Measurement properties and hierarchical item structure of the Epworth Sleepiness Scale in Parkinson's disease

Peter Hagell, RN PhD[1,2], Jan-Erik Broman, RN PhD[3]

[1] Department of Health Sciences, Lund University, Lund, Sweden
[2] Department of Neurology, University Hospital, Lund, Sweden
[3] Department of Neuroscience, Psychiatry, Uppsala University, University Hospital, Uppsala, Sweden

Short title: ESS in PD

Number of words (main text): 3216

Number of words (abstract): 224

Corresponding author:
Peter Hagell
Department of Health Sciences
Lund University
P.O. Box 157
SE-221 00 Lund
Sweden
Tel: +46 46 222 1937
Fax: +46 46 222 1934
E-mail: Peter.Hagell@med.lu.se

Abstract

The aim of this work was to evaluate the measurement properties and hierarchical item structure of the Epworth Sleepiness Scale (ESS) in patients with Parkinson's disease (PD). Data were taken from a cross-sectional study regarding fatigue and sleep related aspects of PD. One hundred and eighteen consecutive patients with neurologist diagnosed PD without significant co-morbidities (54% men; mean age, 64; mean PD duration, 8.4 years) from four Swedish neurological outpatient clinics participated. The ESS displayed good data quality with few missing items (0-2.5%), good reliability (Cronbach's alpha, 0.84), marginal floor and no ceiling effects (1.7 and 0%, respectively), and differentiated between those reporting problems staying awake during the past month and those who did not. Item-total correlations, factor and Rasch analyses indicated that items tap a single underlying construct. Rasch analysis supported basic rating scale assumptions and demonstrated an item hierarchy similar to that previously found in patients with other sleep disorders. Gaps in the levels of sleep propensity covered by ESS items and their response options were identified at the higher and lower ends of the underlying sleepiness continuum. This study provides an evidence base for using the ESS in PD by demonstrating good psychometric properties and a stable hierarchical item structure. However, addition of new items and use of Rasch scoring has potential to further enhance the clinical usefulness of the ESS.

Key words: ESS, excessive daytime sleepiness, Parkinson's disease, Rasch analysis, reliability, validity

Introduction

The Epworth Sleepiness Scale (ESS) was devised to operationalize common situations that differ in their ability to induce dozing (Johns, 1991, 1992, 1994, 2002). Its eight items define various postures, activities and environmental settings that are expected to vary in their likelihood to induce sleepiness. For example, whereas many people (also without abnormal sleepiness) may tend to doze off while lying down to rest in the afternoon (ESS item 5), only those with high levels of sleep propensity would be expected to do so while sitting and talking to someone (ESS item 6). Studies have largely supported these assumptions, although some variations in item hierarchy have been found between patients and non-patients, as well as between cultures (Chen *et al.,* 2002; Johns, 1994, 2002; Violani *et al.,* 2003).

The vast majority of studies on the measurement properties and item hierarchy of the ESS have involved subjects with sleep-disordered breathing, primary sleep disorders and healthy subjects (Bloch *et al.,* 1999; Chen *et al.,* 2002; Johns, 1992, 2002; Violani *et al.,* 2003). However, daytime sleepiness is a common feature of a variety of disorders including, e.g., Parkinson's disease (PD) (Arnulf, 2005). A number of studies have addressed daytime sleepiness in PD, of which most have used the ESS (for review, see Arnulf, 2005). However, in order to fully enable valid interpretation of such studies the measurement properties of the ESS in PD must be considered since psychometric properties cannot be assumed to be stable across respondent groups (McHorney *et al.,* 1994). So far, assessments of psychometric properties of the ESS in PD have been incomplete (Marinus *et al.,* 2003). Given the potential role of dopaminergic and related circuits in daytime sleepiness in PD (Rye, 2004), it appears particularly relevant to consider whether the item hierarchy of the ESS found in other patient groups is stable also among people with PD. If so, this would not only support the validity of the ESS in PD, but also the stability of the construct of sleep propensity across disorders. This study assesses the measurement properties and hierarchical item structure of the ESS in PD.

Methods

Data were taken from a cross-sectional, multi-center study regarding fatigue and sleep related aspects of PD (Hagell *et al.,* in press).

Patients

One hundred and twenty four consecutive Swedish speaking people with neurologist diagnosed PD from four Swedish medical centers were invited to participate in the study. Patients with ongoing infections, psychiatric drug adverse reactions, clinically significant co-morbidities (including depression and cognitive impairment), as determined by patients' attending neurologist and the study assessors at the time of assessment, and those participating in other ongoing studies were excluded. Four eligible patients declined participation, one did not have time to participate, and one was found not to meet inclusion criteria, leaving a total of 118 participants (Table 1). There were no significant differences in dropout rates across study sites. The study was approved by the respective Research Ethics Committees.

