) in the present sample.

Rasch Analysis

The dichotomous Rasch model [30] and its polytomous counterpart, the Andrich Rating Scale Model (ARSM)[1] [31], are members of the unidimensional Item Response Theory (IRT) model family comprising a generalized linear model and the associated statistical procedure that connect the observed responses at the current point where the examinee is on an unmeasured latent trait. The parameters of the probabilistic model represent the position of persons and items on an underlying continuum based on a simple logic: there is a higher probability that respondents endorse 'easier' items compared to 'difficult' ones and there is a higher probability that persons with higher 'abilities' endorse items (or higher categories of items) compared to persons with lower 'abilities' [32]. With respect to the present study, easier (more difficult) items are referred to as items targeting a less (more) intense expression of the underlying variable (fatigue or sleepiness), whereas persons with higher (lower) abilities refer to individuals with a more (less) severe expression of the underlying construct.

Formally, the polytomous ARSM [31] reads as:

$$\log(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_j \quad (1)$$

where:

- P_{nij} is the probability that person n encountering item i is observed in category j,
- B_n is the 'ability' measure of person n, also often referred to as 'agreeability',
- D_i is the 'difficulty' measure of item i, also often referred to as 'endorsability', the point where the highest and lowest categories of the item have equal probabilities,
- F_j is the 'calibration' measure of category j relative to category $j - 1$, the point where categories $j - 1$ and j are equally probable relative to the measure of the item [33].

Rasch measures are 'logits' (log-odd units) that actually measure a difference, a local distance between persons or between items with 0 conventionally assigned to the average endorsability of the items or average person abilities [34]. Put differently, each original observation is modeled to be an amount or a count of qualitative levels of 'performance' (e.g. sleepiness or fatigue). Each person's raw score is modeled to be an accumulation of qualitative levels of performance on the entire test. Then, from these raw score amounts, the location on the latent variable is estimated.

[1] Both terms describing analyses based on the dichotomous Rasch model or the polytomous ASRM are often used in an interchangeable manner and referred to as 'Rasch analysis'.

