

# Repeating administration of Epworth Sleepiness Scale is clinically useful

Denis Martinez · Tiago Cataldo Breitenbach · Magali Santos Lumertz ·  
Denise Lérias Alcântara · Neusa Sicca da Rocha · Cristiane Maria Cassol ·  
Maria do Carmo Sfreddo Lenz

Received: 10 August 2010 / Revised: 20 September 2010 / Accepted: 22 October 2010 / Published online: 10 November 2010  
© Springer-Verlag 2010

## Abstract

**Purpose** We aimed to verify whether it is clinically useful to repeat the Epworth Sleepiness Scale (ESS) in individuals with suspected sleep-disordered breathing (SDB).

**Methods** In this cross-sectional, prospective study, results of the repeated administration of the ESS were analyzed. In 929 consecutive patients, ESS was obtained as usual in the laboratory routine, immediately before the sleep study (ESS1) and was repeated in the morning, after the polysomnography (ESS2). ROC curve, classical psychometry, and item response theory (IRT) Rasch analysis were used to assess measurement properties of ESS.

**Results** The ESS1 score was (mean  $\pm$  SD),  $11 \pm 5.1$ , and the ESS2,  $13 \pm 4.7$  ( $p < 0.001$ ). Change in ESS score is

explained in multivariate analysis by age, but not by gender, educational level, change in Stanford Sleepiness Scale, symptoms of sleep disorder, or polysomnography results. Accuracy of an ESS score  $>10$  to predict apnea–hypopnea index  $\geq 5$  increased from 56% (ESS1) to 72% (ESS2). IRT psychometric properties (unidimensionality, invariance, local independence) were maintained in ESS2. **Conclusions** Repeating the administration of the Epworth Sleepiness Scale in a clinical setting increases its score and diagnostic accuracy and correlation with SDB variables, without changing the psychometric properties of the scale. This experiment indicates the clinical usefulness of repeating the ESS. The scale can be repeated at a negligible cost, before dismissing individual patients on the basis of a low

The project was supported partly by a research grant from institutional agency FIPE-HCPA. Magali Lumertz, a researcher, was supported by a government student grant from governmental agency FAPERGS. Tiago Breitenbach was supported by a government student grant from governmental agency BIC-CNPq-UFRGS, and Cristiane Maria Cassol was supported by government student grant from governmental agency CAPES.

D. Martinez (✉)

Division of Cardiology, Hospital de Clínicas de Porto Alegre (HCPA), Sleep Clinic, Universidade Federal do Rio Grande do Sul (UFRGS),  
Rua Ramiro Barcelos, 2350,  
Porto Alegre, RS, Brazil 90035-903  
e-mail: dm@ufrgs.br

T. Cataldo Breitenbach · N. Sicca da Rocha · C. M. Cassol  
Hospital de Clínicas de Porto Alegre (HCPA), Universidade  
Federal do Rio Grande do Sul (UFRGS),  
Porto Alegre, Brazil

T. Cataldo Breitenbach  
e-mail: tbreitenbach@hcpa.ufrgs.br

N. Sicca da Rocha  
e-mail: neusa-rocha@via-rs.net

C. M. Cassol  
e-mail: ccassol@hcpa.ufrgs.br

M. Santos Lumertz  
Universidade Federal de Ciências da Saúde de Porto Alegre  
(UFCSA),  
Porto Alegre, Brazil  
e-mail: magali.lumertz@gmail.com

D. Lérias Alcântara · M. d. C. Sfreddo Lenz  
Sleep Clinic,  
Porto Alegre, Brazil

D. Lérias Alcântara  
e-mail: dla@portoweb.com.br

M. d. C. Sfreddo Lenz  
e-mail: maria@sono.com.br

ESS score, discontinuing a potentially lifesaving diagnostic and therapeutic process.

**Key words** Epworth Sleepiness Scale · Hypersomnolence · Sleep apnea · Polysomnography · Item response theory · Sleep-disordered breathing

#### Abbreviations used in manuscript

AHI	Apnea–hypopnea index
$\chi^2$	Chi-square
DIF	Differential item functioning
ESS	Epworth Sleepiness Scale
ESS1	ESS administered at evening, before polysomnography
ESS2	ESS administered in the morning, after polysomnography
IRT	Item response theory
OSAHS	Obstructive sleep apnea–hypopnea syndrome
PSG	Polysomnography
SaO <sub>2</sub>	Arterial oxygen saturation
SDB	Sleep-disordered breathing

#### Introduction

Diagnosis of obstructive sleep apnea–hypopnea syndrome (OSAHS) combines an apnea–hypopnea index (AHI) of five or more events per hour of sleep and a symptom of sleep disorder, usually excessive daytime sleepiness [1–3]. Despite the importance of this symptom in the OSAHS, there is considerable controversy regarding the reliability of methods used for sleepiness estimation [4]. Measuring sleepiness is complicated by its multiple dimensions [5–7]. The Epworth Sleepiness Scale (ESS) [8] became a widely used tool to help in the OSAHS diagnosis [9–17], besides detecting sleepiness in Parkinson’s disease [18], mental disorders [19], and insomnia [20].

Miletin and Hanly [21] expressed concern about ESS being used to measure change over time since its test–retest properties had not been properly assessed. The possibility exists that a change is reflecting recall bias instead of actual clinical change. Although Johns applied the ESS to medical students and repeated the scale 5 months later [22], Chervin [23] considered that test–retest reliability has not been adequately established in clinical samples. Psychometric inconsistencies in the factor analysis were reported by Smith et al. [24].

Rasch analysis, originally developed in education, has been employed to assess psychometric properties of questionnaires used for medical diagnosis [25] according with the item response theory (IRT). Three properties are important in IRT: unidimensionality, invariance, local

independence. Unidimensionality means that the items of an additive scale are correlated and represent a single construct or concept. Invariance means that person factors, like age and gender, do not affect the way the instrument is answered; therefore, men or women, young or elderly should have the same performance in the test. Local independence means that the probability of success on an item depends only on item parameters and on testee ability [26].

