

The fairness of the Epworth Sleepiness Scale: two approaches to differential item functioning

Martin Ulander · Kristofer Årestedt · Eva Svanborg · Peter Johansson · Anders Broström

Received: 8 September 2011 / Revised: 29 December 2011 / Accepted: 30 January 2012 / Published online: 26 February 2012
© Springer-Verlag 2012

Abstract

Purpose Differential item functioning (DIF) is said to exist in an item if a subject's response to the item is affected by other aspects than that which the test is intended to assess. DIF might affect the validity of a test. The aim of this study was thus to examine whether any of the items in the Epworth Sleepiness Scale (ESS) exhibits DIF regarding age or gender, and if so, to which degree.

Methods Using previously collected cross-sectional ESS data from 1,168 subjects with different clinical characteristics (61% males, mean age 67.8 year (SD 12.2 year)), ordinal regression as well as Rasch-based DIF analyses were performed.

Results Concerning age, both DIF analyses showed DIF for age in items 3 (inactive in a public place), 4 (passenger in a car), and 8 (in a car that has stopped in traffic). The Rasch model also showed DIF for gender in item 3. The DIF magnitudes as judged by McFadden pseudo- R^2 changes were, however, only minor.

Conclusions ESS has small but reproducible DIF for age in items 3, 4, and 8. The detected DIF might be worth to consider in large-sample studies, although it probably has no effect on an individual basis.

Keywords Epworth Sleepiness Scale · Differential item functioning · Sleep · Daytime sleepiness

M. Ulander · E. Svanborg
Institution of Clinical and Experimental Medicine,
Linköping University,
Linköping, Sweden

M. Ulander (✉) · E. Svanborg · A. Broström
Department of Clinical Neurophysiology, University Hospital,
581 85 Linköping, Sweden
e-mail: martin.ulander@liu.se

K. Årestedt · P. Johansson
Department of Medical and Health Sciences, Division of Nursing,
Linköping University,
Linköping, Sweden

K. Årestedt
School of Human Sciences, Linnaeus University,
Kalmar, Sweden

P. Johansson
Department of Cardiology, University Hospital,
Linköping, Sweden

A. Broström
Department of Nursing Science, School of Health Sciences,
Jönköping University,
Jönköping, Sweden

Introduction

The Epworth Sleepiness Scale (ESS) [1] is one of the most wide-spread inventories to assess subjective daytime sleepiness over time. The respondent is asked to rate the risk of falling asleep on a four-point Likert scale (0–3 points, where 0 indicates no risk of falling asleep) in eight different situations, chosen to differ in their level of sleep-inducing capacity. The original article by Johns has been cited over 3,000 times. Given its wide-spread use, it is important to study its psychometric properties. There are several studies on different aspects of the psychometric properties, e.g. [2–7]. Johns [2] assessed test–retest reliability in medical students, as well as criterion validity in OSAS patients before and after CPAP initiation. He used these two groups to assess dimensionality, finding a single factor solution. This was not confirmed by Smith et al. [7] in a confirmatory factor analysis. Chervin and Aldrich [4], when performing a regression analysis on ESS and multiple sleep latency test (MSLT) data, found that males, on average, had 2.2 points lower ESS score, even after correction for age and mean MSLT sleep latency. Hagell and Broman [6] showed good

psychometric properties and item hierarchy using a Rasch model in patients with Parkinson disease. Violani et al. [5] studied Rasch properties in 146 sleep-disordered patients and found item misfit for item 5, and also that most of the items were located at the extremes of the difficulty spectrum. There is, to our knowledge, no previous study that has examined whether the scale shows differential item functioning (DIF).

A lot of early psychometric work was done on intelligence tests, and this has affected psychometric terminology in a way that might be confusing. The amount of whatever the test is supposed to measure in an individual is referred to as the individual's "ability". Concerning the ESS, a sleepy subject is thus said to have a higher "ability" than a non-sleepy subject, and will probably get a higher ESS score. Not all items in a test measure the same level of the property that is measured. It is easier to fall asleep while resting in the afternoon (i.e., item 5) than to fall asleep while sitting and talking to someone (i.e., item 6). In other words, a person has to have a "higher ability" (i.e., has to be more sleepy) to score on item 6 than on item 5. In Rasch terminology, item 6 is said to be more "difficult" than item 5 [8].