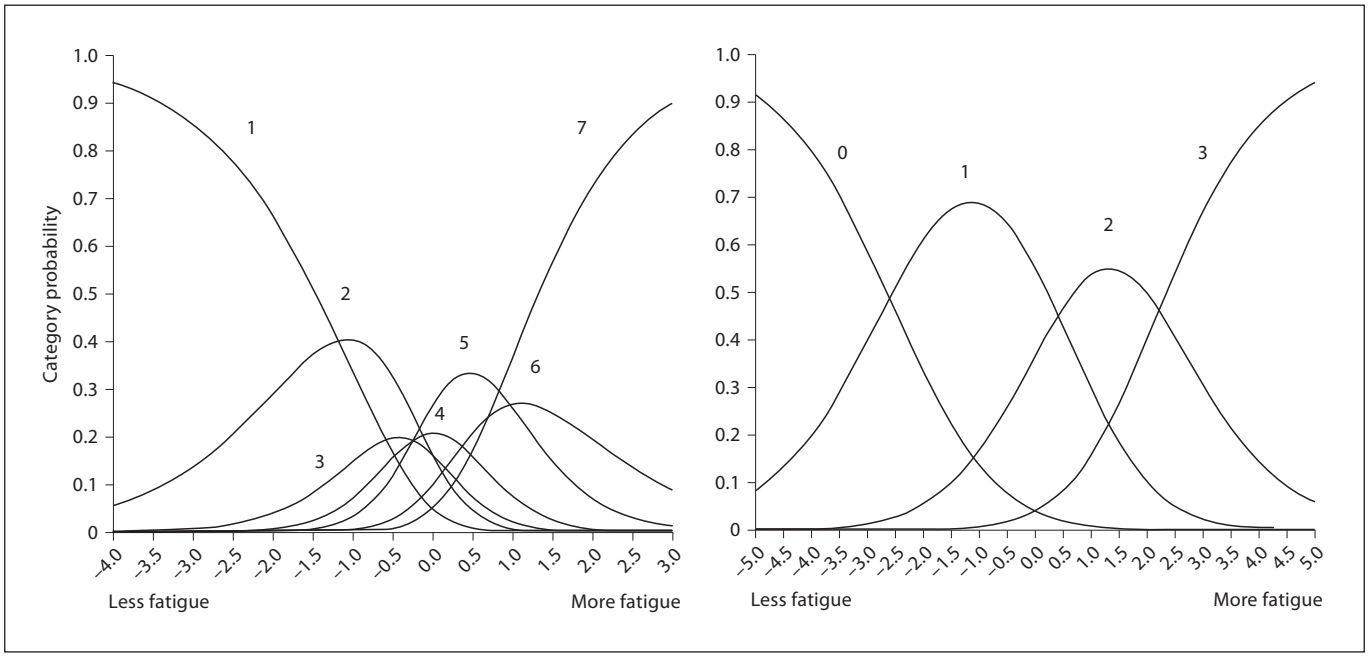


Fig. 2. Rating scale functioning of the FSS. The category probability curves for the 7-point scale recoded into a 4-point rating scale. Step calibrations should be monotonically ordered. Categories never emerging as modal and category threshold disordering violate rating scale assumptions and require collapsing until rating scale diagnostical criteria are met (Hagell et al. [17]).

Rating Scale Functionality

Rating scales are used as a communication tool between the test developer's intention and the respondents' record on the construct. Rasch rating scale diagnostics provide statistical guidance in evaluating how well the response categories function to create an interpretable measure [32]. Generally, regular frequency distributions (e.g. Gaussian, uniform, bimodal) and a comparable number of responses across all categories are recommended for good measurement. Additionally, step calibrations (= estimated endorsement rate for choosing one category over the other) should be ordered (i.e. increase monotonically and preferably by 1.4 logits between categories) and display good fit. Outfit mean squares should not exceed 2 in order to provide sufficient information over misinformation due to the response scale [35]. If problems are diagnosed, collapsing problematic categories into better-behaving adjacent categories is often a good remedy to reduce noise and improve variable clarity [32, 35]. Still, collapsing categories must make sense and translate the perceptive levels of the respondents regarding the underlying construct [32].

Rating scale diagnostics of the ESS indicated a good fit of the rating scale with outfit mean squares ranging from 0.86 to 1.03 (fig. 1). Together with satisfactory fit indices, we observe an ordered rating scale with monotonically increasing step calibrations (n.s., –1.51, 0.29, 1.22). The observed counts of responses in categories 3 and 4 ('moderate' and 'high chance of dozing') are somewhat lower than expected (19 and 15% where we would expect around 25% of the responses in each category). Hence, step calibrations between categories 3 and 4 are less than expected (0.93 logits), implying that a lack of distinction between these categories may exist and that collapsing categories may be considered. However, as categories are not unacceptably underutilized and cover a reasonable length of the scale (fig. 1), collapsing ESS categories seems not to be necessary for a good measurement here. The underutilization of categories 3 and 4 rather illustrates that the population mean is in the lower half of the scale rather than in the center. This simply suggests that, in a community-based sample as opposed to a clinical population [18], the probability of dozing is perceived more frequently as nonexistent or small than moderate or large. The rating scale itself seems to function adequately and par-

ticipants are able to discern 4 distinct levels of dozing probability.

Rating scale analysis of the FSS reveals an adequate fit with outfit mean squares ranging from 0.77 to 1.21 (fig. 2). However, as apparent from figure 2, a number of categories never emerge as modal (categories 3, 4, and 6) and step calibrations severely violate the rating scale assumptions (n.s., –1.13, 0.00, –0.24, –0.27, 0.99, 0.66). Rescoring the category scale into a 4-point category scale yielded satisfactory outfit indices ranging from 0.94 to 1.33 (fig. 2). Category 1 was recoded into category 0 (= 'strongly disagree'), the following 3 categories were recoded into category 1 (= 'disagree'), categories 5 and 6 into category 2 (= 'agree') and the last category was recoded into category 3 (= 'strongly agree'). The recalibrated rating scale is well ordered with well-spaced, monotonically increasing step calibrations (n.s., –2.68, 0.21, 2.48).