Retest may modify the score of a scale [27, 28]. It is difficult to ascertain whether, in the clinical setting, the repetition of a test is useful. Many of the studies repeating the ESS involve small number of cases. Considering the widespread use of the ESS [29], considering the doubts still being raised about its reproducibility [21, 23], we decided to verify, in a large number of patients with suspected sleep-disordered breathing (SDB), whether the score and psychometric properties of ESS change when subjects are tested before and retested after the polysomnography (PSG).

#### Methods

This is a cross-sectional, prospective, descriptive study, in which the intervention was to duplicate the ESS administration before and after the PSG. The study included patients attending a university-affiliated sleep clinic with suspected OSAHS who gave informed consent to the use of their questionnaires, in terms approved by the institutional ethics committee. The inclusion criteria were: age 18 years and over, a full-night, baseline, intervention-free polysomnography, and thoroughly filled ESSs in the evening, on arrival at the sleep laboratory (ESS1) and in the morning, after the PSG (ESS2).

Both at the first and second administrations of the ESS, patients were allowed to interact with the technicians to clarify doubts on filling the questionnaires, as is inherent to sleep laboratory routine, avoiding experimental non-generalizable situations. Patient–technician communication tends to be more concentrated at the evening, during electrodes placement than in the morning when people fill the forms rapidly with rare questions before leaving the laboratory. Occasionally, patients were exposed to posters and handouts concerning sleep disorders, but the exposition occurred mostly at the reception, before replying to ESS1. Questions were divided in soporificity levels, based on the average score of each question in ESS1, as described by Johns [22]. The mean of the scores of the three or two questions in each soporificity category (most, intermediate, and least soporific), were rounded so that the mean scores were integers from 0 to 3. We considered an “inappropriate” answer to the ESS when a patient responded to one of the most soporific situations with a score lower than the

score used in intermediate or least soporific situations. Stanford Sleepiness Scale (SSS) was administered before (SSS1) and after PSG (SSS2); the difference SSS1–SSS2 (diffSSS) was calculated to estimate the reduction in sleepiness obtained by each patient after the night's sleep. Reliable change index, a widely used tool to establish the significance of a change in a test administered in two occasions, was calculated as described before [30, 31].

### Polysomnography

The PSG was carried out according to the standard method as described before [32], with electroencephalogram (EEG; C3-A2, C4-A1), electrooculogram (left eye and right eye), submental and anterior tibial electromyogram, and electrocardiogram. The airflow was measured by nasal cannula attached to a pressure transducer; respiratory effort was detected by respiratory inductance plethysmography (QRIP, Braebon, Kanata, Canada) the arterial oxygen saturation (SaO<sub>2</sub>), by pulse oximeter (Xpod, Nonin, Plymouth, Minnesota, USA).

Apneas were defined as the reduction in the baseline airflow to 10% or less for 10 s or longer; hypopneas were defined as reductions in the airflow by 50% or more associated with a 3-s arousal or SaO<sub>2</sub> fall by 3% or more. The AHI was calculated by dividing the total apneas and hypopneas by number of hours of sleep.

### Rasch analysis

The IRT psychometric properties analyzed were unidimensionality, invariance, and local independence. Because all items in the ESS are polytomous, they were verified to certify whether their response categories were ordered. The boundaries between categories, called thresholds, were checked for “disordered thresholds”, indicating the need to collapse adjacent categories. Following this, data were fitted to the model to determine overall fit, and how well each item suited the model. Three overall statistics were considered to determine model fit. Two were item–person interaction statistics distributed as a Z statistic with mean of 0 and standard deviation of 1, indicating perfect fit to the model. A third was an item–trait interaction statistic reported as chi-square (chi<sup>2</sup>), reflecting the invariance across the trait, indicated by a non-significant chi<sup>2</sup>. Besides, individual item–fit statistics are presented, as residuals (acceptable within the range ±2.5) and as a chi<sup>2</sup> statistic (a non-significant chi<sup>2</sup> is required).

Differential item functioning (DIF) analysis tests the invariance of the ESS, in such a way that individuals with identical sleepiness scores had the same probability of responding uniformly to an item, irrespective of gender, age, etc. All items were checked for DIF by gender, age

(split at median of 40 years), educational level (with or without college education), and AHI (split at five events per hour) as person factors. The item logistic function of DIF analysis, i.e., the proportion of individuals at the same sleepiness level with the same answer to a given item, was considered violating the requirement for unidimensionality and invariability if items did not yield the same item response function for two or more ESS score groups. ANOVA was used for detecting DIF of the person–item deviation residuals with person (age, gender, etc.) and class intervals (e.g., group along the ESS range) as factors. We identified uniform and non-uniform DIF where, in the former, the difference between groups in the probability of affirming an item across the trait was constant and, in the latter, the difference varies across the trait.

Lastly, we considered the items that had the highest correlation with the first principal component in both directions as the set most likely to breach the assumption of local independence. The person estimates of the total fitting item set against the person estimates of the item set for the positively loading items was compared to the person estimates for the total fitting item set against the person estimates of the item set for the negatively loading items. The result is an indication of unidimensionality; being the paired *t* test of this comparison non-significant, at  $p > 0.05$  level, the items were considered unidimensional. Rasch analysis was performed using the RUMM 2020 package (Rumm Laboratory Pty Ltd, Perth, Australia).

### Statistical analysis

Data was analyzed with the program SPSS v.16 (SPSS, Chicago, USA). The results are described as percentages, mean and standard deviation, or median and interquartile interval 25–75%. The statistical tests used—Student's *t* test, chi<sup>2</sup>, Cronbach's alpha, Spearman's rho, and Pearson's correlation—are specified in the description of results. Factor analysis with varimax rotation was used with AHI > 5 as selection variable. Area under the ROC curve was utilized to estimate the diagnostic accuracy of each question and of ESS1 and ESS2 to predict AHI ≥ 5, AHI ≥ 15, AHI ≥ 30. Bivariate correlations and multiple regression analysis were used to assess the relationship between ESS2–ESS1 difference (DiffESS) and the variables gender, age, educational level, diffSSS, and AHI. Due to the asymmetry of the data distribution, natural logarithm transformation of AHI (lnAHI) was employed in regressions.