An item in a psychometric test is considered to exhibit DIF if the responses to the item are affected independently by other factors than the property that the item intends to assess. In other words, in order for the ESS to measure sleepiness reliably, only the sleepiness of the subjects should affect how they score. If, for example, the age or gender of the test taker affects his or her score on an item independently, there is a problem with the validity of the item. If DIF is large and exists in several items, it might give rise to differential test functioning, meaning that the test itself produces different scores for equally sleepy subjects who differ only with regard to background variables. In Rasch analysis, an item is said to be difficult if it measures a high degree of sleepiness. The items in the Epworth Sleepiness Scale depict different everyday life situations, describing different degrees of sleepiness (i.e., item difficulty). There is, however, a lack of information regarding how the items were initially chosen [1]. Activities of everyday life might differ in different groups of subjects. This might affect how people respond to the ESS regardless of how sleepy they are. The aim of this study was thus to examine whether any of the items in the Epworth Sleepiness Scale exhibits differential item functioning with regard to age or gender, and if so, to which degree.

Methods

Design and samples

The present study is based on pooled data ($n=1,168$) collected in five different cross-sectional studies (Table 1). Thus, no patients were specifically recruited for this study.

The populations consisted of OSA patients with CPAP [9] ($n=237$), OSA patients without CPAP [10] ($n=170$), patients with hypertension [10] ($n=214$), patients with heart failure [11] ($n=215$) and home-dwelling elderly patients [12] ($n=331$). The patients had completed the ESS as part of the original studies, and these data were used for the analysis. The version of the ESS used was translated into Swedish by Jan-Erik Broman (personal communication) upon request by the Swedish Society for Sleep Research and Sleep Medicine. It has been back-translated into English and approved by M.W. Johns. All original studies have been ethically approved by the Regional Ethical review Board in Linköping.

Analytical strategy

Ordinal regression DIF analyses can be performed in three steps. In the first step [13], the item score is used as the dependent variable, and the total test score is used as an independent variable. Then, the grouping variable (i.e., the source of DIF one wants to test for, e.g., age or gender) is added as an independent variable. If the grouping variable is significantly related to the item score, when potential between-group differences in total score are taken into account, the item is said to exhibit DIF with regard to the grouping variable. Then, changes in pseudo- R^2 when the grouping variable is added can be studied to assess the magnitude of DIF. In a third step, if DIF is found, interaction terms between total test score and grouping variable can be added to test for non-uniform DIF across the range of total score. In item response theory based DIF analysis [8], the subjects are divided into classes depending on their ability (i.e., their sleepiness as defined by their total score). A two-way ANOVA is then performed to test for significant differences in item response probabilities between the different groups of subjects. The Rasch model is an item response model where only item difficulty is allowed to vary between items.

Statistical analysis

Baseline statistical analysis was performed to assess differences in age between males and females using both a two-sample T test and a chi-square test of dependency. For the chi-square test, the subjects were divided into two age groups (<65 vs. ≥ 65 years).

The method presented by Zumbo [13] was used. A hierarchical ordinal regression model was made for each item, where the item score was dependent and the total ESS score was used as an independent variable. In a second step, age was added as an independent variable. The change in McFadden pseudo- R^2 between the two models for each item was used to assess the importance of DIF where it occurred.

Table 1 Demographic data

	<i>N</i>	Gender Males <i>n</i> (%)	Age Mean (SD)	ESS score Mean (SD)	Reference
Demographic data for each subgroup as well as for the pooled group. Age data was missing for 45 subjects in the Broström et al. study [9] (36 males and 9 females). These 45 subjects were included in the DIF analysis based on gender, but excluded from the analysis based on age					
All groups	1,168	715 (61.2%)	67.8 (12.2)	7.7 (4.5)	
Home-dwelling elderly	331	164 (49.5%)	78.5 (3.2)	6.6 (3.7)	[12]
OSA patients without CPAP	170	132 (77.6%)	55.9 (11.4)	11.0 (4.9)	[9]
OSA patients with CPAP	238	194 (81.4%)	60.2 (9.3)	7.3 (4.6)	[29]
Heart failure patients	215	131 (60.9%)	75.9 (9.3)	6.8 (3.8)	[11]
Hypertension patients	214	95 (44.4%)	58.8 (5.8)	7.9 (4.4)	[10]

For the DIF analysis of gender, a similar approach was used. A model including total ESS score, age and gender was used to differ between age- and gender-related DIF in items where both age and gender were significant predictors. Non-uniform DIF was tested for by adding an interaction term between age and total ESS score or gender and total ESS score, respectively. The ordinal regression DIF analysis was performed with Stata version 10.0 (Stata Corporation, TX, USA).