Item Fit and Dimensionality

The fit analysis reports how well the actual data agree with the linear measures obtained from the Rasch model. Item outfit mean squares are outlier-sensitive fit statistics based on a χ^2 statistic. Outfit mean squares are sensitive to unexpected responses on very easily and/or very difficultly endorsable items. Item infit mean squares are 'inlier' pattern sensitive, meaning that this measure is more sensitive to unexpected responses by respondents on items with difficulties (roughly) matching their abilities [33]. Both infit and outfit mean squares are expected to be 1.0. Values less than 1.0 indicate too predictable values for the Rasch model (overfit or redundancy) and values greater than 1.0 indicate noise (underfit). As mean squares generally average 1.0, higher and lower values are common. Values between 0.5 and 1.5 are the most productive for measurement. Values above 1.5 and below 2.0 are unproductive for measurement but not degrading. Items with fit values exceeding 2.0 degrade the measurement system and should be excluded.

Our results show good outfit mean square statistics for all items of the ESS, ranging from 0.62 to 1.22. Infit mean squares range from 0.77 to 1.21 (table 2). Moreover, all observed values are less than 1.4, which indicates the unidimensional nature of the sleepiness measure from the ESS.

Table 3 displays the item statistics for the FSS. The results from the analysis show outfit mean square statistics ranging from 0.62 to 1.66 and infit mean squares ranging from 0.59 to 1.46. Because the items 'My motivation is lower when I am fatigued' and 'Exercise brings on my fatigue' have both outfit mean squares indicating that they are not productive for measurement, standardized fit statistics are investigated. Standardized fit statistics for both items exceed 2 (outfit tz = 4.6 and 4 and infit tz = 3.6 and 3.7, respectively). These values designate too much variation in the responses (underfit). For instance, these values may reflect individuals disagreeing that exercise brings on fatigue or that motivation is lower when fatigued while their fatigue scores on the Rasch model predict that they would. Because some items show fit values exceeding 1.4, the overall dimensionality of the instrument is assessed by means of Rasch Factor Analysis (RFA). The RFA consists of a regular Rasch analysis procedure as described so far, thus using the ordinal raw data to construct a linear measure, followed by an unrotated principal factor analysis of the (ordinal-level) residual data after extraction of the Rasch measure. This procedure is used to identify possible common variance among the unexplained or unmodeled data by the general Rasch dimension [36]. Items (attributes) that have substantial variance unexplained by the overall Rasch dimensions have high (>0.40) positive or negative loadings, thus potentially indicating the existence of a secondary dimension in the data, characterized by the content of the contrast of these high positive and negative factor loadings. The importance of the secondary dimension can be derived from the eigenvalues of the unmodeled variance. If the unexplained variance of the contrast is below 3.0, the unidimensionality of the data is well preserved. Ideally, we expect values as low as random normal deviates would predict, which is around 1.4. Additionally, we consider the unexplained variance of the contrast being less than 5% of the total unexplained variance as an excellent indication of unidimensionality. Other tentative guidelines for the RFA include checks of the variance explained by items (>4 × 1st contrast is good) and by the Rasch measures (>50% is good) [33].

Our results show that 56.1% of variance is explained by the Rasch dimension (35.7% by the person agreeability range and 20.4% by item endorsability range). The variance explained by the first contrast (9.2%) has the strength of the unmodeled variance in less than 3 items (eigenvalue 1.9) and in less than the model-predicted chance values (1.5–2.0). In conclusion, the results of the RFA show a fairly well-preserved unidimensionality of the construct and as both items do not degrade measurement, they remained included for further analyses.

Invariance Assessment

Measurement invariance between groups implies that, in order to execute between-group comparisons on latent

Table 2. Rasch item statistics of the ESS

Item content	Logits	SE	Infit mean squares	Outfit mean squares
(5) Lying down to rest in the afternoon when circumstances permit	–1.49	0.12	0.95	1.02
(2) Watching TV	–0.51	0.11	1.12	1.19
(4) As a passenger in a car for 1 h without a break	0.14	0.11	1.21	1.22
(1) Sitting and reading	0.39	0.11	1.00	0.96
(7) Sitting quietly after lunch without alcohol	1.04	0.12	0.77	0.77
(3) Sitting inactively in a public place	1.32	0.12	0.90	0.98
(8) In a car, while stopped for a few minutes in traffic	3.33	0.20	0.89	0.62
(6) Sitting and talking to someone	3.94	0.24	0.95	1.00