### Results

We obtained data from 929 patients (Table 1). Educational level was elementary school, in 19.1% of the cases; high

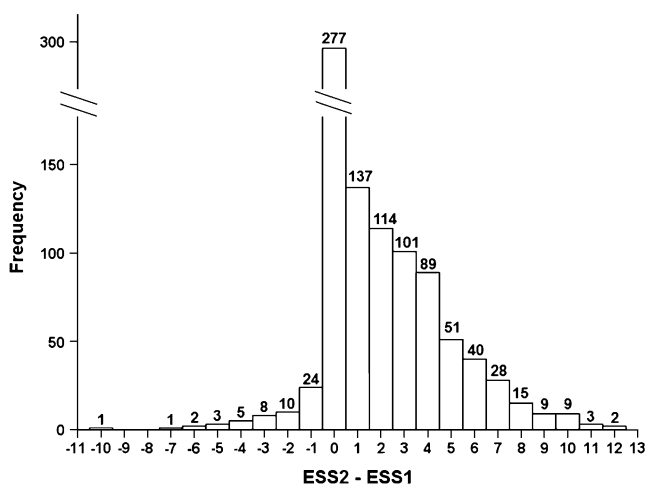
**Table 1** Means, standard deviations, medians, and interquartile range of the anthropometric and polysomnographic characteristics of the sample

Variables	Mean $\pm$ SD	Median [IR]
Male gender [ <i>n</i> (%)]	590 (63.5%)	
Age (years)	46 $\pm$ 14	46 [37–56]
Body mass index (kg/m <sup>2</sup> )	27 $\pm$ 5.3	26 [24–31]
Time in bed (min)	476 $\pm$ 33	480 [463–493]
Sleep efficiency (%)	83 $\pm$ 12	86 [80–92]
Stage N1 latency (min)	17 $\pm$ 22	9 [4–18]
Stage N2 latency (min)	25 $\pm$ 28	16 [8.2–27]
Stage REM latency (min)	112 $\pm$ 70	92 [63–145]
Stage N3 (%)	14 $\pm$ 6.4	15 [9–18]
Stage REM (%)	15 $\pm$ 6	15 [11–19]
Apnea–hypopnea index (AH/h)	24 $\pm$ 22	17 [7–41]
Lowest SaO <sub>2</sub> (%)	82 $\pm$ 8.8	84 [75–88]
Evening Stanford Sleepiness Scale	2.6 $\pm$ 1.3	2 [2–3]
Morning Stanford Sleepiness Scale	2.4 $\pm$ 1.4	2 [1–3]
Evening Epworth Sleepiness Scale	10 $\pm$ 5.1	10 [7–14]
Morning Epworth Sleepiness Scale	13 $\pm$ 4.6	13 [10–16]
Difference in Epworth Sleepiness Scale	2.3 $\pm$ 2.8	2 [0–4]
Difference in Stanford Sleepiness Scale	0.07 $\pm$ 1.6	0.00 [0–1]

SD standard deviation, IR 25–75% interquartile range, AH total apnea–hypopnea, SaO<sub>2</sub> arterial oxygen saturation

school degree, in 47.1%; and college degree, in 31.4%; 2.4% were health professionals. OSAHS was mild in 24.7%, moderate in 21.8%, and severe in 36.3% of the patients. ESS2 increased in 64.4%, remained the same in 29.8%, and diminished in 5.8% of cases (Fig. 1).

The means of the scores for each of the ESS items and the total ESS score in the morning, ESS2, were significantly greater than the score in the evening, ESS1 (Table 2, Fig. 2). Disagreement between ESS1 and ESS2 is greater for items exploring highly soporific situations as in questions 5 and 7.

**Fig. 1** Histogram of the difference in Epworth Sleepiness Scale (ESS) scores, before (ESS1) and after (ESS2) polysomnography

The Bland–Altman plot depicts the reproducibility of the ESS in Fig. 2. DiffESS was not significantly correlated with gender, diffSSS, educational level, and AHI; when all these regressors were included in a multiple regression model, the only significant variable was age.

In Fig. 3, the percentage of ESS scores >10 changed from 37 to 49 (12 percent points) in the group with AHI<5 and more than 20 points in the groups with AHI $\geq$ 5. The average ESS score in cases with AHI<5 increased from 9.5 to 11.2 (1.7 points) in the group with AHI<5 and more than 2 points in the groups with AHI $\geq$ 5.

Diagnostic performance of ESS1>10 and ESS2>10 to predict AHI $\geq$ 5 is displayed in Table 3. Several properties, such as sensitivity, specificity, accuracy changed significantly from ESS1 to ESS2.

Performance of ESS1 and ESS2 to predict AHI $\geq$ 5, AHI $\geq$ 15, and AHI $\geq$ 30, assessed by the area under the ROC curve, is displayed in Table 4. Area under the ROC curve was unchanged or changed non-significantly from ESS1 to ESS2 for all questions. The best performance is seen in the prediction of AHI $\geq$ 5. In the prediction of AHI $\geq$ 30, questions 3, 5, and 8 displayed a small non-significant reduction in area.

The test–retest reliability of the ESS, assessed using Chronbach’s alpha, was 0.92 ( $p=0.00001$ ) with single measures intraclass correlation of 0.85. Kappa coefficients for each question are displayed in Table 5. Question 5 has the lowest kappa.

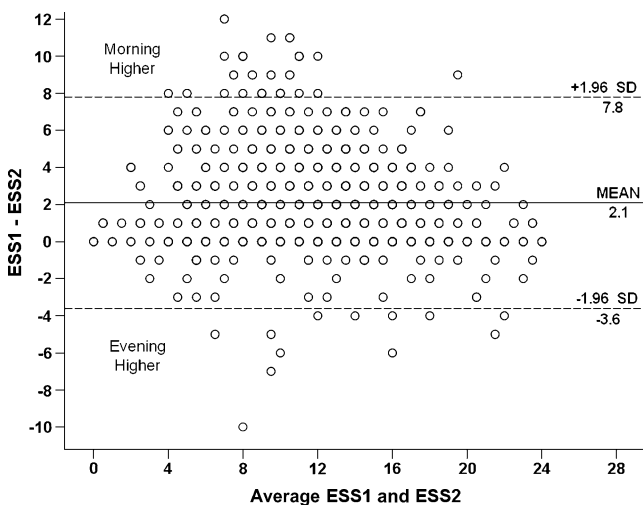
In the evening, the three most soporific situations were those described in questions 5, 2, and 4; the three