The item response theory DIF analysis was based on a polytomous Rasch model. Person and item parameters were estimated using the joint maximum likelihood method [8]. The analyses were conducted with the sample grouped into eight class intervals. Each class interval consisted of people with similar ability level (i.e., similar level of daytime sleepiness). The number of subjects in each class interval ranged from 93 to 288. Global fit was assessed by examining whether item and person fit residuals diverged from expected values (i.e., mean 0 and standard deviation 1). In addition, global as well as individual item fit was also assessed by the item–trait interaction (χ^2 statistic). A good model fit was defined as a non-significant χ^2 test (Bonferroni-corrected to $p < 0.00625$ due to the eight class intervals) when comparing the model and the data. In order to avoid detecting statistically but not practically significant item misfit, a visual inspection of the item characteristic curves (ICC), as described by Hagqvist [14], was also used. This was done prior to the decision to proceed with the DIF analysis.

The targeting (i.e., whether the items and response categories cover the different levels of sleepiness experienced by the subjects) was evaluated by comparing person ability levels with the item threshold distributions. The item thresholds estimates were also used to evaluate the response scale of the ESS. In the present study, each class interval was split according to gender and age and then analyzed separately. Age was dichotomized into two groups, below 65 years and 65 years or older, respectively. A main effect of age and gender was used to detect uniform DIF. An interaction effect between class intervals across the latent trait (i.e., level of daytime sleepiness) and age and gender respectively, indicates non-uniform DIF. To reduce the risk for type I error

inflation due to multiple tests, the analysis was Bonferroni adjusted to a significance level of $p < 0.0021$ (due to the number of ANOVAs, i.e., eight items, and two main effects and one interaction effect). Global reliability was assessed by Cronbach's alpha and person separation index. Individual item fit (i.e., how well persons respond to each item) was evaluated by item fit residuals ($-2.5 < \text{residual} < 2.5$), chi-square statistics (Bonferroni adjusted to $p < 0.00625$), and by inspecting the ICC. The Rasch analysis was performed with the RUMM2020 software (RUMM Laboratories Pty Ltd, Australia).

Results

The participants

The total sample consisted of 1,168 subjects. Twenty-five subjects obtained a total score of 0, meaning that their ability parameters cannot be estimated, and they were therefore excluded from the DIF analysis. No subject scored 24 points. Forty-five subjects for whom age data were missing were excluded from the DIF analysis for age, but included in the DIF analysis for gender. The item response theory based DIF analyses concerning age are thus based upon 1,098 subjects and the DIF analysis for gender is based on 1,143 subjects. The study populations and characteristics are described in Table 1. The χ^2 test of independence showed a significantly ($p < 0.001$) higher mean age for female participants than for male participants.

Ordinal regression analysis

The results from the ordinal regression analyses are shown in Table 2. The model showed significant DIF for age for items 3, 4, and 8 and concerning gender for item 3. When both age and gender were used as regressors, the significance remained for age but not for gender in item 3. No multicollinearity problems were detected as judged from the variance inflation factor, but the parallel lines assumption was violated.

Table 2 Ordinal regression location estimates, DIF significances, and DIF magnitudes

Item	Variables in ordinal regression model	Location estimate	Significance	McFadden pseudo- R^2 (change)
Item 1	ESS total			0.2663
	ESS total, age	0.015 (0.005)	0.004	0.2665 (0.0002)
	ESS total, sex	0.005 (0.122)	0.965	0.2663 (0.0)
Item 2	ESS total			0.2312
	ESS total, age	0.002 (0.005)	0.682	0.2318 (0.0006)
	ESS total, sex	-0.288 (0.118)	0.015	0.2331 (+0.0019)
Item 3	ESS total			0.3285
	ESS total, age	-0.022 (-0.006)	<0.001	0.3318 (+0.0033)
	ESS total, sex	0.500 (0.141)	<0.001	0.3336 (+0.0051)
Item 4	ESS total			0.2883
	ESS total, age	-0.037 (0.005)	<0.001	0.3032 (+0.0149)
	ESS total, sex	-0.181 (0.131)	0.167	0.2890 (+0.0007)
Item 5	ESS total			0.1521
	ESS total, age	0.011 (0.005)	0.029	0.1554 (0.0033)
	ESS total, sex	-0.019 (0.122)	0.875	0.1521 (0.0)
Item 6	ESS total			0.2882
	ESS total, age	0.021 (0.009)	0.017	0.2934 (0.0052)
	ESS total, sex	0.473 (0.213)	0.026	0.2925 (0.0043)
Item 7	ESS total			0.2831
	ESS total, age	0.009 (0.005)	0.111	0.2888 (+0.0057)
	ESS total, sex	0.180 (0.130)	0.169	0.2838 (+0.0009)
Item 8	ESS total			0.3186
	ESS total, age	0.047 (0.009)	<0.001	0.3442 (+0.0256)
	ESS total, sex	-0.079 (0.200)	0.694	0.3187 (+0.0001)