Table 3. Rasch item statistics of the FSS

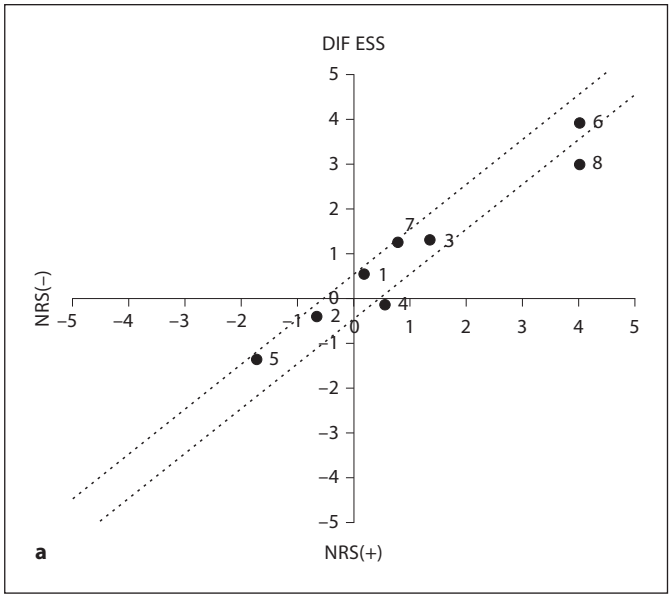
Item content	Logits	SE	Infit mean squares	Outfit mean squares
(a) My motivation is lower when I am fatigued	–1.75	0.14	1.46	1.66
(b) Exercise brings on my fatigue	–0.48	0.14	1.47	1.51
(c) Fatigue interferes with my physical functioning	–0.33	0.14	0.59	0.62
(d) I am easily fatigued	–0.29	0.14	0.73	0.76
(e) Fatigue interferes with my work, family, or social life	0.20	0.14	0.97	0.93
(f) Fatigue is among my three most disabling symptoms	0.28	0.14	1.22	1.15
(g) My fatigue prevents sustained physical functioning	0.52	0.14	0.84	0.81
(h) Fatigue interferes with carrying out certain duties and responsibilities	0.92	0.15	0.88	0.84
(i) Fatigue causes frequent problems for me	0.92	0.15	0.90	0.85

trait scores, the numerical values in use must be on the same measurement scale. If measurement invariance is not achieved, the between-group mean levels are potentially biased and lead to flawed interpretations [37]. DIF analysis allows for the investigation of disagreement regarding the location of an item's endorsability estimate. When this location varies across groups by more than the modeled error, evidence of DIF is suspected [32]. In the current DIF analysis, item measures (DIF sizes) are computed for the NRS complaint group against the endorsability estimates of the group of individuals with no complaints of NRS. The DIF analysis is executed separately for the ESS (fig. 3a) and the FSS (fig. 3b). The DIF contrast is the pairwise difference in endorsability of the item between groups and should be statistically significant and/or at least 0.5 logits to be noticeable [32].

The DIF analysis of the ESS shows a partial difference in item endorsability estimates between individuals with NRS [NRS(+)] and without NRS complaints [NRS(–)]. Item 4 ('As a passenger in a car for 1 h without a break') and item 8 ('In a car, while stopped for a few minutes in traffic') display different endorsability estimates for the NRS(+) group, exceeding the 0.5 logit DIF contrast level (fig. 3a). This difference is supported statistically (item 4: $\chi^2 = 9.49$, $p < 0.005$ and item 8: $\chi^2 = 6.15$, $p < 0.05$). DIF analysis of the FSS shows no statistically significant difference in item endorsability estimates between groups regarding the presence or absence of NRS complaints (fig. 3b).