**Table 2** Means, standard deviations, medians, and interquartile range of the scores for each Epworth question before (ESS1) and after (ESS2) the polysomnography and for the difference between scores

ESS question	Evening ESS (ESS1)		Morning ESS (ESS2)		Difference ESS2–ESS1			<i>t</i>	<i>p</i>
	Mean±SD	Median [IR]	Mean±SD	Median [IR]	Mean±SD	Median [IR]	Range		
1. Reading	1.6±1	2 [1–2.5]	1.9±1	2 [1–3]	0.25±0.59	0 [0–0]	–2–3	12.7	0.000
2. Watching TV	1.9±0.95	2 [1–3]	2.2±0.9	2 [2–3]	0.36±0.64	0 [0–1]	–2–3	17.2	0.000
3. In public	1.2±1	1 [0–2]	1.3±1	1 [1–2]	0.17±0.48	0 [0–0]	–2–3	11.0	0.000
4. As a passenger	1.7±1.1	2 [1–3]	1.9±1.0	2 [1–3]	0.28±0.61	0 [0–1]	–2–3	13.7	0.000
5. Lying down	2.2±0.96	3 [2–3]	2.6±0.7	3 [3–3]	0.42±0.79	0 [0–1]	–2–3	16.3	0.000
6. Talking	0.48±0.8	0 [0–1]	0.59±0.8	0 [0–1]	0.11±0.46	0 [0–0]	–2–3	7.1	0.000
7. Sitting quietly	1.5±1.0	1 [1–2]	1.9±1.0	2 [1–3]	0.43±0.76	0 [0–1]	–3–3	16.9	0.000
8. In the traffic	0.39±0.7	0 [0–1]	0.44±0.7	0 [0–1]	0.06±0.47	0 [0–0]	–3–3	3.3	0.001
Total	11±5.1	11 [7–14]	13±4.7	13 [10–16]	2.1±2.7	1 [0–4]	–10–12	23.2	0.000

ESS Epworth sleepiness scale, SD standard deviation, IR 25–75% interquartile range.

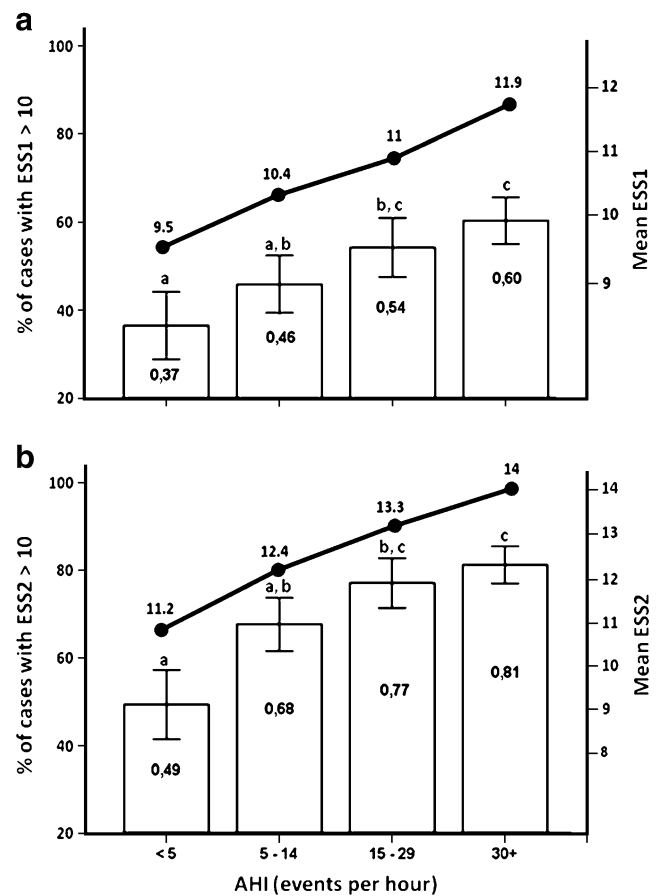
intermediate soporific situations were in questions 1, 3, and 7, and the two least soporific situations in questions 6 and 8. Considering that scores in questions with different levels of soporificity for one individual are supposed to be in a logical order, i.e., more soporific situations receive higher scores than less soporific situations, the results were arranged in tables comparing soporificity levels: (1) most and intermediate soporific, and (2) most and least soporific, as displayed in Table 6. The lower the Kappa coefficients for the agreement between two levels, the more likely the logic of the scale is preserved. Thus, the significant reduction of the kappa coefficients in ESS2 indicates better performance in terms of the “appropriateness” of the whole set of answers, as can be seen by the reduction in the



**Fig. 2** Bland–Altman plot of the difference and average values of the Epworth Sleepiness Scale (ESS) score, before (ESS1) and after (ESS2) polysomnography. Horizontal lines represent mean±1.96 standard deviations of the ESS2 to ESS1 difference. The largest variability is seen in average scores from 7 to 12, a range in which the diagnosis of sleepiness is more likely to change

number of inappropriate answers from 70 to 62 and from 8 to 6.

Two components were identified in the rotated matrix of factor analysis for the eight items both in ESS1 and



**Fig. 3** Mean ESS score and percentage of cases with Epworth Sleepiness Scale (ESS) score >10 before (panel a; ESS1) and after the polysomnography (panel b; ESS2). Different letters mean significant difference in post hoc Tukey’s test

**Table 3** Diagnostic performance of the Epworth Sleepiness Scale score >10 to predict AHI >5/h in the evening and in the morning

	Evening ESS (ESS1)		Morning ESS (ESS2)	
	Estimated value	95% CI	Estimated value	95% CI
Prevalence (%)	84	81–86	84	81–86
Sensitivity (%)	54	50–58	76 <sup>a</sup>	73–79
Specificity (%)	63	55–71	50 <sup>a</sup>	42–59
Overall accuracy (%)	56	53–58	72 <sup>a</sup>	70–74
Probability of a test being	Positive (%)	51	48–55	71 <sup>a</sup>
	Negative (%)	49	45–52	28 <sup>a</sup>
Probability of a positive test being	True positive (%)	88	85–91	89
	False positive (%)	11	09–15	11
Probability of a negative test being	True negative (%)	21	18–25	29 <sup>a</sup>
	False negative (%)	79	75–82	71 <sup>a</sup>
Likelihood ratios (conventional)	Positive	1.5	1.2–1.8	1.5
	Negative	0.72	67–78	0.48 <sup>a</sup>
Likelihood ratios (weighted by prevalence)	Positive	7.7	6.0–9.8	7.9
	Negative	3.7	3.5–3.9	2.5 <sup>a</sup>

<sup>a</sup> Characteristics that changed significantly from ESS1 to ESS2

ESS2. In ESS1, component 1 included items 1, 2, 4, 5, 7 and component 2 included items 3, 6, 8. In ESS2, component 1 incorporated items 1, 3, 6, 8 and component 2 included items 2, 4, 5, 7. This could indicate that the ESS is bi-dimensional.

