Psychiatry Research 238 (2016) 100–108

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Tired of blunt tools? Sharpening the clinical assessment of fatigue and sleepiness

Olivier Mairesse [a,b,c,d], Daniel Neu [a,b,*]

[a] Brugmann University Hospital, Sleep Laboratory and Unit for Chronobiology U78, Free University of Brussels-Université Libre de Bruxelles, U.L.B., Brussels, Belgium
[b] UNI Neuroscience Institute, Faculty of Medicine, Laboratory for Medical Psychology ULB312 and Faculty of Motor Sciences, Université Libre de Bruxelles (U.L. B.), Brussels, Belgium
[c] Department of Experimental and Applied Psychology (EXTO), Vrije Universiteit Brussel (V.U.B.), Brussels, Belgium
[d] Royal Military Academy, Department LIFE, Brussels, Belgium

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form
21 December 2015
Accepted 7 February 2016
Available online 8 February 2016

Keywords:
Fatigue
Sleepiness
Rasch analysis
Common person equating
Sleep quality
Fatigue Severity Scale
Epworth Sleepiness Scale
Pittsburgh Sleep Quality Inventory

ABSTRACT

Fatigue and sleepiness are ubiquitous symptoms in various conditions and are frequently associated to impaired sleep quality. While separate fatigue and sleepiness scales exist, both constructs are often confused. Unraveling this issue requires estimating the instruments' measurement properties, potential scale recalibration and re-evaluation of symptom intensities on a comparable basis. This study aims at improving the assessment of these symptoms and quantifying their degree of overlap using common-person-equating (CPE). One hundred fifty-nine patients, either with complaints of fatigue, sleepiness and/or non-restorative sleep, addressed to an academic sleep unit for a full-night polysomnography (PSG), enrolled in the study. Symptom levels were measured with the Fatigue Severity (FSS) and Epworth Sleepiness (ESS) scales. Sleep quality was assessed by the Pittsburgh Sleep Quality Index, defining 'good' and 'poor' sleeper groups. Good and poor sleepers did not differ statistically regarding demographics and PSG parameters. Rasch analysis revealed that, considering proper calibration, the ESS and FSS generate reliable and valid, unidimensional linear measures and to be invariant to perceived sleep quality. CPE showed predominantly fatigued, rather than sleepy patients, being more likely to present as poor sleepers. A concordance diagram based on scale scores is provided, in order to improve the differentiation of both symptoms.

1. Introduction

Although that sleepiness and fatigue are very different concepts from physiological or patho-physiological points of view, semiological and semantic confusion frequently persists in patients, clinicians and researchers alike (Shen et al., 2006a; Neu et al., 2010a).

Sleepiness is a physiological phenomenon having primarily a behavioral function: signaling the need for sleep. It is regulated by circadian and homeostatic drives and further influenced by environmental and behavioral stimuli or situations (Borbély, 1982; Johns, 1998, 2002; Mairesse et al., 2010). We generally refer to excessive daytime sleepiness (EDS) in pathological conditions (Neu et al., 2010a; Ohayon, 2008). EDS presents as a higher daytime sleep pressure and increased sleep propensity (Shen et al., 2006a; Neu et al., 2010a; Johns, 2002). EDS is mainly related to sleep deprivation or can occur as a symptom of certain primary sleep disorders.

Fatigue also has a physiological counterpart, as in acute and very energy demanding tasks. In contrast to sleepiness, it resolves with rest and does not require sleep to recover from. In pathological conditions, we therefore generally refer to chronic fatigue that is by definition not alleviated with rest and exacerbated by mental or physical tasks (Shen et al., 2006a; Neu et al., 2010a). Chronic fatigue is rather related to systemic conditions (inflammatory processes, immune disorders, cancer, major depression e.g.), to insomnia (a hyper-arousal condition with a relative absence of sleepiness) and it is the key symptom of the chronic fatigue syndrome (Neu et al., 2010a). Both chronic fatigue and EDS present very heavy burdens for public health care (Shen et al., 2006b; Neu et al., 2010b). Although fatigue and sleepiness can be related to non-restorative sleep complaints, (Shen et al., 2006a; 2006b; Neu et al., 2010b) both have however also been shown to

[*] Correspondence to: Brugmann University Hospital, Université Libre de Bruxelles (U.L.B.), Sleep Laboratory and Unit for Chronobiology U78, Arthur Van Gehuchten Square 4-Building Hh, 1020 Brussels, Belgium.
E-mail addresses: olivier.mairesse@chu-brugmann.be (O. Mairesse), daniel.neu@chu-brugmann.be (D. Neu).

http://dx.doi.org/10.1016/j.psychres.2016.02.005
0165-1781/© 2016 Elsevier Ireland Ltd. All rights reserved.

be potentially independent consequences from sleep disorders (Hossain et al., 2005). The relationships between both symptomatic conditions and how they relate to perceived sleep quality remain therefore poorly understood (Guilleminault et al., 2006). Moreover, adequate care and adapted treatment strategies depend on the ability to clinically differentiate between both conditions. An essential step towards a better disentanglement of fatigue and sleepiness is, first of all, exploring the psychometric properties of commonly used clinical assessment tools and their potential dimensional overlap (Shen et al., 2006a; Neu et al., 2010a).

The most widely used self-report instrument for sleepiness in sleep research is the Epworth Sleepiness Scale (ESS) (Neu et al., 2010a). Global levels of fatigue symptoms are commonly measured by means of the Fatigue Severity Scale (FSS). The FSS is currently one of the most widely used fatigue scales in general internal medicine (Neu et al., 2010a). Both instruments have shown to be significantly correlated, but strengths of association varied with respect to the studied population. Correlations ranged from .18 in an observational study of sleep disordered patients with various diagnoses (Hossain et al., 2005), .33 in shift workers (Shen et al., 2006b), .48 in Parkinson's disease (Valko et al., 2010), .49 in a general population sample (Neu et al., 2010b) and .50 in multiple sclerosis patients (Braley et al., 2012). These findings further add to the confusion about how sleepiness and fatigue relate, since correlation statistics provide no indication of measurement error and consequently fail to supply information on how precisely each construct is measured separately (Bond and Fox, 2007). Rasch analysis (Rasch, 1980) however, allows us to do just that.

Rasch analyses of the Fatigue Severity Scale (FSS) and the Epworth Sleepiness Scale (ESS) were recently performed separately in samples of Parkinson's disease patients (Hagell et al., 2006, Hagell and Broman, 2007). Both instruments showed to possess good psychometric properties and stable hierarchical item structures. However, in order to provide insight about the overlap of both constructs, a combined approach is required. Rasch-based common person equating (CPE) allows for such associative exploration. CPE refers to a process of determining comparable scores on different instruments measuring the same hypothetical construct, when administered to a common group of persons (Yu and Popp, 2005). By analogy, if fatigue and sleepiness are undifferentiated concepts, Rasch-derived person estimates on the ESS and the FSS should – within error – return similar after CPE. This approach has yet only been investigated once in a community sample, however without psychometric evaluation of perceived sleep quality (Neu et al., 2010b). This study showed that, although related, subjective fatigue and sleepiness seem to represent different underlying concepts, evidenced by a smaller than expected convergence between ESS and FSS measures with the assumption that they measured similar constructs. Nevertheless, it remains to be determined whether the observed similarities were simply a result of fatigue and sleepiness levels below clinical thresholds in a general population sample, or if insufficient experience (Lane et al., 2011) with either one or both of the symptoms attenuates the individuals' discrimination ability and contributes to the observed overlap of person estimates (Neu et al., 2010b). Additionally, given that both symptom clusters can be associated with complaints of impaired sleep quality (Neu et al., 2010a), measurement instruments of fatigue and sleepiness might therefore not be invariant to it.