Common Person Equating

Common person equating was performed by plotting the Rasch person estimates (logit scale) of the ESS and the FSS against each other (fig. 4). The ESS person estimates were recalculated omitting the items displaying DIF in order to improve the quality of the person-linking operation. FSS scores were adjusted for the mean difference between ESS and FSS scores (–0.66) because the FSS is an



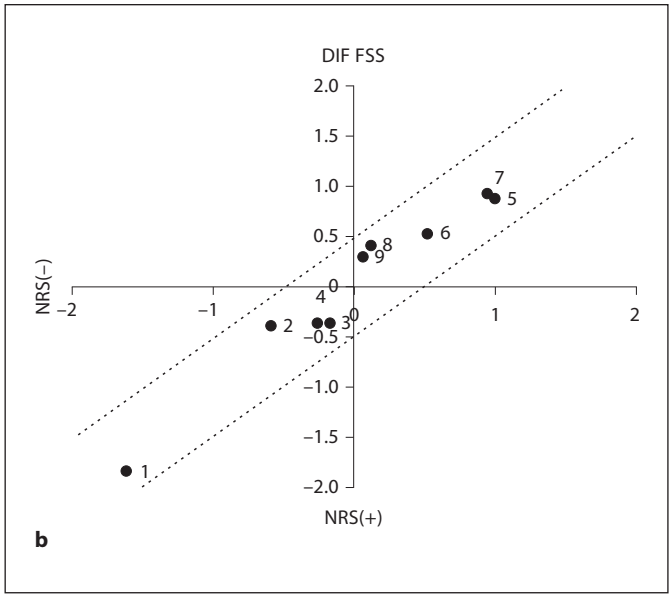


Fig. 3. a Measurement invariance of the ESS in individuals with and without NRS complaints. DIF analysis shows invariance for most items on the ESS regarding NRS. However, item 4 ('As a passenger in a car for 1 h without a break') and item 8 ('In a car, while stopped for a few minutes in traffic') display different difficulty estimates for the NRS(+) group. The dotted lines represent the 0.05 logit boundaries. **b** Measurement invariance of the FSS in individuals with and without NRS complaints. DIF analysis of the FSS shows complete invariance and therefore no significant difference in item difficulty estimates between groups, regarding NRS complaints. The dotted lines represent the 0.05 logit boundaries.

'easier' instrument. The 95% CI control lines were derived from the person estimate standard errors. Under ideal circumstances of concurrent validity, the empirical regression line should have a slope of 1 and an intercept of 0 (ideal fit diagonal in fig. 4). The slope of the empirical regression line is 0.33 and the intercept 0.17. This suggests that both instruments measure a different construct. However, our results show that grossly 80% of the cross-plotted estimates fall within the 95% CI, suggesting that a majority of the investigated sample either fails to make a distinction between subjective fatigue and sleepiness or possess, within error, comparable levels of sleepiness and fatigue.

Regarding group differences between individuals with and without NRS complaints, the plot reveals a similar horizontal spread between the agreeability estimates of the different groups, with no systematic clustering other than a slight predominance of estimates for individuals with NRS complaints higher on the logit scale. This vertical discrepancy is expected as individuals with NRS tend indeed to report higher fatigue and sleepiness levels (table 1).

Discussion

As stated before, the literature investigating the relationships between fatigue and sleepiness, in different population samples, remains sparse [11, 13–16]. In shift workers for instance, Hossain et al. [16] reported only a weak correlation between fatigue and sleepiness, whereas Bailes et al. [11] reported highly significant correlations in a community sample of older individuals. In a more recent study of a clinical population in a sleep laboratory, Hossain et al. [3] reported a 'non-significant association between continuous variables of subjective fatigue and sleepiness' and the authors concluded that subjective fatigue and sleepiness can be independent manifestations of sleep disorders and should therefore require independent assessments.

The above-mentioned previous studies reporting associations between fatigue and sleepiness mainly used correlational approaches. A potential flaw in their approach is that correlations do not take measurement error into account that can result in attenuated r values [32]. A more comprehensive assessment of discriminant validity can be achieved by considering the IRT [37, 38]. Within