Table 7 shows significant correlation coefficients of ESS1 and ESS2 with OSAHS variables. The increases in  $r$ , from ESS1 to ESS2, are small and within the 95% confidence interval. The increase of the correlation coefficients between lnAHI and ESS, from 0.15 in ESS1 to 0.21 in ESS2, is marginally significant ( $p \sim 0.06$ ).

The calculated reliable change index for ESS1 was 4.1. Then, changes larger than  $\pm 4$  points reflect actual variation in the appraisal of sleepiness. In our sample, only seven cases (0.75%) displayed a reduction larger than 4 points or more and 157 cases (16.9%) an increase of 5 points or more. These patients' PSG results are not significantly different from the results of the remaining ones.

IRT psychometric properties were assessed in the Rasch model (Table 8). The analysis of response threshold showed that the category responses 0, 1, 2, 3

**Table 4** Area under the ROC curve and significance for each question of ESS1 and ESS2 to predict AHI >5/h, >15, and >30 in the evening and in the morning

Questions	IAH >5 events/h				IAH >15 events/h				IAH >30 events/h			
	Evening ESS (ESS1)		Morning ESS (ESS2)		Evening ESS (ESS1)		Morning ESS (ESS2)		Evening ESS (ESS1)		Morning ESS (ESS2)	
	Area	$p$	Area	$p$	Area	$p$	Area	$p$	Area	$p$	Area	$p$
1. "Reading"	0.58	0.001	0.59	0.001	0.56	0.001	0.57	0.001	0.55	0.007	0.56	0.004
2. "Watching TV"	0.56	0.02	0.59	0.001	0.59	0.003	0.59	0.000	0.57	0.001	0.59	0.000
3. "In public"	0.57	0.01	0.58	0.002	0.57	0.001	0.57	0.000	0.56	0.003	0.55	0.007
4. "As a passenger"	0.57	0.01	0.58	0.003	0.53	0.09	0.55	0.020	0.53	0.15	0.55	0.014
5. "Lying down"	0.57	0.01	0.57	0.01	0.55	0.02	0.56	0.002	0.57	0.001	0.56	0.005
6. "Talking"	0.56	0.01	0.58	0.003	0.56	0.006	0.57	0.000	0.54	0.04	0.55	0.008
7. "Sitting quietly"	0.57	0.004	0.58	0.002	0.56	0.002	0.56	0.001	0.59	0.00	0.58	0.00
8. "In the traffic"	0.53	0.21	0.54	0.09	0.55	0.023	0.55	0.02	0.54	0.07	0.53	0.20
Total ESS	0.61	0.000	0.62	0.000	0.59	0.000	0.60	0.000	0.59	0.00	0.60	0.00

**Table 5** Kappa coefficients between evening and morning scores for each ESS question

	Morning ESS2: Q1—“Reading”						Morning ESS2: Q5—“Lying down”						
Evening ESS1; question 1; Kappa 0.63	0	1	2	3	Total	Evening ESS1; question 5; Kappa 0.39	0	1	2	3	Total		
	0	91	47	17	4	159		0	17	8	22	21	68
	1	3	190	75	11	279		1	5	47	32	65	149
	2	1	10	177	71	259		2	0	5	78	139	222
	3	0	3	9	220	232		3	0	1	13	476	490
	Total	95	250	278	306	929		Total	22	61	145	700	929
	Morning ESS: Q2—“Watching TV”						Morning ESS2: Q6—“Talking”						
Evening ESS1; question 2; Kappa 0.49	0	1	2	3	Total	Evening ESS1; question 6; Kappa 0.73	0	1	2	3	Total		
	0	33	30	13	4	80		0	515	78	8	3	604
	1	1	116	85	35	237		1	11	199	14	2	226
	2	1	7	169	137	314		2	2	5	61	7	75
	3	0	0	11	287	298		3	0	3	2	19	24
	Total	35	153	278	463	929		Total	528	285	85	31	929
	Morning ESS: Q3—“In public”						Morning ESS2: Q7—“Sitting quietly”						
Evening ESS1; question 3; Kappa 0.72	0	1	2	3	Total	Evening ESS1; question 7; Kappa 0.51	0	1	2	3	Total		
	0	210	66	10	1	287		0	88	42	33	15	178
	1	4	250	65	2	321		1	5	145	99	44	293
	2	2	11	168	25	206		2	2	6	152	84	244
	3	0	0	3	112	115		3	1	2	5	206	214
	Total	216	327	246	140	929		Total	96	195	289	349	929
	Morning ESS2: Q4—“As a passenger”						Morning ESS2: Q8—“In the traffic”						
Evening ESS1; question 4; Kappa 0.62	0	1	2	3	Total	Evening ESS1; question 8; Kappa 0.72	0	1	2	3	Total		
	0	81	57	21	4	163		0	604	63	4	3	674
	1	5	165	78	16	264		1	18	147	13	2	180
	2	2	4	177	61	244		2	3	3	40	1	47
	3	0	2	7	249	258		3	3	4	1	20	28
	Total	88	228	283	330	929		Total	628	217	58	26	929

are significantly different in most questions. However, for question 5, the levels 2 and 3 are marginally similar. In the ESS1, question 5 does not fit the model for presenting a residual of 3.24, higher than the acceptable limit of 2.5. In the ESS2, the item 5 presented disordered response thresholds and, to solve this, the codes 0123 were substituted with 0112, reducing the  $\chi^2$  of the item–trait interaction of highly significant value ( $p=0.00002$ ) to values near the optimum ( $p=0.02$ ). Both in ESS1 and in ESS2 a redundancy was noted in question 5, which was reduced, but was not eliminated on the second administration. For ESS1, person fit residual was  $-0.280$ , PSI was 0.84, and paired  $t$  test 95% confidence interval was 0.03–0.06; for ESS2, person fit residual was  $-0.310$ , PSI was 0.83, and paired  $t$ -test 95% confidence interval was 0.05–0.07. In ESS1, there were statistically significant differences in scores, representing lack of invariance in gender (item 3; sleeping in public), age (items 1, 2, 4, and 8) and educational level (item 1; reading). In ESS2, the lack of invariance was present only for age, in

the same questions. The Rasch analysis identified 13 cases in the ESS1 and 12 in the ESS2 that did not fit the model.