Ordinal regression models. Three ordinal regression models were made for each item, differing in the included independent variables. Significance values less than 0.006 are significant after Bonferroni adjustment for eight items. McFadden pseudo- R^2 changes are compared to the regression model where total ESS score is used as the only independent variable

Rasch analysis

Item parameters from the Rasch model are shown in Table 3, and DIF parameters are shown in Table 4. Person–item distribution is shown in Fig. 1. This indicates whether the degree of sleepiness measured by the test matches the

degree of sleepiness in the sample. The mean fit residual for persons was -0.309 (SD 0.785) and for items -0.633 (SD 2.716). Cronbach’s alpha was 0.80 and the person separation index was 0.79. The global fit in item–trait interaction (chi-square) showed a significant model misfit for the whole test (the “trait” in question being sleepiness).

Table 3 Rasch model item parameters, model fit, and distribution of responses

Item	Location	Residual	Sig.	No chance		Small chance		Moderate chance		High chance
				N (%)	Threshold	N (%)	Threshold	N (%)	Threshold	
1	-0.391	-0.971	<0.001	397 (34.6%)	-1.161	381 (33.2%)	0.097	250 (21.8%)	1.064	119 (10.4%)
2	-1.515	1.055	0.413	148 (12.9%)	-1.078	297 (25.9%)	0.039	376 (32.8%)	1.039	326 (28.4%)
3	0.403	-3.301	<0.001	621 (54.1%)	-1.069	323 (28.2%)	-0.027	153 (13.3%)	1.096	50 (4.4%)
4	-0.084	-0.896	0.015	585 (51.0%)	-0.575	284 (24.8%)	0.057	166 (14.5%)	0.518	112 (9.8%)
5	-2.087	5.206	<0.001	65 (5.7%)	-0.561	172 (15.0%)	0.048	320 (27.9%)	0.513	590 (51.4%)
6	1.623	-1.644	0.039	975 (85.0%)	-0.423	135 (11.8%)	0.039	30 (2.6%)	0.384	7 (0.6%)
7	0.11	-1.521	<0.001	548 (47.8%)	-1.01	333 (29.0%)	-0.063	198 (17.3%)	1.073	58 (5.1%)
8	1.94	-2.995	0.001	961 (83.8%)	-0.828	141 (12.3%)	-0.527	40 (3.5%)	1.354	5 (0.4%)

Location on the logit scale, item fit residual with significances and response distribution as well as thresholds for each response alternative for each item. Bonferroni adjusted significances are highlighted

Table 4 Rasch model DIF parameters

Item	Gender			Age		
	Class interval	Gender	Gender×CInt	Class interval	Age	Age×CInt
1. Sitting and reading	0.000	0.739	0.718	0.000	0.032	0.090
2. In a public place eg a theater or a meeting	0.371	0.006	0.492	0.451	0.994	0.925
3. Watching TV	0.000	0.000	0.969	0.000	0.000	0.999
4. Passenger in a car for an hour w/o break	0.103	0.230	0.374	0.140	0.000	0.056
5. Lying down in the afternoon	0.000	0.592	0.918	0.000	0.053	0.039
6. Sitting and talking to someone	0.237	0.097	0.794	0.232	0.229	0.653
7. Sitting quietly after a lunch w/o alcohol	0.000	0.159	0.442	0.000	0.994	0.890
8. In a car, that has stopped for a few minutes	0.000	0.772	0.501	0.000	0.000	0.009

Rasch DIF parameters for gender and age. Significant values are bold-faced and rounded to four significant digits. Significance levels have been Bonferroni-corrected for multiple comparisons. Gender×CInt (Class interval) and Age×CInt represent significance tests for non-uniform DIF. No significant non-uniform DIF was found, however

Bonferroni-corrected item fit residuals showed significant misfit for items 1, 3, 5, 7, and 8 (Table 3). The visual inspection indicated the largest misfit in item 5 (Fig. 2). An example of an item with good model fit (item 2) is presented in Fig. 3. When mean trait location was fixed at 0 on the logit scale, the mean person location of the sample was -1.093 (SD 1.186).

There were no reversed response alternatives in any of the items. Item difficulties ranged from -2.087 (item 5) to 1.98 (item 8). The figure shows a gap in item coverage in the upper part of the logit scale, between items 6 and 8 and the rest of the items.

DIF was demonstrated concerning age for items 3, 4, and 8 and concerning gender for item 3 when the item response

theory approach was used (Table 3). Young people tended to get higher scores on items 3 and 4, and lower scores on item 8 than older people. Females tended to get lower scores on item 3 than males.