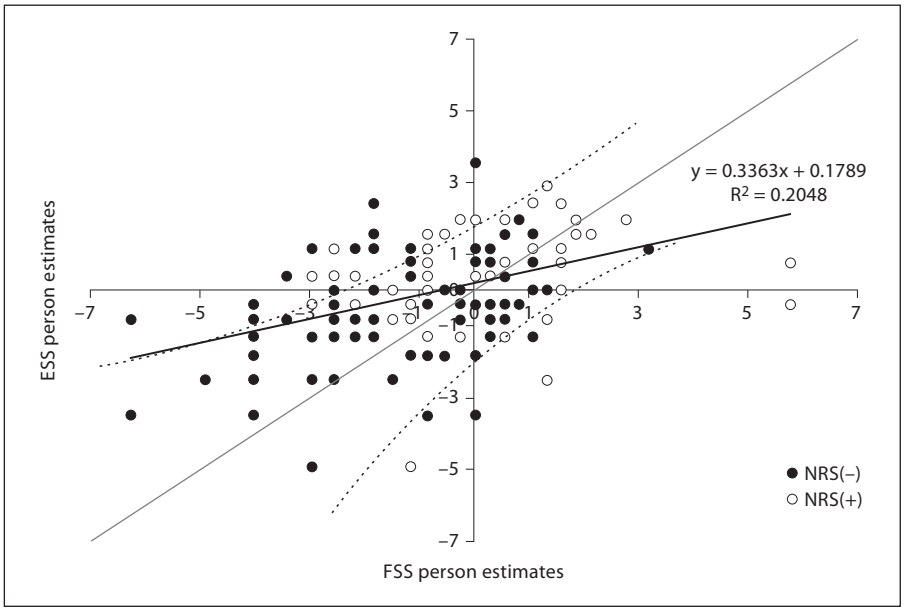


Fig. 4. Common person equating. White dots represent subjects with an NRS complaint and black dots represent subject without NRS. The black line represents the empirical regression line, the gray diagonal represents the ideal fit line. The dotted lines represent the estimates' 95% CI.

the here proposed framework of IRT, a person's performance on a test is related to quantitative and qualitative components: a person estimate (i.e. measure) and its error term. It is necessary when linking person estimates of different measurement instruments that those two components are included to reveal the extent of invariance [32].

Descriptively, we found here that the perception of higher sleepiness and fatigue levels can be related to complaints of NRS and to reported lower habitual sleep duration. Scores on all psychometric scales showed higher levels in participants with an NRS complaint. Nevertheless, with the exception of fatigue, the 'clinical significance' of the respective intensities on the other psychometric scales remains rather negligible. In contrast to others, we found a strong relationship between fatigue and sleepiness, disregarding NRS, perhaps underlining the lack of distinction and discrimination of fatigue and sleepiness in nonclinical population samples. Hossain et al. [3] recently showed in a clinical population that fatigue and sleepiness can be independent of diagnosed sleep disorders and more interestingly they showed only a weak correlation of ESS and FSS scores in their sample. Furthermore, they showed very different overlap proportions of associated sleepiness in patients with a fatigue complaint and vice versa. Indeed, in a presently ongoing study in our sleep laboratory, we also find, in contrast to the present results, very similar results to those of Hossain et al. in a clinical population [unpublished data]. These findings seem to indicate that the discrimination and differentiation of the perceptions of fatigue and sleepiness in a clinical and in a nonclinical population could be dissimilar.

However, within the IRT, it is the applied Rasch analysis methodology that permitted us to further describe the consistency of the items in the ESS and the FSS and their relationships to NRS and to each other. Indeed our results showed a good consistency in the response patterns among symptom groups of both scales and a unidimensional structure and invariance of the FSS. DIF analysis of the ESS showed partial differences in item endorsability estimates regarding NRS. From a perceptional point of view, one could speculate that sleepiness has perhaps a different sensitivity to NRS. We also showed that the underutilization of some response categories, on the ESS in the present sample, needed recalibration and the use of a Rasch-transformed person score. The latter can also be related to the fact that 'high chances of dozing' are naturally less frequent in nonclinical populations. Nevertheless, the fact that, disregarding descriptive differences (with higher FSS and ESS scores in subjects with an NRS complaint), further analyses only showed little distinction in the perceptions (response pattern) of these concepts (fatigue and sleepiness) in both categories [NRS(+) and NRS(–)] could of course underline that there is indeed a 'real' difference in fatigue and sleepiness levels in subjects with NRS here. Hence, the numerical differences on the measurement scales mostly revealed only weak clinical significances.