**Discussion**

Based on extensive literature revision, this is the first study to assess the effect of short-term ESS repetition on its diagnostic performance and both classical and IRT psychometric properties in a large sample. The IRT psychometric properties show small improvement from ESS1 to ESS2.

Validation studies show that the ESS score is relatively maintained after 1 year [33] or months [34, 35]. In such long intervals, the learning effect may be attenuated. The present study did not aim to identify the effect of different time intervals on the ESS score. It may be useful testing the effect of ESS repetition in the office, within a shorter time interval of one or two hours.

**Table 6** Four by four tables of the rounded mean scores in the three most soporific situations (questions 5, 2, and 4) by three intermediate soporific situations (questions 1, 3, and 7), and by the two least

soporific situations (questions 6 and 8), in the evening (ESS1) and in the morning (ESS2)

	Most soporific situations—ESS1						Most soporific situations—ESS1						
Intermediate soporific situations ESS1; Kappa 0.21; inappropriate answers: 70	0	1	2	3	Total	Least soporific situations ESS1; Kappa 0.17; inappropriate answers: 8	0	1	2	3	Total		
	0	29	68	36	5	138		0	40	180	228	78	526
	1	<b>12</b>	120	187	44	363		1	<b>2</b>	38	162	104	306
	2	<b>2</b>	<b>27</b>	172	117	318		2	<b>1</b>	<b>2</b>	26	46	75
	3	0	<b>5</b>	<b>24</b>	81	110		3	0	0	<b>3</b>	19	22
	Total	43	220	419	247	929		Total	43	220	419	247	929
	Most soporific situations—ESS2						Most soporific situations—ESS2						
Intermediate soporific situations ESS2; Kappa 0.14 <sup>a</sup> ; inappropriate answers: 52	0	1	2	3	Total	Least soporific situations ESS2; Kappa 0.03 <sup>a</sup> ; inappropriate answers: 6	0	1	2	3	Total		
	0	8	37	27	4	76		0	10	88	226	126	450
	1	<b>3</b>	57	165	55	280		1	<b>1</b>	18	152	192	363
	2	0	<b>12</b>	176	216	404		2	0	<b>1</b>	22	69	92
	3	0	<b>1</b>	<b>36</b>	132	169		3	0	0	<b>4</b>	20	24
	Total	11	107	404	407	929		Total	11	107	404	407	929

“Inappropriate” answers are displayed in bold type.

<sup>a</sup> Significantly different from the Kappa coefficient in ESS1

Pien et al. considered as evidence of increased sleepiness during pregnancy the ESS changing from 8.6 to 10.2 [36]; this is less than the 2.3 points overnight change in the present report. Chin et al. described as proof of alteration in perception of sleepiness an increase in the pre-treatment ESS, recalled after CPAP use, from 8.5 to 11.1 [37], indistinguishable from our findings. On the other hand, reductions of ESS score after treatment of OSAHS [31, 38] could be greater than reported if the effect of repetition was controlled for.

The Bland–Altman diagram (Fig. 2) shows a wide range of individual variation in ESS score, implying directly to the utility of the ESS in clinical practice. In 50% of the

cases, score changed between 0 and 4 points. The range of two standard deviations, from −10 to 12, indicates that although the bias is small, significant individual variability is seen.

In Fig. 3, the average ESS and the percentage of ESS scores >10 increased more in the groups with AHI>5 than in the negative group, with AHI<5. This non-significant finding suggests that the increase in ESS score is more pronounced in the abnormal AHI bins. The group with AHI<5, is not a healthy control group since subjects underwent PSG to investigate a sleep complaint. Normal AHI does not exclude sleepiness secondary, for instance, to upper airway resistance syndrome or sleep deprivation. Considering that all subjects feel that their sleep needs medical attention, it is more likely that a higher ESS score in the second administration correctly captured actual sleepiness, even in people with AHI<5.

ESS questions in which the patients scored higher indicate a more soporific nature of the situation. One can consider an inappropriate answer responding “would never doze” to the more soporific question 5 and mark a “high chance of dozing” to question 7, less soporific. In our sample (Table 6), for ESS1, 8.9% of the testees had a higher score in question 3 than in question 5; in the repeated test, only 3% maintained this “inappropriate” pattern. Thus, the reduction in inappropriate answers indicates that the retest leads to a more judicious choice of options.

In the Rasch analysis, age separated at median of 40 years, affects questions 1, 2, 4, and 8, violating the

**Table 7** Pearson’s Moment Correlation Coefficient Between Several Polysomnographical and Clinical Outcomes, Including Questions of the Berlin Questionnaire, and First and Second Epworth Sleepiness Scale Administrations

	ESS1	ESS2
	<i>r</i> ( <i>p</i> )	<i>r</i> ( <i>p</i> )
Age	0.03 (0.44)	0.06 (0.10)
BMI	0.15 (0.001)	0.16 (0.001)
Stage N2 latency	−0.18 (0.000)	−0.21 (0.000)
Sleep efficiency	0.19 (0.000)	0.20 (0.000)
Percent time in apnea–hypopnea	0.18 (0.000)	0.21 (0.000)
ln Apnea–hypopnea index	0.15 (0.000)	0.21 (0.000)
Lowest O <sub>2</sub> saturation	−0.21 (0.000)	−0.24 (0.000)