In the present study we therefore propose to use Rasch-based CPE to investigate subjective fatigue and sleepiness, in a polysomnography (PSG) monitored clinical sample, with respect to perceived sleep quality. We hypothesize that both instruments (1) possess satisfying psychometric properties (i.e. rating scale functioning, item fit and undimensionality), (2) lack measurement invariance to sleep quality, (3) possess high convergent validity disregarding sleep quality and (4) converge differently with respect to perceived sleep quality.

2. Methods

We recruited 159 hypnotic-free consecutive patients addressed to the sleep unit (mean age 47 +/− 13 years, 23 females) in a cross-sectional study protocol. Inclusion criteria mainly comprised primary care referral for clinical complaints of either fatigue or sleepiness or non-restorative sleep alone, or combinations of the aforementioned symptoms. All patients were instructed to withdraw from any hypnotic drugs, for at least 2 weeks, prior to admission to the sleep laboratory. In addition, no neuropsychopharmacological treatment was administered during the hospitalization period. All patients underwent semi-structured clinical interviews in order to rule out significant mental disorders (including major depression, bipolar or psychotic disorders) according to DSM IV criteria (APA, 2000). In addition, all patients in our lab completed questionnaires about life style and drinking habits. Substance abuse and a consumption of more then two units of alcohol per day were excluded. Daytime napping was not allowed during the hospitalization period. Any treatment essay, comprising application of continuous positive airway pressure (CPAP) was excluded.

2.1. Instrumentation

We assessed fatigue and sleepiness with the *Epworth Sleepiness* (ESS) and the *Fatigue Severity* (FSS) scales respectively. The *ESS* consists of 8 items that describe specific situations for which individuals should indicate their chance of dozing off by means of a 4-point Likert scale ranging from 0 (never doze) to 3 (high chance of dozing). The summed scores range from 0 to 24 and scores above 10 are commonly interpreted as increased global sleep propensity (Johns, 1991). The FSS is a self-report scale for the assessment of daytime fatigue and its impact on daily functioning. The FSS was introduced on individuals with multiple sclerosis and systemic lupus erythematosus (Krupp et al., 1989), it has since then been used in studies investigating fatigue in chronic conditions like obesity, Parkinson's disease, hepatitis C infection, Chronic Fatigue Syndrome (Olson et al., 2003; Neu et al., 2007) and in general population samples (Neu et al., 2010b; Lerdal et al., 2005). The FSS is a 9 item, 7-point Likert scale. 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. A mean score of 4 (Krupp et al., 1989) or 5 (Lerdal et al., 2005) have been proposed as thresholds for clinically significant or pathological chronic fatigue.

Sleep quality perception was measured with the *Pittsburgh Sleep Quality Index* (PSQI). The PSQI is by far the most widely used assessment of subjective sleep quality. This instrument contains 10 questions reflecting 7 different sleep related components: sleep quality, latency, duration, habitual sleep efficiency, sleep disturbance, use of hypnotics and daytime functioning. The component scores are then summed to give the global PSQI score. A global PSQI score above 5 has been shown to indicate severe difficulties in at least two of the above-mentioned components, or moderate difficulty in more than three areas (Buysse et al., 1989). Clinically significant alteration of perceived sleep quality was also defined here as a PSQI score >5. In order to control for sleep quality, comparisons of patients' descriptive variables and respective fatigue and sleepiness levels, within the included sample, were conducted between ‘poor sleepers’ (PSQI >5, $n=102$) and ‘good sleepers’ (PSQI ≤ 5, $n=47$) according to PSQI scores (Buysse et al., 1989).

Table 1
Descriptive variables.

	Good sleepers		**Poor sleepers**		**MANOVA**		
	(PSQI ≤ 5)		(PSQI > 5)				
	(*n*= 47, 15 females)		(*n*=102, 52 females)				
	M	SD	M	SD	*F*	*p*	η^2
Age	43.83	13.72	48.38	12.74	3.806	0.053	0.026
BMI (kg/m^2)	27.80	5.52	28.15	5.62	0.122	0.727	0.001
TIB (min)	521.15	59.44	519.03	67.20	0.034	0.855	0.000
TST (min)	405.24	86.29	385.92	84.96	1.603	0.208	0.011
SEI (%)	77.45	12.66	74.37	13.60	1.674	0.198	0.012
SOL (min)	27.92	24.08	37.54	42.23	2.065	0.153	0.014
WASO (min)	64.41	40.45	72.52	51.39	0.885	0.348	0.006
N1 (min)	49.03	38.38	40.60	21.63	2.832	0.095	0.020
N2 (min)	217.36	59.76	223.00	70.68	0.219	0.641	0.002
SWS (min)	84.99	44.43	73.42	45.23	2.072	0.152	0.014
REM (min)	53.86	26.06	48.89	28.57	0.999	0.319	0.007
REMLAT (min)	141.35	99.76	149.47	95.16	0.221	0.639	0.002
AHI (/h)	16.17	18.57	14.39	20.48	0.251	0.617	0.002
ODI (/h)	13.33	16.49	13.66	20.90	0.009	0.926	0.000
ArI (/h)	29.48	15.36	33.54	22.68	1.208	0.274	0.008
FSS	3.08	1.65	4.66	1.53	31.886	< **0.001**	0.183
ESS	9.04	5.70	9.58	5.69	0.280	0.598	0.002
PSQI	3.57	1.29	11.09	3.63	185.788	< **0.001**	0.567

'Good' and 'poor' Sleepers according to PSQI cut-offs (PSQI ≤ 5 or > 5); Mean (M); Standard Deviation (SD); Body mass index (BMI) in kg/m^2; time in bed (TIB), sleep period time (SPT), total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), slow wave sleep (SWS), Rapid Eye Movement (REM) Sleep, REM latency (REMLAT) in minutes (min); sleep efficiency index (SEI=(TST/TIB)*100) in percent (%); apnea-hypopnea index (AHI), oxygen desaturation index (ODI) and arousal index (ArI) are ratios per hour without units; fatigue severity scale (FSS) expressed as mean scores, Epworth sleepiness scale (ESS), hospital anxiety and depression rating scale (HAD tot) and anxiety/depression sub scores (HAD-A/HAD-D), Pittsburgh sleep quality index (PSQI). Values are expressed as means +/- (standard deviation); trends are given between brackets.