Discussion

The main finding of this study was that three items (3, 4, and 8) of the ESS exhibits DIF with regard to age, and one item (3) showed possible DIF with regard to gender. Older people tend to score lower on items 3 and 4, and higher on item 8, and women tended to score lower than men on item 3. We found similar results as in earlier studies [5, 6] with regard to

Fig. 1 Person-item distributions. Person distribution is presented to the left and item distribution to the right of the logit scale. Items are presented as item numbers followed by a threshold (after the dot)

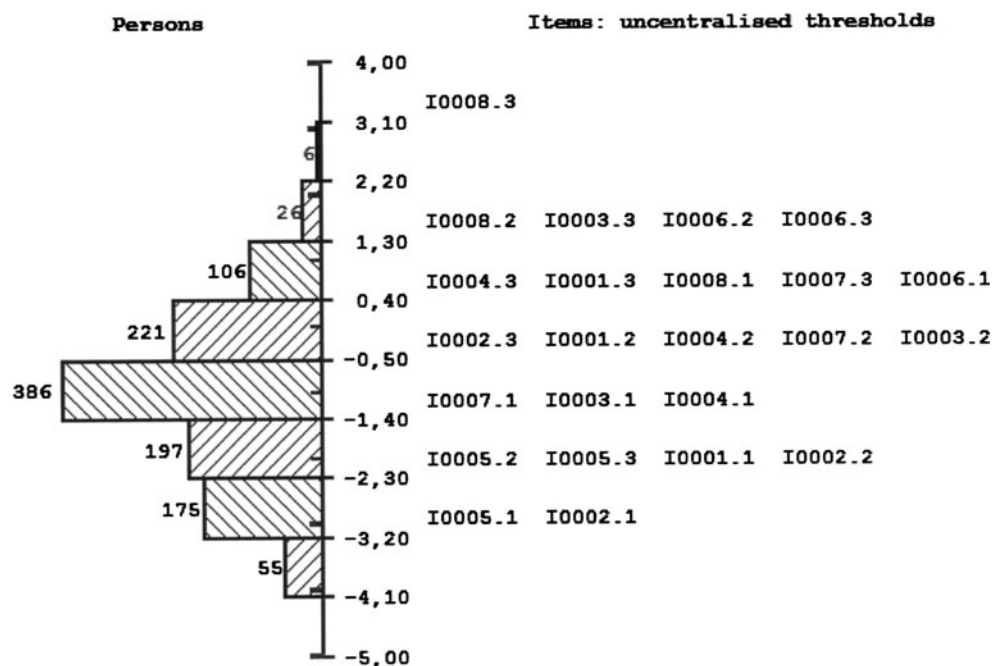
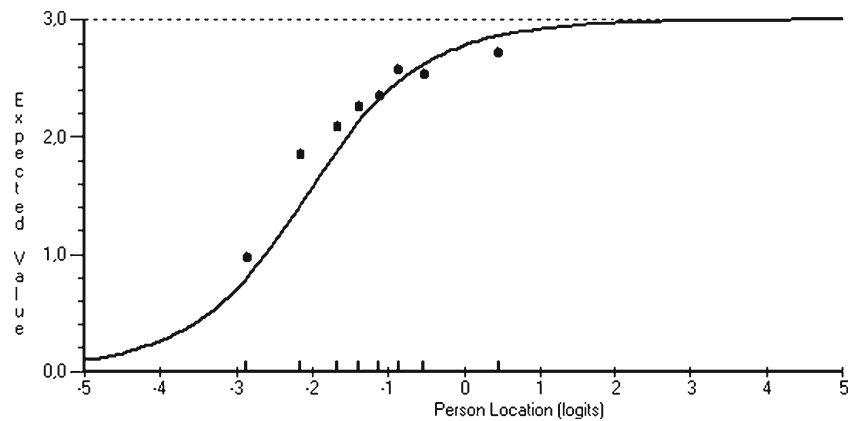


Fig. 2 Item characteristic curve for item 5. The item characteristic curve for item 5, which had the largest misfit to the Rasch model. Subjects are divided into eight classes depending on their ability. The *dots* represent the expected values for the classes of subjects



item hierarchy. Our reservation regarding the DIF for gender is due to the fact that only one of the analyses (i.e., the Rasch analysis) indicated this, while the ordinal regression did not. In our material, the females were significantly older than the males, and thus the detected DIF for gender might be caused by age differences.