At last, following common person equating, the cross-plotted estimates showed that a large majority (about 80%) of participants failed to distinguish between subjective fatigue and sleepiness or possess, for whatever reason, absolutely similar levels of both symptom complexes although both seem to describe different concepts. Put differently, the characteristics of the Rasch-derived empirical regression rather suggest that despite the lack of differentiation but similar 'agreeability (ability) scores' for fatigue and sleepiness in most participants, the presented fatigue and sleepiness scales measure distinct underlying constructs.

Limitations of the present study are related to the relatively small sample size for a community-based study. The exploration of the data was therefore limited and we mainly focused on the Rasch methodology-based analyses of the relationships between subjective fatigue and sleepiness intensities regarding NRS. The respective prevalences of excessive sleepiness (ESS score >10), moderately severe to severe fatigue (FSS score >4) and NRS are relatively high in our sample. Nevertheless, these characteristics do not influence the measured relationships between the explored concepts. Moreover, different cutoff points (ESS score >11 and FSS score >5) did not influence the main outcome results (data not shown). Furthermore, prevalences of sleepiness in previous community-based studies present with a very high variability ranging from 0.3 to 43%. This large variability is often explained by the assessment types of sleepiness (simple questions rather than ESS) or a probably very frequent confusion with the concept of daytime fatigue [1, 5, 11, 13].

Although some recent findings pointed out the fact that the differentiation of sleepiness and fatigue, and therefore also their interrelation, could depend on the setting of a psychometric investigation, the underlying physiological processes that lead to sensations of exhaustion, fatigue, asthenia, drowsiness or sleepiness should nevertheless remain the same. However, it seems as if the symptomatic assessment of the intensity of the related complaints looses precision when nonclinical subjects are investigated.

If the improvement of the understanding and differentiation of underlying physiological processes are the goals of future investigations, then Rasch modelization should be applied to different sample types. Hence, further research should investigate the relationships between these perceptional and semiological concepts and observable variables such as vigilance, psychomotor performance, effort measures or polysomnographic recording for instance.

Acknowledgements

The present work was supported by CRSB, a private fund dedicated to research in sleep medicine. D.N. was supported by a research grant from the Ministry of Research, Culture and Superior Education of the Grand-Duchy of Luxembourg. J.-B.V.is supported by an F.R.I.A. research grant (Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture), Belgium. P.L. is supported by the Belgian National Funding for Scientific Research (F.N.R.S.), Belgium.

References

1 Ohayon MM: Prevalence and correlates of nonrestorative sleep complaints. Arch Intern Med 2005;165:35–41.
2 Pawlikowska T, Chalder T, Hirsch SR, Wallace P, Wright DJ, Wessely SC: Population based study of fatigue and psychological distress. BMJ 1994;308:763–766.
3 Hossain JL, Ahmad P, Reinish LW, Kayumov L, Hossain NK, Shapiro CM: Subjective fatigue and subjective sleepiness: two independent consequences of sleep disorders? J Sleep Res 2005;14:245–253.
4 Kim H, Young T: Subjective daytime sleepiness: dimensions and correlates in the general population. Sleep 2005;28:625–634.
5 Young TB: Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. J Clin Psychiatry 2004;65(suppl 16):12–16.
6 Pigeon WR, Sateia MJ, Ferguson RJ: Distinguishing between excessive daytime sleepiness and fatigue: toward improved detection and treatment. J Psychosom Res 2003;54:61–69.
7 Leger D: The cost of sleepiness. Sleep 1995; 18:281–284.
8 Borbély AA: A two process model of sleep regulation. Hum Neurobiol 1982;1:195–204.
9 Lavie P: Ultrashort sleep-waking schedule. 3. 'Gates' and 'forbidden zones' for sleep. Electroencephalogr Clin Neurophysiol 1986;63: 414–425.
10 Neu D, Hoffmann G, Moutrier R, Verbanck P, Linkowski P, Le Bon O: Are patients with chronic fatigue syndrome just 'tired' or also 'sleepy'? J Sleep Res 2008;17:427–431.
11 Bailes S, Libman E, Baltzan M, Amsel R, Schondorf R, Fichten CS: Brief and distinct empirical sleepiness and fatigue scales. J Psychosom Res 2006;60:605–613.
12 Leibowitz SM, Brooks SN, Black JE: Excessive daytime sleepiness: considerations for the psychiatrist. Psychiatr Clin North Am 2006;29:921–945.
13 Shen J, Barbera J, Shapiro CM: Distinguishing sleepiness and fatigue: focus on definition and measurement. Sleep Med Rev 2006; 10:63–76.
14 Lichstein KL, Means MK, Noe SL, Aguillard RN: Fatigue and sleep disorders. Behav Res Ther 1997;35:733–740.
15 Chervin RD: Sleepiness, fatigue, tiredness, and lack of energy in obstructive sleep apnea. Chest 2000;118:372–379.