2.2. *Polysomnography*

The recordings included at least three or more electroencephalograms recorded at least from Fp2-A1, C4-A1 and O2-A1 sites, two electro-oculograms, submental and bilateral anterior tibial electromyograms. Oral and nasal airflow were recorded by an oro-nasal cannula (Pro-Flow Plus™ Pro-Tech® Mukilteo, WA, USA), respiratory effort was measured by thoracic and abdominal belts (Pro-Tech® CT2™, Mukilteo, WA, USA). Capillary oxygen saturation was monitored by photosensitive finger-oxymetry (Nonin® Flexi-Form® II 7000A Nonin Medical Inc, Minneapolis, MN USA and LINOP® Adt Masimo corp. Irvine, CA, USA). All PSG recordings were analyzed on 21″ screens displaying 30 s polysomnographic epochs (Philips Respironics Inc™ Alice5®, Philips Healthcare™, Eindhoven, The Netherlands, European Union). Sleep Onset Latency (SOL) was defined as the time between Lights Out and the first 30 s epoch of sleep. Wake time did not include sleep latency (Wake after sleep onset, WASO). Sleep efficiency (SEI) was defined by the ratio of Total Sleep Time (TST) and Time In Bed (TIB). NREM (Non Rapid Eye Movement) sleep included sleep stages N1, N2 and N3 (or slow wave sleep, SWS). REM (Rapid Eye Movement) sleep latency (REMLAT) was defined as the time between sleep onset and the first epoch of REM sleep. An episode of sleep apnea was defined as a more than an 80% reduction in airflow for at least 10 s during sleep. A sleep hypopnea was defined as a 50% to 80% reduction of airflow amplitude accompanied by either a 3% or greater reduction in oxygen saturation or an arousal. Oxygen desaturation index (ODI) was defined by the number of 3% (or greater) drops of oxygen saturation per hour of sleep, as measured by photo-oxymetry. Arousals were defined according to the American Academy of Sleep Medicine (AASM) criteria (AASM, 2005). The arousal index (ArI) represented the number of arousals per hour of sleep. AASM criteria were used for sleep stage scoring. Group comparisons of descriptive PSG results are reported below (see Section 3 and Table 1).

The study was approved by the local ethical committee and conducted in accordance with the rules and regulations for the conduct of clinical trials stated by the World Medical Assembly in Helsinki.

2.3. *Rasch analyses and statistics*

The Rasch model family comprises generalized linear models and associated statistical procedures that connect observed questionnaire item responses, to an individual's location on a measurement dimension (Andrich, 1978). The parameters of this probabilistic model represent both the position of persons and items on a latent continuum based on a simple logic: respondents have a higher probability of endorsing “easier” items compared to “difficult” ones, and persons with higher “abilities” will have a higher probability of endorsing items than persons with lower “abilities” (Bond and Fox, 2007). Rasch analysis allows obtaining linear (additive) measures, qualified by standard errors and fit statistics from stochastic observations of ordered category responses. Total raw scores of persons and items are used to estimate these measures. The necessary and sufficient transformation of ordered qualitative observations into additive measures is a Rasch model. Under the conditions of this model, these measures are also item-distribution-free and person-distribution-free. This implies that these measures are statistically equivalent for items regardless of which persons (from the same population) are analyzed, and vice versa. The dichotomous Rasch model and its polytomous counterpart, the Andrich Rating Scale Model (ARSM) are members of the Rasch Model Family.

Formally, the polytomous ARSM (Andrich 1978) reads as: log $(\mathrm{P_{nij}}/\mathrm{P_{ni(j-1)}})=\mathrm{B_n}-\mathrm{D_i}-\mathrm{F_j}$, 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*,
- D_i is the “difficulty” measure of item *i*, 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 (Linacre, 2009).

With respect to the ESS and the FSS, each person's raw-score is modeled to be an accumulation of qualitative levels of respectively sleepiness and fatigue on the entire test. Then, from these raw-score amounts, the location on the latent variable is estimated and expressed on a linear (equal-interval) logarithmic scale. As such, the Rasch-calibration transforms raw scores into linear sleepiness and fatigue measures for each scale. Moreover, Rasch-modeled person ability or item difficulty measures are estimated within error, allowing them to be interpreted in the context of precision (Bond and Fox, 2007).

We assessed Rasch based rating scale diagnosis, dimensionalities of fatigue and sleepiness by means of Rasch based Factor Analysis and measurement invariance via differential item functioning (DIF). Convergent validity of fatigue and sleepiness, and with respect to sleep quality was assessed through Rasch based common person equating (CPE). Full details of the methodology and the mathematical basis of the Rasch approach were previously described in detail elsewhere (Bond and Fox, 2007; Neu et al., 2010b; Linacre, 2009). With exception of Rasch analyses, all statistics were computed using SPSS 22® (Industrial Business Machines© Inc., SPSS©, *Armonk, NY, USA*). The main outcomes from the Rasch analyses were carried out with Winsteps® for Windows (Winsteps 3.72.3©, SWREG® Inc., Minnetonka, MN, USA).

3. Results

3.1. Descriptives

In order to investigate potential gender differences and determine the strength of association between gender and sleep quality (as in whether being a poor ($PSQI > 5$) or a good sleeper ($PSQI \leq 5$)), we calculated the Lambda measure of association. Our results show a Lambda value of. 030 ($p = .843$), implying a weak, non-significant, association between sleep quality and gender. Therefore, gender was excluded from further analyses. A MANOVA with single-factor (group) was performed to compare age, body mass index (BMI), PSG data and symptom intensities. Demographics and PSG data did not show any significant differences between groups. Besides worse sleep quality (higher PSQI scores), poor sleepers showed higher levels of fatigue and affective symptoms, but not of sleepiness (Table 1). Explorative pairwise correlations showed significant relations between fatigue and sleepiness ($r = .214$, $p < 0.01$) and between fatigue and sleep quality impairment ($r = 0.426$, $p < 0.001$). Sleepiness on the other hand, was not significantly associated with sleep quality ($r = -0.089$, $p = 0.284$).

3.2. Rating scale diagnostics

Rating scale diagnosis showed that response categories of both instruments overlapped or were insufficiently spaced. The estimated endorsement rate for choosing one category over the other (step calibrations) should increase monotonically and be spaced around approximately 1.1 logits for a 4-point scale such as the ESS and around 0.6 logits for a 7-point scale such as the FSS (Linacre, 2009; Kim et al., 2010). Rating scale diagnostics of the Epworth Sleepiness Scale indicate a good fit of the rating scale with OUTFIT mean squares ranging from 0.80 to 1.13. Together with satisfactory fit indices, we observe an ordered rating scale with monotonically increasing step calibrations (n/a, −0.65, 0.08, 0.74). However, step calibrations between category 1 and 2 and 2 and 3 are less than expected (0.57 and 0.82 logits), implying that a lack of distinction between these categories may exist and that collapsing categories may be considered, especially regarding categories 1 and 2. However, both categories are not unacceptably underutilized, covering a reasonable length of the scale (see Fig. 1). Collapsing ESS categories seems therefore not to be compulsory for good measurement. Rating scale analysis of the Fatigue Severity Scale reveals an adequate fit with OUTFIT mean squares ranging from 0.83 to 1.29. However, as apparent from Fig. 1b, a number of categories never emerge as modal (category 2, 3, 4, 6) and step calibrations violate the rating scale assumptions (n/a, −0.66, 0.24, −0.25, −0.22, 0.55, 0.82). However, rescoring the category scale into a 4-point category scale (1234567→1223344) yields satisfactory OUTFIT indices ranging from 0.87 to 1.31 (see Fig. 1c). The recalibrated rating scale is now well ordered with well-spaced, monotonically increasing step calibrations (n/s, −1.33, 0.02, 1.31).