An important assumption when using a questionnaire to assess a theoretical construct is that the responses given by a test taker only depend on the degree of that construct in the subject. DIF can be seen as a violation of this assumption. The present study is, as far as we know, the first one that has examined the ESS with regards to DIF in age and gender. Using two different approaches [8, 13], we found similar results. We split the material into two age groups for the Rasch analysis. Another possibility would be to split the material into smaller age groups. We did split the material into three age groups (i.e., <65 years, 65–74 years, and ≥ 75 years) based on the age distribution of the material. However, this did not change the findings (i.e., the same items showed DIF for age) (data not shown). The use of two groups with 65 year as cutoff was therefore chosen as this is the common age of retirement in Sweden. Differences in everyday life activities might be a cause of the DIF for age. One potentially important practical difference between working and being retired is the possibility to take naps.

Rather than falling asleep in different situations, such as meetings, daytime sleepiness might simply result in not taking part in such activities or compensating by caffeine consumption. These factors are not taken into account by the ESS, and differences in the possibilities to cope with daytime sleepiness in different ages might be one reason for the existence of DIF. Young people might get higher scores on items 3 and 4 than similarly sleepy older people, as young people might be more exposed to such situations. This raises a question regarding the reversed DIF in item 8, where older people scored higher than similarly sleepy young people. Both items 4 and 8 have to do with traffic, but unlike item 4, item 8 does not specify whether the subject drives or not. Retired people might drive less than non-retired people, and the reason for DIF in item 8 might be that young subjects respond to the question regarding the risk of falling asleep while driving, while old people respond regarding the risk of falling asleep as passengers. This is however only a hypothesis, and we have not tested it specifically.

The Rasch model showed significant misfit. Earlier studies [5, 6] have however shown good fit to Rasch models. The difference might be due to our much larger sample size. The order of the items in our study are similar to earlier findings [5, 6]. Visual inspection of the item characteristic curves shows generally acceptable fit. Item 5 shows the

Fig. 3 Item characteristic curve for item 2. The item characteristic curve for item 2, with good Rasch model fit. Subjects are divided into classes as in Fig. 2. The *dots* represent the expected values for the classes of subjects



largest misfit, but no DIF was detected in that item. It is important to combine the traditional, χ^2 -based goodness-of-fit test with a visual inspection of the ICC curves in order to avoid rejecting valid Rasch models, as not all statistically significant misfit is practically significant [14]. We believe that the visual inspection of the curves, the fact that the ordinal regression model (which is not influenced by Rasch misfit) indicated DIF in the same three items, and the fact that others have found a good model fit for Rasch models [5, 6], all together indicate that the DIF found in the present study was not simply secondary to Rasch model misfit.

DIF is, by definition, a form of model misfit. The Rasch model uses the total score of the scale as an estimation of the subject's sleepiness, and if some of the items are biased (e.g., by means of DIF), this might theoretically affect the total score and thus increase the risk for false DIF detection. However, as the DIF found in this study is relatively small, this is unlikely to be a cause of the DIF. Violani et al. [5] also found misfit problems with item 5, although in their study it did not reach statistical significance. Another issue in DIF research, as stressed by Teresi [15] is unidimensionality. We have not specifically studied the dimensionality of the ESS, but others have. While Johns [2] reported a one-factor solution both in presumably healthy medical students and in a patient group with obstructive sleep apnea syndrome, Smith et al. [7] were not able to confirm a single factor solution using confirmatory factor analysis. Hagell and Broman [6] found two factors with eigenvalues > 1.0 in an exploratory factor analysis. The item difficulty gap between items 6 and 8 and the other items might give rise to a spurious factor due to a relatively low degree of endorsement as compared to the other items [16]. As Hagell and Broman [6] point out, the seemingly conflicting results regarding factor structure might thus not contradict a single factor solution.

The fact that both the Rasch model and the ordinal regression model showed DIF with regard to age in the same items, strengthens this finding. However, statistically significant DIF might not be practically significant. This is especially important in large studies. Several methods have been proposed to assess DIF magnitude (reviewed in [15]). Cutoff values to assess DIF magnitude from pseudo- R^2 changes when grouping variables are added to the ordinal regression model have been suggested, but different authors have suggested strikingly different cutoff values, as reviewed by Jodoin and Gierl [17]. However, using any of these cutoff values, the DIF detected in the ESS in this study is probably not practically significant on an individual level. However, as the ESS is commonly used in large-scale epidemiologic studies, where even small DIF without practical significance for individual subjects might give rise to statistical significance, there might still be reasons for concern.