- Table 1 about here -

Procedures

After providing written informed consent, patients were assessed clinically by means of the Unified PD Rating Scale, the Hoehn & Yahr staging of PD, the Schwab & England activities of daily living scale, and the Mini-Mental State Exam (Fahn *et al.,* 1987; Hoehn and Yahr, 1967; Folstein *et al.,* 1976). Patients then completed the ESS and questionnaires tapping fatigue, sleep quality, depression, anxiety, perceived adjustment to illness, and illness-

related distress. All assessments were performed during the "on" phase (i.e., periods of good drug response and no or minimal PD-related disability). Hoehn & Yahr and Schwab & England were also estimated for the "off" phase (i.e., periods of poor drug response and increased PD-related disability) based on patient reported history and medical records. Aspects beyond the objectives of this study are not considered here.

Epworth Sleepiness Scale (ESS)

The ESS specifies eight everyday situations for which respondents are requested to indicate how likely they have been to doze off or fall asleep during recent times (Johns, 1991). Response options range from 0 (= would never doze) to 3 (= high chance of dozing). Summation of the eight item scores yields a total ESS score between 0 and 24, indicating the subject's average sleep propensity in daily life (24 = more sleep propensity). In this study, the Swedish version of the ESS was used (Broman *et al.,* 2000).

Analyses

ESS data were analyzed regarding data quality, reliability, floor- and ceiling effects, construct validity, unidimensionality, rating scale functionality and item hierarchy using traditional and Rasch measurement approaches (McHorney *et al.,* 1994; Nunnally and Bernstein, 1994; Rasch, 1960; Wright and Masters, 1982).

Item completion rates were examined to assess data quality, which is considered high when the percentage missing data is low (McHorney *et al.,* 1994). Reliability was assessed by Cronbach's coefficient alpha, a measure of item interrelatedness and an estimate of reliability, that preferably should be around 0.8-0.9, and not below 0.7 (Streiner, 2003).

Floor and ceiling effects represent the percentage of respondents obtaining the lowest and highest possible scores, respectively. Floor/ceiling effects influence scale responsiveness adversely and should preferably not exceed 15% (McHorney and Tarlov, 1995).

Construct validity refers to whether an instrument measures what it is intended to measure and was assessed by means of "known-groups" comparison. Respondents were grouped into two groups depending on their responses to item 8 of the Pittsburgh Sleep Quality Index (PSQI), which inquires about problems staying awake during daily activities (Buysse *et al.,* 1989). ESS scores were compared between patients indicating that they had not experienced any such problems during the past month (score 0 on PSQI item 8) and those indicating that they had. Significantly higher ESS scores were expected for the latter group.

Unidimensionality, i.e., the extent to which items tap a single underlying latent construct, is a basic assumption behind the use of summed scores and was evaluated by several approaches. First, principal component exploratory factor analysis (EFA) with determination of the number of factors to extract by means of parallel analysis (Thompson and Daniel, 1996) was performed. One thousand parallel sets of random ESS data (i.e., the same number of items and persons) were thus generated followed by independent parallel EFAs of the empirical and random data matrices (O'Connor, 2000). For each consecutive empirical eigenvalue that exceeded the 95th percentile of the distribution of random data eigenvalues, a factor was retained (O'Connor, 2000). Extraction of only one such factor was interpreted in support of unidimensionality. Second, the extent to which each item is substantially related to the underlying construct was assessed by determining the corrected item-total correlations (i.e., correlation between each item and the total score of the other items; Howard and Forehand, 1962), which should exceed 0.40 (McHorney *et al.,* 1994). Third, the ESS was subjected to analyses according to the Rasch rating scale measurement model (Wright and Masters, 1982). A fundamental assumption in this model is that each item

contributes to the measurement of a single underlying latent construct. Unidimensionality was thus assessed by determining each item's model fit, expressed as item mean-square (MNSQ) information-weighted goodness-of-fit (INFIT) statistics. MNSQ has an expected value of 1; for rating scales, values ≤1.4 are considered appropriate and were interpreted as support for unidimensionality (Wright and Linacre, 1994).