3.3. Item fit

Both calibrated instruments showed good fit of the data to the model, as evidenced by good person reliability (ESS: 0.81 and FSS: 0.87) and good item reliability (ESS: 0.99 and FSS: 0.94) . Rasch measures and fit statistics show good construct validity and unidimensionality (Table 2). Item 2 of the FSS has an OUTFIT mean square > 2, which indicates unexpected observations by persons or DIF on this item. Removing item 2 did not significantly affect person measures ($r = 0.98$) and was therefore included in the analysis.

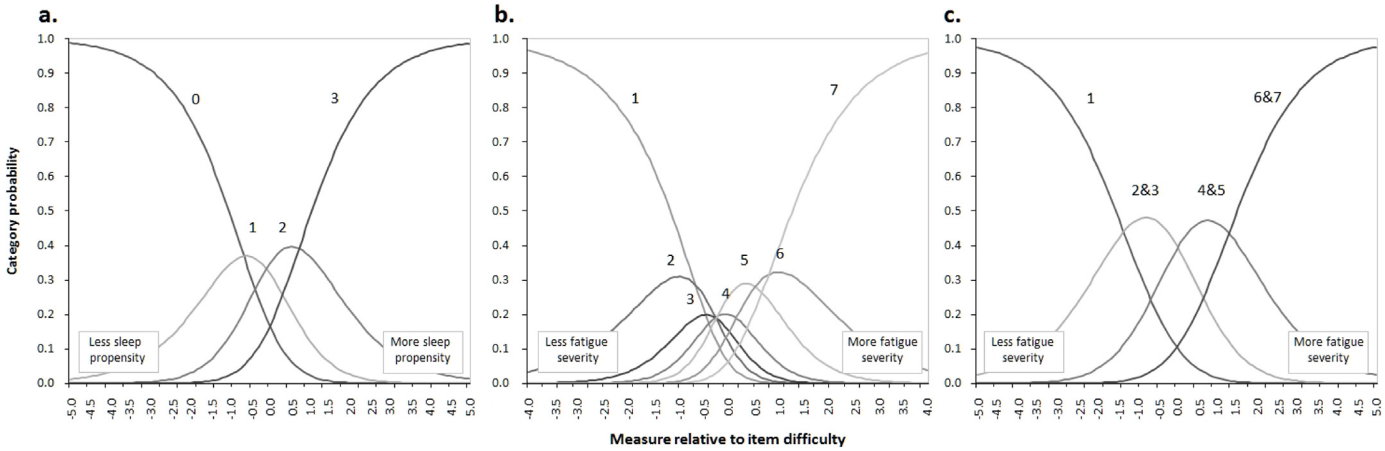


Fig. 1. Rating scale diagnostics. (a) Rating scale functioning of the ESS, (b) rating scale functioning of the 7 category FSS. Category probability curves of the 7-point scale show categories never emerging as modal and violated step calibrations assumptions. Collapsing categories into a 4-point scale (c) meets rating scale diagnostic criteria.

Table 2
Item statistics.

	Logits	SE	INFIT mnsq	OUTFIT mnsq
Epworth sleepiness scale: item content				
5. Lying down to rest in the afternoon when circumstances permit	−1.82	0.12	1.42	1.44
2. Watching TV	−1.12	0.11	1.16	1.37
4. As a passenger in a car for 1 h without a break	−0.43	0.10	1.01	1.00
1. Sitting and reading	−0.26	0.10	1.06	0.93
7. Sitting quietly after lunch without alcohol	0.00	0.10	0.88	0.89
3. Sitting inactively in a public place	0.47	0.11	0.88	0.95
8. In a car, while stopped for a few minutes in traffic	1.30	0.13	0.81	0.62
6. Sitting and talking to someone	1.86	0.16	1.04	0.74
Fatigue severity scale: item content				
1. My motivation is lower when I am fatigued	−0.80	0.13	1.44	1.66
3. I am easily fatigued	−0.56	0.13	0.76	0.65
4. Fatigue interferes with my physical functioning	−0.31	0.13	0.76	0.70
8. Fatigue is among my three most disabling symptoms	−0.17	0.13	0.68	0.59
9. Fatigue interferes with my work, family, or social life	−0.09	0.13	0.94	0.98
2. Exercise brings on my fatigue	**0.13**	**0.12**	**1.75**	**2.13**
6. My fatigue prevents sustained physical functioning	0.38	0.12	0.99	0.86
5. Fatigue causes frequent problems for me	0.69	0.12	0.81	0.72
7. Fatigue interferes with carrying out certain duties and responsibilities	0.75	0.12	0.86	0.82

Rasch measures are expressed in logits. Negative logits indicate ‘easier’ to endorse items and vice versa. INFIT and OUTFIT mean squares should be less than 2.

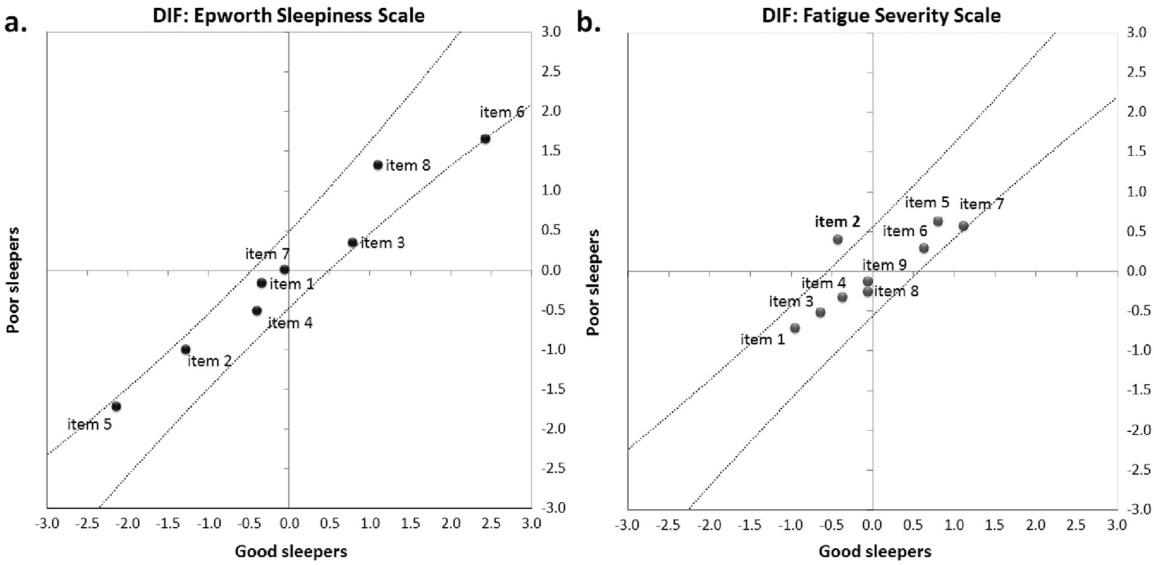


Fig. 2. Differential item functioning with respect to perceived sleep quality. Item difficulty estimates are expressed in logits. Dotted lines represent the 95% CI intervals. (a) Measurement invariance of the ESS with respect to perceived sleep quality (good sleepers vs poor sleepers). DIF analysis shows complete invariance and therefore no significant difference in item difficulty estimates between groups with respect to subjective sleep quality. (b) Measurement invariance of the FSS. Item 2 of the FSS falls outside the confidence limits indicating DIF with respect to sleep quality. Poor sleepers show higher than expected and good sleepers lower than expected difficulty estimates for item 2.