There are large differences in reported ESS scores between the general populations of different countries, which

might be understood as cultural and climatic differences [18]. Gender-related DIF might also differ between different cultures. In Sweden, a significant percentage of non-retired women work full time [19]. In a society where women traditionally stay at home, their ability to cope with daytime sleepiness might differ. Whether DIF between different cultures exists in the ESS has not been studied. However, Hayes et al. [20] have shown that African Americans score significantly higher on the ESS than white Americans. This was especially evident in items 2, 6, and 7, where the differences remained even after correction for different sleepiness risk factors as well as the subjects' responses to the other items. The authors do not clearly state that this is due to DIF, although they discuss it as one of several possibilities.

When Chervin and Aldrich [4] compared ESS scores to MSLT sleep latency, they found that male gender was associated to a 2.2-point lower ESS score when age and MSLT sleep latency were corrected for. This might indicate differential item functioning with regard to gender.

Chervin and Aldrich [4] did not use total ESS score in their regression model, but MSLT sleep latency. While some studies have found a weak correlation (reviewed in [3]), Chervin and Aldrich [4] found no significant correlation at all. MSLT and ESS might very well measure two different constructs [21]. Most of the items in the ESS concern situations when sleep is undesired. Conversely, in the MSLT, the subject is explicitly asked to close the eyes and relax, and external factors that might disturb sleep are minimized. While Chervin and Aldrich [4] used total ESS score as the dependent variable, we used ESS scores on individual items. Olson et al. [22] have shown that ESS correlates independently with Global Severity Index as well as with all subscales except psychoticism in the SCL-90, an inventory to assess subjective psychiatric symptoms [23]. The authors speculate that the explanation might be an underlying tendency to give affirmative responses to any questions. It might also be explained by some sort of DIF, although we have not tested for DIF based on psychological symptoms. It might be argued, for example, that depressive symptoms could be related to altered patterns of daily activities, and that this might be a cause of DIF.

The item hierarchy, and the gap in item difficulty between items 6 and 8 and the rest of the items, have been described before, both by Violani et al. [5] and by Hagell and Broman [6]. Hagell and Broman examined the psychometric properties of the ESS in Parkinson Disease patients, and Violani studied a sleep clinic cohort with various sleep-related disorders. They both report similar item hierarchies with item 5 measuring the lowest degree of sleepiness and items 6 and 8 the highest. They both report a gap between items 6 and 8 and the others, and the only difference between the two studies is the order of items 4 and 7. Our findings are in line with theirs.

A potential cause for DIF that has not been studied is related to the ambiguous wording of item 8 (“in a car that has stopped for a few minutes in traffic”). As the item does not specify whether the respondent is the driver or a passenger, and as these situations probably differ with regard to soporificity, DIF due to the respondent’s interpretation of the item might exist. Another potential problem with the item might be that respondents are tempted to “fake good” in order to keep their driver’s licenses. A Peruvian version of the ESS, especially aimed at non-driving subjects, replaced item 8 with “standing and leaning or not[sic] on a wall or furniture”, but they did not report the degree of sleepiness needed to score on this item [24].

Other factors that are not directly related to sleepiness might affect a subject’s ESS score as well. Chin et al. [25] found differences in ESS scores in OSAS patients before and after initiation of CPAP-treatment. When they completed the ESS prior to CPAP initiation and then retrospectively completed it again when they were on CPAP (i.e., they were asked to remember their degree of sleepiness prior to treatment). The authors interpreted this finding as a response shift, i.e., that the patients did not fully recognize their degree of sleepiness before they had access to therapy. Kaminska et al. [26] reported that patients tend to get a lower ESS score if the physician scored the ESS based on an interview with the patients, as compared to when the test was self-administered. The reason for this is unclear, but it the problem was not restricted to individual items. The authors speculate that social desirability bias or recall bias might have affected the results in the physician-administered test. The ESS data in our study are from self-administered tests. Kumru et al. [27] have compared self-rated sleepiness to how a partner rated the patient’s sleepiness and found that partners rated the probability of the patient falling asleep significantly higher than the patients, especially regarding intermediately difficult items (i.e., items 1–4 and 7). This too might indicate that social desirability bias or recall bias might affect the ESS score.

Others e.g., [28] have pointed to the lack of a clear description of the item generation process in the ESS. Items were chosen to represent different levels of soporificity, but the process is not described in detail. One way to address this would be to use a qualitative approach to interview subjects with different levels of sleepiness according to the MSLT or MWT, and use information from these interviews to generate potential items for a sleepiness assessment questionnaire. In conclusion, while the DIF detected in the ESS in this study is probably too small to affect patients on an individual basis, it might be a cause of concern in large-scale studies. As there is a great need for a quick, easy, and cheap way to assess daytime sleepiness, future research should focus on developing new methods to do so.