16 Hossain JL, Reinish LW, Kayumov L, Bhuiya P, Shapiro CM: Underlying sleep pathology may cause chronic high fatigue in shift-workers. J Sleep Res 2003;12:223–230.
17 Hagell P, Höglund A, Reimer J, Eriksson B, Knutsson I, Widner H, Cella D: Measuring fatigue in Parkinson's disease: a psychometric study of two brief generic fatigue questionnaires. J Pain Symptom Manage 2006; 32:420–432.
18 Hagell P, Broman JE: Measurement properties and hierarchical item structure of the Epworth Sleepiness Scale in Parkinson's disease. J Sleep Res 2007;16:102–109.
19 Kolen MJ, Brennan RL: Test Equating: Methods and Practices. New York, Springer, 2004.
20 Yu CH, Osborn Popp SE: Test Equating by Common Items and Common Subjects: concepts and applications. Pract Assess Res Eval 2005;10:1–19.
21 Johns MW: A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–545.
22 Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD: The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989;46:1121–1123.
23 Lerdal A, Wahl A, Rustøen T, Hanestad BR, Moum T: Fatigue in the general population: a translation and test of the psychometric properties of the Norwegian version of the fatigue severity scale. Scand J Public Health 2005;33:123–130.
24 Stone P, Richards M, A'Hern R, Hardy J: A study to investigate the prevalence, severity and correlates of fatigue among patients with cancer in comparison with a control group of volunteers without cancer. Ann Oncol 2000; 11:561–567.
25 Flachenecker P, Kümpfel T, Kallmann B, Gottschalk M, Grauer O, Rieckmann P, Trenkwalder C, Toyka KV: Fatigue in multiple sclerosis: a comparison of different rating scales and correlation to clinical parameters. Mult Scler 2002;8:523–526.
26 Kos D, Nagels G, D'Hooghe MB, Duportail M, Kerckhofs E: A rapid screening tool for fatigue impact in multiple sclerosis. BMC Neurol 2007;6:27.
27 Hoddes E, Dement WC, Zarcone V: The development and use of the Stanford Sleepiness Scale. Psychophysiology 1972;9:150.
28 Zigmond AS, Snaith RP: The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67:361–370.
29 McCue P, Buchanan T, Martin CR: Screening for psychological distress using internet administration of the Hospital Anxiety and Depression Scale (HADS) in individuals with chronic fatigue syndrome. Br J Clin Psychol 2006;45:483–498.
30 Rasch G: Probabilistic Models for Some Intelligence and Attainment Tests (expanded edition). Chicago, University of Chicago Press, 1980.
31 Andrich D: A rating formulation for ordered response categories. Psychometrika 1978;43: 561–573.
32 Bond TG, Fox CM: Applying the Rasch Model: Fundamental Measurement in the Human Sciences, ed 2. Philadelphia, Erlbaum, 2007.
33 Linacre JM: User's Guide and Program Manual to WINSTEPS: Rasch Model Computer Programs. Chicago, MESA Press, 2009.
34 Tesio L: Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105–115.
35 Linacre JM: Investigating rating scale category utility. J Outcome Meas 1999;3:103–122.
36 Linacre JM: Detecting multidimensionality: which residual data-type works best? J Outcome Meas 1998;2:266–283.
37 Reise SP, Widaman KF, Pugh RH: Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114: 552–566.
38 Hambleton RK, Swaminathan H, Rogers HJ: Fundamentals of Item Response Theory. Newbury Park, Sage Press, 1991.