3.4. Dimensionality

Most items possess OUTFIT mean square values indicating unidimensionality (Table 2). However, for both the ESS and the FSS some values exceed the 1.4 mean square limit. Unidimensionality is then further investigated by means of Rasch Factor Analysis, a procedure used to identify possible common variance among the unmodeled data by the general Rasch sleepiness or fatigue dimensions. Eigenvalues of the secondary dimensions (contrasts) should be less than 3 (Linacre, 2009). The eigenvalue of the largest contrast for the ESS equals 1.6 and 2.1 for the FSS.

3.5. Differential item functioning

Measurement invariance in Rasch analysis is performed by investigating disagreement regarding item measures. When measures vary across groups by more than the modeled error, evidence of differential item functioning (DIF) is suspected. Results from the DIF analysis are shown in Fig. 2a and b. DIF analysis of the ESS shows complete invariance and therefore no significant difference in item difficulty estimates between groups with respect to subjective sleep quality. Item 2 of the FSS however, falls outside the confidence limits indicating DIF with respect to sleep quality (DIF item 2=0.81 logits, S.E.=0.27). Poor sleepers show higher than expected and good sleepers lower than expected difficulty estimates for item 2. However, we would expect for DIF to be significant in altering measurement if it exceeds 0.97 logits according to the following formula: $|\mathrm{DIF}| > 0.43$ logits+2*DIF S.E. (Zwick et al., 1999). As this is not the case, it confirms that removing item 2 does not significantly alter person measures and may therefore remain included in the final instrument.

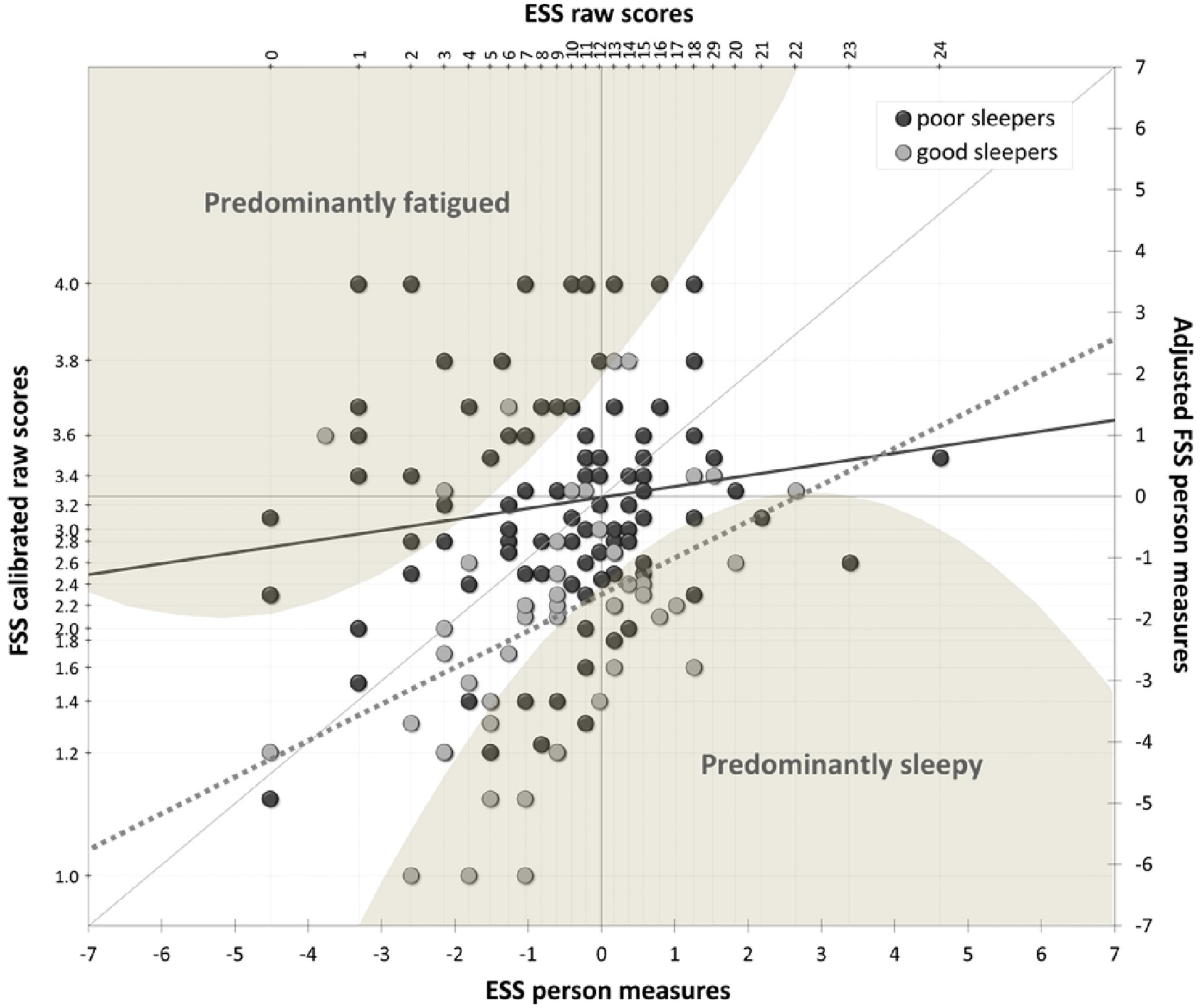


Fig. 3. Common person equating. Common Person Equating (CPE). Black dots represent poor sleepers' person measures, gray dots represent good sleeper's person measures. FSS and ESS person measures were criterion-referenced (adjusted for mean difficulty differences between instruments). The light gray line represents the identity line, i.e. maximal concordance between Rasch person estimates. The highlighted gray areas represent the areas outside the 95% CI limits of the Rasch person estimates. Persons falling outside the upper confidence limit are considered predominantly fatigued, persons falling outside the low confidence limit are considered predominantly sleepy. The full black line represents the person measure regression line for poor sleepers; the dotted gray line represents the person measure regression line for good sleepers.