**Table 8** Individual item fit in the Rasch model

Item	Evening ESS (ESS1)					Morning ESS (ESS2)				
	Location	SE	Fit residual	chi <sup>2</sup>	<i>p</i>	Location	SE	Fit residual	chi <sup>2</sup>	<i>p</i>
1. Reading	−0.51	0.04	−0.82	14.7	0.10	−0.45	0.04	−1.6	14.4	0.11
2. Watching TV	−0.99	0.05	1.7	7.3	0.60	−1.2	0.05	1.0	16.8	0.05
3. In public	0.26	0.04	−0.5	14.9	0.09	0.44	0.04	−0.55	19.2	0.02
4. As a passenger	−0.57	0.04	1.3	13.7	0.13	−0.53	0.04	1.9	16.5	0.06
5. Lying down	−1.4	0.04	<b>3.24</b>	13.0	0.16	−1.8	0.06	1.4	7.5	0.58
6. Talking	1.8	0.06	−1.6	21.2	0.01	1.9	0.05	−1.8	17.5	0.04
7. Sitting quietly	−0.38	0.04	−1.2	21.0	0.01	−0.54	0.04	−1.5	23.4	0.01
8. In the traffic	1.8	0.06	−1.1	14.3	0.11	2.2	0.06	−0.2	17.3	0.04
Overall measures of fit			0.12	132.7	0.0002			−0.2	128.5	0.0005

ESS Epworth sleepiness scale, SE standard error, *chi*<sup>2</sup> chi-square, fit residual above acceptable limit is in boldtype

invariance assumption, in both ESS1 and ESS2. As the two age groups were similarly sleepy, it is possible that from middle age on, people are more aware of sleepiness sensation. This is consistent with the report of more sleepiness-related car accidents among younger drivers because of them being more often unaware of their sleepiness [39]. The effect of age on lack of invariance on ESS answers is similar in ESS1 and ESS2. The fact that items 1 and 3 started to display invariance for educational level and gender in the ESS2 indicates an improvement of psychometric properties assessed in the Rasch model.

The reliable change index defines the variation in a score that is necessary and sufficient for the change to reflect an actual alteration in the measured variable. In the present study, the calculated reliable change index is 4.1, meaning that a five-point change is a reliable change of the ESS score. This is similar to the index of six points reported by Smith and Sullivan [31]. Chin et al.'s [37] finding of post-treatment reduction in ESS from 8.5 to 4.3 points is within the reliable change index. Several studies report changes in ESS without a control group for comparison. Having patients being their own controls in studies involving ESS is a problematic method often used. For instance, ESS reduction with weight loss repeated four times over 24 months, from 7.9 to 4.8 points, was not controlled for repetition effect [40]. Evidence based on such comparisons might need reassessment.

The correlation coefficients between clinical and PSG outcomes and ESS increase from ESS1 to ESS2 (Table 7). This improvement suggests that ESS2 is a better predictor of sleepiness-related polysomnographic outcomes.

Patient–technician communication during electrodes placement may have amplified the morning ESS score. The possibility that sleep inertia in the morning made the patients more prone to respond with higher ESS scores is

countered by the negligible change in the SSS. Although in the morning patients filled the forms swiftly, allowing less time for deliberation before leaving the laboratory, the correlation between AHI and ESS is higher in the second administration.

Repeating the scale using results obtained from the patients' bed partners is another approach to increasing the diagnostic accuracy of the ESS [33, 41]. The results in small samples (60 and 82 subjects) did not show clinical utility of duplicating the scale administration. The increases of 1.2 and 1.8 in the ESS by the bed partner are less than what is here reported.

Polysomnography is paramount for OSAHS diagnosis, but a full sleep study is requested usually when a symptom, more often sleepiness, is present and preferably confirmed by a valid instrument. An abnormal score in the ESS will prompt the physician to order PSG and motivate the patient to undergo the exam. One determination of ESS score below 10, however, should not preliminarily exclude the existence of sleepiness, before the physician avouches the correctness of the ESS result by repeating the scale administration or even questioning the bed partner.

For the individual patient, to whom treatment for OSAHS may be denied on the basis of a low ESS score, repeating the scale may be lifesaving, at a negligible cost. Repeating ESS with bed partners or with patients may be equally beneficial before dismissing patients as "asymptomatic". As was demonstrated, in individual cases, ESS may be inappropriately understood by some patients, reducing their opportunity for diagnosis and treatment.

Is it adequate to repeat the ESS? Vis-a-vis our results, in a sleep laboratory population, the answer is yes. It does not change the psychometric properties of the scale, improves the diagnostic performance of the test and the correlation

with OSAHS variables. Further research is required to complete the understanding of cost-effectiveness of ESS repetition, measuring its effect on primordial outcomes such as mortality reduction.