References

1. Johns MW (1991) A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. *Sleep* 14:540–545
2. Johns MW (1992) Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep* 15:376–381
3. Johns MW (2000) Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: failure of the MSLT as a gold standard. *J Sleep Res* 9:5–11
4. Chervin RD, Aldrich MS (1999) The Epworth Sleepiness Scale may not reflect objective measures of sleepiness or sleep apnea. *Neurology* 52:125–131
5. Violani C, Lucido F, Robusto E, Devoto A, Zucconi M, Ferini Strambi LF (2003) The assessment of daytime sleep propensity: a comparison between the Epworth Sleepiness Scale and a newly developed Resistance to Sleepiness Scale. *Clin Neurophysiol* 114:1027–1033
6. 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
7. Smith SS, Oei TPS, Douglas JA, Brown I, Jorgensen G, Andrews J (2008) Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnea. *Sleep Med* 9:739–744
8. Hambleton RK, Swaminathan H, Rogers HJ (1991) *Fundamentals of Item Response Theory*. Sage, Newbury Park
9. Broström A, Arestedt KF, Nilssen P, Strömberg A, Ulander M, Svanborg E (2010) The side effects to CPAP treatment inventory: the development and initial validation of a new tool for the measurement of side-effects to CPAP treatment. *J Sleep Res* 19:603–611
10. Ståhlkrantz A, Sunnergren O, Ulander M, Albers J, Mårtensson J, Johansson P, Svanborg E, Broström A (2009): Difference in insomnia symptoms, daytime sleepiness, fatigue and global perceived health between hypertensive patients with or without risk of obstructive sleep apnea in a primary care setting. *Eur J Cardiovasc Nurs* 8S1, S24
11. Broström A, Strömberg A, Dahlström U, Fridlund B (2004) Sleep difficulties, daytime sleepiness, and health-related quality of life in patients with chronic heart failure. *J Cardiovasc Nurs* 19:234–242
12. Johansson P, Alehagen U, Svanborg E, Dahlström U, Broström A (2009) Sleep disordered breathing in an elderly community-living population: relationship to cardiac function, insomnia symptoms and daytime sleepiness. *Sleep Med* 10:1005–1011
13. Zumbo BD (1999) *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa, ON, Directorate of Human Resources Research and Evaluation. Department of National Defense
14. Hagquist C (2001) Evaluating composite health measures using Rasch modelling: an illustrative example. *Soz Präventivmed* 46:369–378
15. Teresi J (2006) Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Med Care* 44:S152–S170
16. Nunnally JC, Bernstein IH (1999) *Psychometric theory*, 3rd edn. McGraw-Hill, New York
17. Jodoin MG, Gierl MJ (2001) Evaluating type I error and power using an effect size measure with the logistic regression procedure for DIF detection. *Appl Meas Educ* 14:329–349
18. Soldatos RS, Allaert FA, Ohta T, Dikeos DG (2005) How do individuals sleep around the world? Results from a single-day survey in ten countries. *Sleep Med* 6:5–13
19. SCB (Statistics Sweden): Labour Force Survey, August 2010; pp 1–270.

20. Hayes AL, Spilsbury JC, Patel SR (2009) The Epworth score in African American populations. *J Clin Sleep Med* 5:344–348
21. Sangal RB, Mitler MM, Sangal JM (1999) Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy. *Clin Neurophysiol* 110:2131–2135
22. Olson L, Cole M, Ambrogetti A (1998) Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms. *J Sleep Res* 7:248–253
23. Derogatis LR, Rickels K, Rock AF (1976) The SCL-90 and the MMPI: a step in the validation of a new self-report scale. *Br J Psychiatry* 128:280–289
24. Rosales-Mayor E, Rey de Castro J, Huayanay L, Zagaceta K (2011) Validation and modification of the Epworth Sleepiness Scale in Peruvian population. *Sleep Breath* [Epub ahead of print] doi: [10.1007/s11325-011-0485-1](https://doi.org/10.1007/s11325-011-0485-1)
25. 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
26. Kaminska M, Jobin V, Mayer P, Amyot R, Perraton-Brillon M, Bellemare F (2010) The Epworth Sleepiness Scale: self-administration versus administration by the physician, and validation of a French version. *Can Respir J* 17:e27–34
27. 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
28. Miletin MS, Hanly PJ (2003) Measurement properties of the Epworth sleepiness scale. *Sleep Med* 4:195–199
29. Broström A, Strömberg A, Mårtensson J, Ulander M, Harder L, Svanborg E (2007) Association of Type D personality to perceived side effects and adherence in CPAP-treated patients with OSAS. *J Sleep Res* 16:439–447