3.6. *Common person equating*

CPE permits the assessment of convergent validity by cross-plotting Rasch person estimates for each instrument and investigating concordance between the measures (Fig. 3). Measures are criterion-referenced, meaning that the FSS average ability (0.66 logits) is equated to the ESS average ability (−0.69) by subtracting each individual FSS measure by the difference of the average abilities of both scales (1.35 logits). Person measures on the 45° diagonal (identity line) indicate perfect concordance. Under ideal circumstances of convergent validity, the empirical regression line should overlap with the identity line (slope=1 and intercept=0). Additionally, at least 95% of person measures should fall within confidence limits. As apparent in Fig. 3, a large proportion (66 out of 159 measures or 41.5%) of person measures fall outside of the 95% CI limits. The slope of the empirical regression line from the total sample (not shown in Fig. 3) equals 0.255 and the intercept −0.512 and differs significantly from the identity line (t=6.773, $p < 0.001$). This suggests that both instruments measure different but related constructs.

With respect to between group differences and in line with the MANOVA results, poor sleepers seem to experience higher fatigue levels than good sleepers, as further evidenced by the significant difference between the intercepts of individual regressions within each group depicted in Figs. 3 (t=−4.264, $p < 0.001$). Conversely, the relationships between respective fatigue and sleepiness intensities expressed by the group-specific regression slopes differ only marginally (t=1.807, p=0.0728). Both slopes however, deviate significantly from the identity line (respectively t=6.835, $p < 0.001$ and t=2.066, $p < 0.05$ for poor and good sleepers). Further analysis of the proportions of individuals falling outside the confidence limits (i.e. predominantly fatigued or sleepy), shows that predominantly fatigued individuals are mainly poor sleepers (30 poor sleepers vs. 3 good sleepers). However, the proportions of good sleepers (n=17) and poor sleepers (n=16) are similar within the predominantly sleepy group (p=0.0003; Fisher's Exact Test, two-tailed). Regarding differences in convergent validity, we find similar proportions of good sleepers (20 individuals or 42.6% of poor sleepers) and poor sleepers (46 individuals or 45.1% of poor sleepers) falling outside confidence limits (z=0.03, p=0.775). Taken together, these results suggest that despite differences in fatigue symptom intensities between poor and good sleepers, the convergence of fatigue and sleepiness as constructs is similar between groups.

4. Discussion

The need to differentiate between fatigue and sleepiness stems from the fact that both have a distinct physiological and/or pathophysiological basis but often remain used interchangeably by patients, clinicians and researchers (Shen et al., 2006a; Neu et al., 2010a). Aside from semantics, widely used instruments aimed at measuring fatigue and sleepiness (i.e. the Fatigue Severity Scale (FSS) and the Epworth Sleepiness Scale (ESS) respectively) seem to be significantly correlated in various conditions (Hossain et al., 2005, Shen et al., 2006b, Valko et al., 2010) and in the general population (Neu et al., 2010b). These findings suggest that, to disentangle between both constructs, a psychometric evaluation of their respective overlap is also in order. The purpose of this study was to evaluate the psychometric properties of the ESS and the FSS and their potential overlap using Rasch methodology. Additionally, we assessed if both instruments were invariant to sleep quality with respect to measurement properties and convergent validity.

Overall, our results show that both scales appear to be reliable instruments as indicated by Rasch reliability statistics (Linacre, 2009). With respect to rating scale functioning, results indicate that the middle categories of the ESS are insufficiently spaced along the underlying latent variable. The rating scale seems nevertheless to function adequately as each category emerges as modal and step calibrations increase monotonically. Participants are thus able to discern four distinct levels of dozing probability. Insufficient space between category thresholds simply indicate that the probability of dozing is perceived less frequently as “slight” or “moderate” or than “no” or “high”, regardless of which ESS item is presented. Collapsing categories seemed therefore unnecessary. In line with previous studies (Neu et al., 2010b; Hagell and Broman, 2007) these findings suggest that the original 4-category rating structure of the ESS is sufficiently adequate to obtain good measurement. The original 7-category rating structure of the FSS did not function effectively. Four categories never emerge as modal, step calibrations are insufficiently spaced and do not increase monotonically. This implies that respondents might not be able to discern 7 distinct levels of fatigue. To meet rating scale assumptions, a 4-category structure was proposed consisting in keeping category 1 and collapsing every second category with the previous uncollapsed adjacent category (1234567 → 1223344 or “strongly disagree”, “disagree”, “agree” and “strongly agree”). This category structure appears to be preferable to the original in order to obtain linear measurement. Regarding item functioning, fit statistics indicate good construct validity and undimensionality of the respective instruments. Yet, item 2 of the FSS “Exercise brings on my fatigue” displays OUTFIT statistics above 2, indicating too much variation in the responses (underfit). This may reflect individuals agreeing that exercise brings on fatigue while their Rasch-modeled fatigue measure predicts they would rather disagree. While some items display OUTFIT mean square values exceeding 1.4 indicating possible multidimensionality, Rasch Factor Analysis confirmed than both instruments measure unidimensional constructs. These results are thus in line with our previous findings in a general population sample (Neu et al., 2010b).

Regarding the stability of the item structure of the ESS and the FSS, we found complete invariance to sleep quality for the ESS and little evidence of differential item functioning in item 2 of the FSS. Poor sleepers show higher than expected and good sleepers lower than expected difficulty estimates for this item. However, the amount of DIF is smaller than expected considering the S.E. of measurement of the DIF estimates for the FSS. As such, we may also consider the FSS to be invariant to perceived sleep quality.

Our next step was to investigate the convergent validity of the ESS and the FSS, and to assess its similarity across groups of individuals with or without significant sleep quality impairment. We achieved this by applying an equating procedure based on Rasch person estimates (CPE): both the FSS and ESS administered to the same sample of respondents are Rasch-analyzed separately and person measures are cross-plotted after equating for differences in mean difficulty in both instruments. By quantifying the amount of person measures falling within 95% confidence limits and by evaluating regression statistics, inferences about convergent validity can be made. Assuming that fatigue and sleepiness are often confused constructs by patients and clinicians alike (Shen et al., 2006a; Neu et al., 2010a), we hypothesized that our patients should have similar Rasch-derived person estimates on the ESS and the FSS. Consequently, instruments assessing levels of fatigue and sleepiness should yield high levels of convergent validity. Our results however, show that more than 40% of estimates fall outside confidence limits and that the slope of the empirical regression significantly differs from the identity line. These results suggest that, while both instruments are related to each other, as also evidenced by a significant correlation, the ESS and the FSS measure distinct constructs of sleepiness and fatigue inasmuch as the ESS and FSS represent these concepts, respectively and accurately.

Although comparable, our results in a general population sample were less distinctive (Neu et al., 2010b). There, about 80% of Rasch person estimates fell within the 95% confidence limits, compared to less than 60% in the present clinical sample. Taken together, this suggests that the overlap of perceived sleepiness and fatigue might be different whether individuals experience good or poor sleep. Therefore, we subdivided our sample of patients in good and poor sleepers according to PSQI thresholds. Remarkably, we found no evidence of differential convergent validity across groups, as similar proportions of poor and good sleepers fall outside confidence limits. These results mainly imply two things: (1) poor and good sleepers alike may experience either predominantly fatigue or sleepiness symptoms and in about 60% of the cases, both symptoms overlap at equivalent intensities; and/or (2) discriminating between feelings of fatigue and sleepiness is independent from sleep quality.