**Acknowledgement** The authors declare that they have no conflict of interest.

## References

- Flemons WW (2002) Obstructive sleep apnea. *N Engl J Med* 347:498–504
- American Academy of Sleep Medicine (2005) The international classification of sleep disorders: diagnostic and coding manual, 2nd edn. American Academy of Sleep Disorders, Westchester
- Epstein LJ, Kristo D, Strollo PJ Jr, Friedman N, Malhotra A, Patil SP, Ramar K, Rogers R, Schwab RJ, Weaver EM, Weinstein MD (2009) Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J Clin Sleep Med* 5:263–276
- Johns M (1999) Rethinking the assessment of sleepiness. *Sleep Med Rev* 2:3–15
- Martinez D, Lumertz MS, Lenz Mdo C (2009) Dimensions of sleepiness and their correlations with sleep-disordered breathing in mild sleep apnea. *J Bras Pneumol* 35:507–514
- Tachibana N, Taniguchi M (2007) Why do we continue to use Epworth sleepiness scale? *Sleep Med* 8:541–542
- George PT (2001) Sleepiness, trioka of consciousness cycle, and the epworth sleepiness scale. *Sleep Breath* 5:181–191
- Johns MW (1991) A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 14:540–545
- Leineweber C, Kecklund G, Janszky I, Akerstedt T, Orth-Gomer K (2004) Snoring and progression of coronary artery disease: the Stockholm Female Coronary Angiography Study. *Sleep* 27:1344–1349
- Tsara V, Serasli E, Amfilochiou A, Constantinidis T, Christaki P (2004) Greek version of the epworth sleepiness scale. *Sleep Breath* 8:91–95
- Izci B, Ardic S, Firat H, Sahin A, Altinors M, Karacan I (2008) Reliability and validity studies of the turkish version of the epworth sleepiness scale. *Sleep Breath* 2:161–168
- Thomas GN, Jiang CQ, Lao XQ, McGhee SM, Zhang WS, Schooling CM, Adab P, Lam TH, Cheng KK (2006) Snoring and vascular risk factors and disease in a low-risk Chinese population: the Guangzhou Biobank Cohort Study. *Sleep* 29:896–900
- Beiske KK, Kjelsberg FN, Ruud EA, Stavem K (2009) Reliability and validity of a Norwegian version of the epworth sleepiness scale. *Sleep Breath* 13:65–72
- Chan EYT, Ng DK, Chan C, Kwok K, Chow P, Cheung JM, Leung SY (2009) Modified Epworth Sleepiness Scale in chinese children with obstructive sleep apnea: a retrospective study. *Sleep Breath* 13:59–63
- Bausmer U, Gouveris H, Selivanova O, Goepel B, Mann W (2010) Correlation of the Epworth Sleepiness Scale with respiratory sleep parameters in patients with sleep-related breathing disorders and upper airway pathology. *Eur Arch Otorhinolaryngol*. doi:10.1007/s00405-010-1250-y
- Jiménez-Correa U, Haro R, González-Robles RO, Velázquez-Moctezuma J (2010) How is the Epworth Sleepiness Scale related with subjective sleep quality and polysomnographic features in patients with sleep-disordered breathing? *Sleep Breath*. doi:10.1007/s11325-010-0372-1
- Cho YW, Lee JH, Son HK, Lee SH, Shin C, Johns MW (2010) The reliability and validity of the korean version of the epworth sleepiness scale. *Sleep Breath*. doi:10.1007/s11325-010-0343-6
- Kumar S, Bhatia M, Behari M (2003) Excessive daytime sleepiness in Parkinson's disease as assessed by Epworth Sleepiness Scale (ESS). *Sleep Med* 4:339–342
- DeZee KJ, Jackson JL, Hatzigeorgiou C, Kristo D (2006) The epworth sleepiness scale: Relationship to sleep and mental disorders in a sleep clinic. *Sleep Med* 7:327–332
- Sanford SD, Lichstein KL, Durrence HH, Riedel BW, Taylor DJ, Bush AJ (2006) The influence of age, gender, ethnicity, and insomnia on epworth sleepiness scores: a normative US population. *Sleep Med* 7:319–326
- Miletin MS, Hanly PJ (2003) Measurement properties of the Epworth sleepiness scale. *Sleep Med* 4:195–199
- Johns MW (1994) Sleepiness in different situations measured by the Epworth Sleepiness Scale. *Sleep* 17:703–710
- Chervin RD (2003) Epworth sleepiness scale? *Sleep Med* 4:175–176
- Smith SS, Oei TP, Douglas JA, Brown I, Jorgensen G, Andrews J (2008) Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnoea. *Sleep Med* 9:739–744
- Wirtz M, Böcker M (2007) Properties and benefits of applying the Rasch model in clinical diagnostics. *Rehabilitation* 46:238–245
- Embretson SE, Reise SP (2000) Item response theory for psychologists. Erlbaum, Mahwah, NJ
- Guyatt GH, Kirschner B, Jaeschke R (1992) Measuring health status: what are the necessary measurement properties? *J Clin Epidemiol* 45:1347–1351
- Wright JG, Feinstein AR (1992) A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol* 45:1201–1218
- Takegami M, Suzukamo Y, Wakita T, Noguchi H, Chin K, Kadotani H, Inoue Y, Oka Y, Nakamura T, Green J, Johns MW, Fukuhara S (2009) Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. *Sleep Med* 10:556–565
- Maassen GH (2004) The standard error in the Jacobson and Truax Reliable Change Index: the classical approach to the assessment of reliable change. *J Int Neuropsych Soc* 10:888–893
- Smith S, Sullivan KA (2007) A reliable change index (RCI) for the Epworth Sleepiness Scale (ESS). *Sleep Med* 9:102
- Lenz MC, Martinez D (2007) Awakenings change results of nighttime ambulatory blood pressure monitoring. *Blood Press Monit* 12:9–15
- Kumru H, Santamaria J, Belcher R (2004) Variability in the Epworth Sleepiness Scale score between the patient and the partner. *Sleep Med* 5:369–371
- Nguyen AT, Baltzan MA, Small D, Wolkove N, Guillon S, Palayew M (2006) Clinical reproducibility of the Epworth Sleepiness Scale. *J Clin Sleep Med* 2:170–174
- Knutson KL, Rathouz PJ, Yan LL, Liu K, Lauderdale DS (2006) Stability of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Questionnaires over 1 year in early middle-aged adults: the CARDIA study. *Sleep* 29:1503–1506
- Pien GW, Fife D, Pack AI, Nkwuo JE, Schwab RJ (2005) Changes in symptoms of sleep-disordered breathing during pregnancy. *Sleep* 28:1299–1305
- Chin K, Fukuhara S, Takahashi K, Sumi K, Nakamura T, Matsumoto H, Niimi A, Hattori N, Mishima M, Nakamura T (2004) Response shift in perception of sleepiness in obstructive sleep apnea–hypopnea syndrome before and after treatment with nasal CPAP. *Sleep* 27:490–493

38. Zimmermann C, Kohler D, Schonhofer B (2000) Value of retrospective assessment of the Epworth Sleepiness Scale after long-term CPAP therapy in obstructive sleep apnea disorder. *Pneumologie* 54:572–574
39. Philip P, Taillard J, Quera-Salva MA, Bioulac B, Akerstedt T (1999) Simple reaction time, duration of driving and sleep deprivation in young versus old automobile drivers. *J Sleep Res* 8:9–14
40. Holty JE, Parimi N, Ballesteros M, Blackwell T, Cirangle PT, Jossart GH, Kimbrough ND, Rose JM, Stone KL, Bravata DM (2010) Does surgically induced weight loss improve daytime sleepiness? *Obes Surg*. doi:10.1007/s11695-010-0213-0
41. Walter TJ, Foldvary N, Mascha E, Dinner D, Golish J (2002) Comparison of Epworth Sleepiness Scale scores by patients with obstructive sleep apnea and their bed partners. *Sleep Med* 3:29–32