What further appears from our analysis is that predominantly fatigued individuals (i.e. individuals with person estimates of fatigue higher than sleepiness and falling outside confidence limits) are mainly patients experiencing poor sleep, whereas the proportion of good and poor sleepers is similar in predominantly sleepy individuals. With respect to group differences in regression parameters, this is further evidenced by the significantly higher intercept of the poor sleeper group (Fig. 3). Indeed, patients with complaints of invalidating daytime fatigue usually tend to complain about their sleep quality or present with non-restorative sleep complaints (Neu et al., 2010a; Guilleminault et al., 2006; Neu et al., 2011). In the present study as well, altered perceived sleep quality is associated with subjective daytime fatigue, but not with sleepiness, and this regardless of specific findings in PSG derived data (Neu et al., 2010b; 2011). The PSQI and the ESS can even measure orthogonal dimensions of sleep-wake symptoms that are not or weakly related to objective sleep measures, whether in a community sample (Buysse et al., 2008) or in a clinical sample (Ferreira et al., 2006). This may come as no surprise, given that sleep architecture, sleep stage durations and sleep efficiency are global concepts with limited physiological meaning. Hence, self-reported good sleepers, may present with objective sleep variables (sleep fragmentation, respiratory disturbance e.g.) similar to poor sleepers. It has previously been mentioned that classical PSG variables may not be able to express potential underlying disturbances involved in the non-recovering sensations of reported poor sleep quality (Neu et al., 2007). Consequently, the present findings raise questions about what influences complex perceptions like sleep quality and which dimensional parts of it are within the effective reach of standard polysomnography (Buysse et al., 2008). Needless to say, in order to provide the most adequate care within a clinical routine setting, we cannot rely on PSG data alone but must deal with patients' complaints and try to unequivocally understand reported symptoms. The latter is inherent to sound clinical decision making and therefore of major importance in selecting appropriate treatment strategies.

The cross-plot of raw scale-scores and calibrated measures displaying margins wherein both constructs overlap (Fig. 3), allows clinicians to determine which condition significantly prevails over the other. For clinical purposes, we provide a blank version of this concordance diagram using raw scale scores that can be used as a simple tool to sharpen the psychometric differentiation of both constructs (Appendix A).

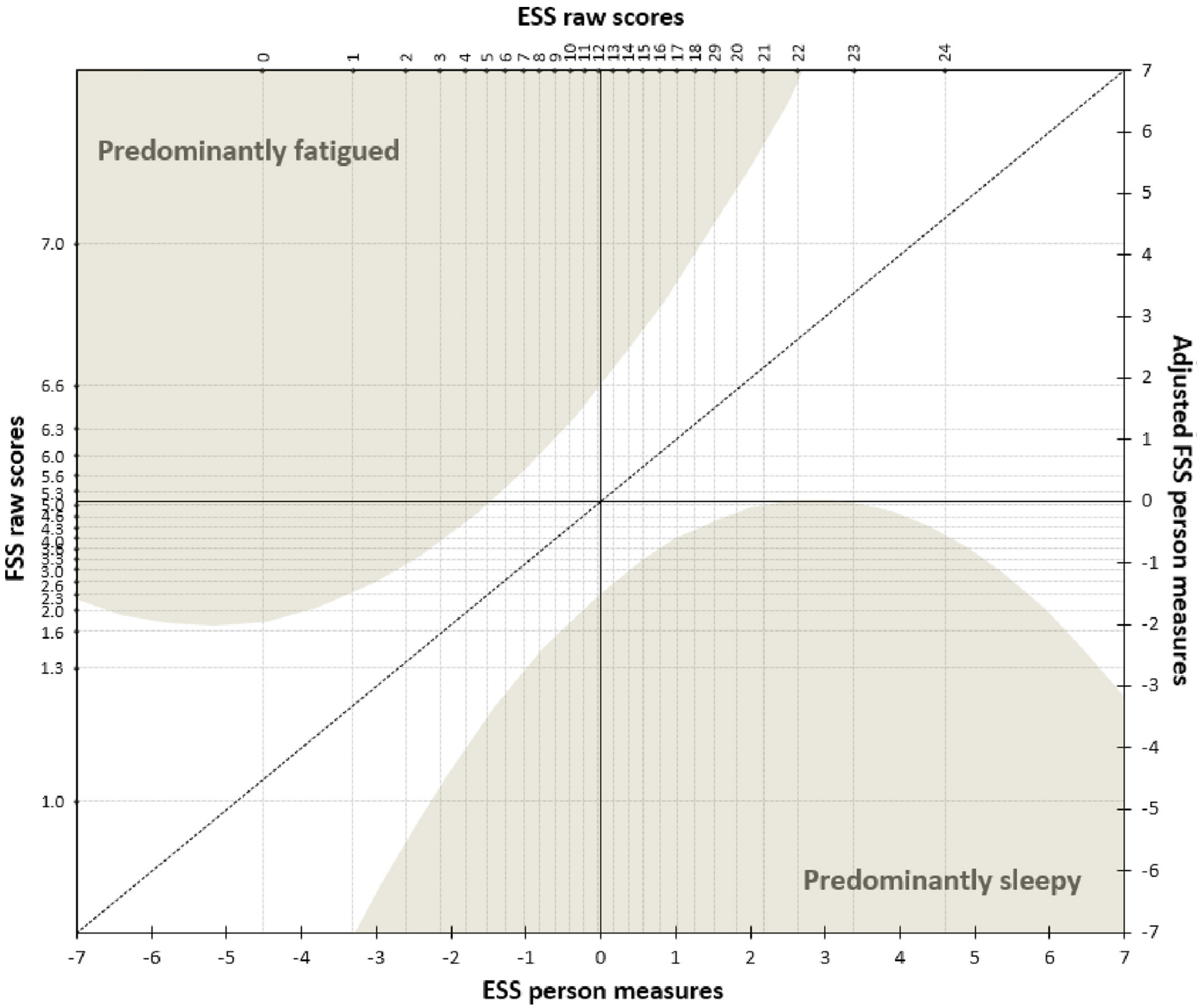
In summary, our results show that both the ESS and FSS are reliable and valid, unidimensional linear instruments to measure subjective sleepiness and fatigue respectively, considering proper calibration. The instruments display limited convergent validity suggesting that they measure distinct, but related constructs. Additionally, the ESS and the FSS appear to be invariant to perceived sleep quality, with respect to item hierarchy and convergent validity. Regarding group differences, we also found that predominantly fatigued individuals in particular experience impaired sleep quality, however individuals experiencing poor or good sleep quality can be found equally in predominantly sleepy individuals. Whether these dissimilarities mainly depend on differences between symptom intensities or are a result of individual discrimination abilities remains to be determined.

Acknowledgments

We would like to thank J. Michael Linacre for his constructive assistance and valuable advice.

Appendix A

Concordance diagram based on raw scale scores



Instruction:

1) ESS raw score: add all 8 item scores together.
2) FSS raw mean score: add all 9 item scores together and divide by 9.
3) Plot the ESS score on the X-axis and the FSS mean score on the Y-axis.

References

American Academy of Sleep Medicine (AASM), 2005. International Classification of Sleep Disorders (ICSD), 2nd edition: Diagnostic and Coding Manual. Westchester, IL: AASM.

American Psychiatric Association (APA), 2000. Diagnostic and Statistical Manual of Mental disorders, Text Revision (DSM-IV-TR), 4th ed. American Psychiatric Association, Washington DC.

Andrich, D., 1978. A rating formulation for ordered response categories. Psychometrika 43, 561–573.

Bond, T.G., Fox, C.M., 2007. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed. NJ: Lawrence Erlbaum, Mahawah.

Borbély, A.A., 1982. A two process model of sleep regulation. Human Neurobiology 1, 195–204.

Braley, T.J., Chervin, R.D., Segal, B.M., 2012. Fatigue, tiredness, lack of energy, and sleepiness in multiple sclerosis patients referred for clinical polysomnography. Mult. Scler. Int. 2012, 673936.

Buysse, D.J., Hall, M.L., Strollo, P.J., Kamarck, T.W., Owens, J., Lee, L., Reis, S.E., Matthews, K.A., 2008. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and Clinical/Polysomnographic Measures in a Community Sample. J. Clin. Sleep Med. 4, 563–571.

Buysse, D.J., Reynolds, C.F., Monk, T.H., Berman, S.R., Kupfer, D.J., 1989. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. 28, 193–213.

Ferreira, J.J., Desboeuf, K., Galitzky, M., Thalamas, C., Brefel-Courbon, C., Fabre, N., Senard, J.M., Montastruc, J.L., Sampaio, C., Rascol, O., 2006. Sleep disruption, daytime somnolence and 'sleep attacks' in Parkinson's disease: a clinical survey in PD patients and age-matched healthy volunteers. Eur. J. Neurol. 13, 209–214.

Guilleminault, C., Poyares, D., Rosa, A., Kirisoglu, C., Almeida, T., Lopes, M.C., 2006. Chronic fatigue, unrefreshing sleep and nocturnal polysomnography. Sleep Med. 7, 513–520.

Hagell, P., Broman, J.E., 2007. Measurement properties and hierarchical item structure of the Epworth Sleepiness Scale in Parkinson's disease. J. Sleep Res. 16, 102–109.

Hagell, P., Höglund, A., Reimer, J., Eriksson, B., Knutsson, I., Widner, H., Cella, D., 2006. Measuring fatigue in Parkinson's disease: a psychometric study of two brief generic fatigue questionnaires. J. Pain Symptom Manag. 32, 420–432.

Hossain, J.L., Ahmad, P., Reinish, L.W., Kayumov, L., Hossain, N.K., Shapiro, C.M., 2005. Subjective fatigue and subjective sleepiness: two independent consequences of sleep disorders? J. Sleep Res. 14, 245–253.

Johns, M.W., 1991. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. Sleep 14, 540–545.

Johns, M.W., 1998. Rethinking the assessment of sleepiness. Sleep Med. Rev. 2, 3–15.

Johns, M.W., 2002. Sleep propensity varies with behaviour and the situation in which it is measured: the concept of somnificity. J. Sleep Res. 11, 61–67.

Kim, D.H., Wang, C., Ng, K.M., 2010. A Rasch Rating Scale Modeling of the Schutte self-report emotional intelligence scale in a sample of international students. Assessment 17, 484–496.

Krupp, L.B., La Rocca, N.G., Muir-Nash, J., Steinberg, A.D., 1989. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus [erythematosus]. Arch. Neurol. 46, 1121–1123.

Lane, R.D., Carmichael, C., Reis, H.T., 2011. Differentiation in the momentary rating of somatic symptoms covaries with trait emotional awareness in patients at risk for sudden cardiac death. Psychosom. Med. 73, 185–192.

Lerdal, A., Wahl, A., Rustøen, T., Hanestad, B.R., Moum, T., 2005. 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 33, 123–130.

Linacre, J.M., 2009. User's Guide and Program Manual to WINSTEPS: Rasch Model Computer Programs. IL: MESA Press, Chicago.

Mairesse, O., Hofmans, J., Neu, D., de Oliveira, A.L.D.M., Cluydts, R., Theuns, P., 2010. The algebra of sleepiness: investigating the interaction of homeostatic (S) and circadian (C) processes in sleepiness using linear metrics. Psicológica 31, 541–559.

Neu, D., Kajosch, H., Peigneux, P., Verbanck, P., Linkowski, P., Le Bon, O., 2011. Cognitive impairment in fatigue and sleepiness associated conditions. Psychiatry Res. 189, 128–134.

Neu, D., Linkowski, P., Le Bon, O., 2010a. Clinical complaints of daytime sleepiness and fatigue: How to distinguish and treat them, especially when they are"-excessive' or"chronic'? Acta Neurol. Belg. 110, 15–25.

Neu, D., Mairesse, O., Hoffmann, G., Dris, A., Lambrecht, L.J., Linkowski, P., Verbanck, P., Le Bon, O., 2007. Sleep quality perception in the chronic fatigue syndrome. Correlations with sleep efficiency, affective symptoms and intensity of fatigue. Neuropsychobiology 56, 40–46.

Neu, D., Mairesse, O., Hoffmann, G., Valsamis, J.B., Verbanck, P., Linkowski, P., Le Bon, O., 2010b. Do 'sleepy' and 'tired' go together? Rasch analysis of the relationships between sleepiness, fatigue and nonrestorative sleep complaints in a nonclinical population sample. Neuroepidemiology 35, 1–11.

Ohayon, M.M., 2008. From wakefulness to excessive sleepiness: what we know and still need to know. Sleep Med. Rev. 12, 129–141.

Olson, L.G., Ambrogetti, A., Sutherland, D.C., 2003. A pilot randomized controlled trial of dexamphetamine in patients with chronic fatigue syndrome. Psychosomatics 44, 38–43.

Rasch, G., 1980. Probabilistic models for some intelligence and attainment tests (Copenhagen, Danish Institute for Educational Research), Expanded Edition. The University of Chicago Press, Chicago, IL.

Shen, J., Barbera, J., Shapiro, C.M., 2006a. Distinguishing sleepiness and fatigue: focus on definition and measurement. Sleep Med. Rev. 10, 63–76.

Shen, J., Botly, L.C., Chung, S.A., Gibbs, A.L., Sabanadzovic, S., Shapiro, C.M., 2006b. Fatigue and shift work. J. Sleep Res. 15, 1–5.

Valko, P.O., Waldvogel, D., Weller, M., Bassetti, C.L., Held, U., Baumann, C.R., 2010. Fatigue and excessive daytime sleepiness in idiopathic Parkinson's disease differently correlate with motor symptoms, depression and dopaminergic treatment. Eur. J. Neurol. 17, 1428–1436.

Yu, C.H., Popp, S.E., 2005. Test equating by common items and common subjects: concepts and applications. Practical assessment. Res. Eval. 10, 1–19.

Zwick, R., Thayer, D.T., Lewis, C., 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. J. Educ. Meas. 36, 1–28.