We also used Rasch analysis to examine basic rating scale assumptions (Linacre, 2002). Rasch analysis estimates the locations of each person, item, and response category on the underlying latent variable according to a logit (log-odd unit) scale that is a logarithmic metric ranging from minus infinity to plus infinity (with mean item location set at zero). To meet basic rating scale assumptions, the locations of transition thresholds between categories (i.e., the points where there are 50/50 probabilities of endorsing adjacent categories) and average category measures should be ordered in an expected manner (from less to more) on the underlying latent continuum. Disordering indicates that the rating scale does not work as intended in the sample (Linacre, 2002).

Similarly, to investigate the hierarchical item structure of the ESS, the amount of sleep propensity represented by each item was determined by means of the Rasch calibrated locations of each item on the underlying latent variable. As opposed to studying the rank ordering of item scores, the logit metric measures at the interval level, which provides a means to assess the distances between items (Wright and Masters, 1982). We also explored the ESS' coverage of the underlying latent sleep propensity continuum by examining the location of the Rasch calibrated transition threshold points between the rating scale categories for each item.

Variables were checked regarding assumptions underlying the use of parametric and non-parametric statistics, and analyzed accordingly. Analyses were performed using SPSS

version 12 (SPSS Inc., Chicago, IL), ScoreRel CI (Barnette, 2005), and WINSTEPS version 3.55 (Winsteps.com, Chicago, IL). The alpha-level of significance was set at 0.05 (2-tailed).

Results

Table 2 summarizes descriptive and psychometric statistics of the ESS. No patients expressed any difficulties understanding or completing the questionnaire. Data quality was good, with few missing item responses. Item 1 ("sitting and reading") had three missing responses (2.5%) and item 6 ("sitting and talking to someone") had none. Other items had one missing response (0.8%) each. Total ESS scores could be computed for all but four patients (3.4%). Coefficient alpha was 0.84 and remained stable (0.80-0.83) when deleting one item at a time. Floor and ceiling effects were marginal and absent, respectively, and known-groups comparison of ESS scores showed a highly significant ($P<0.0001$) difference in the expected direction (Table 2).

- Table 2 about here -

Results provided support for the unidimensionality of the ESS according to all pre-defined criteria. Only the first EFA derived factor (items 1, 2, 5 and 7; Fig. 1) had an eigenvalue exceeding that from random data. The eigenvalues of the first two empirical factors were 3.86 and 1.15, and from the corresponding random data they were 1.54 and 1.34. Hence, the first empirical and random factors explained 48.2% and 19.2% of the total variance, respectively. The second factor explained an additional 14.4% (ESS data) and 16.8% (random data) of the total variance. Corrected item-total correlations ranged between 0.46 and 0.71 (mean, 0.57). Rasch analyses did not identify any items displaying signs of

misfit to the unidimensional measurement model (Table 3), and supported basic rating scale assumptions (Fig. 2).

- Fig. 1, Table 3 and Fig. 2 about here -

The right hand side of Fig. 3 depicts the Rasch-derived sample distribution and the middle section of the same Figure depicts the hierarchical item structure of the ESS, according to their locations along the sleep propensity continuum (y-axis). Table 3 gives the logit measures corresponding to the item locations depicted in Fig. 3. The measurement range and coverage of the ESS is illustrated to the left in Fig. 3, which plots the thresholds between each rating scale category (i.e, the intersections between categories 0-to-1, 1-to-2, and 2-to-3) for the eight items. The measurement range covered 7.64 logits (-3.92 to 3.72). Two major gaps were detected towards the respective ends of the measurement range (solid arrows in Fig. 3); at the end representing more sleep propensity there was a 1.45 logit gap and at the other end there was a 1.08 logit gap. The distance between other rating scale category thresholds ranged between 0 and 0.64 logits, of which four exceeded 0.5 logits (range, 0.52-0.64 logits; dashed arrows in Fig. 3).

- Fig. 3 about here -

Discussion

This is the first study to comprehensively document the psychometric properties and hierarchical item structure of the ESS in PD. The questionnaire appears highly acceptable among respondents, as indicated by excellent data quality, reliability was good and floor/ceiling effects were minimal. Construct validity, unidimensionality and basic rating

scale assumptions of the ESS were also supported, and its item hierarchy was found very similar to that in previously reported populations.

The sample studied here exhibited ESS scores similar to those previously reported in PD (Arnulf, 2005) and represented a wide range of PD stages and durations. However, the overall PD severity was somewhat skewed towards less severe stages and enrollment criteria excluded patients with clinically significant co-morbidities. While it appears unlikely that this has had major influences on the main findings, it may pose some limitations to the generalizability of results. Additional studies are therefore encouraged.

Marinus *et al.* (2003) have reported some measurement properties of the ESS in PD. In agreement with their findings, our principal component EFA grouped items into two components with eigenvalues above 1. The identification of two factors with an eigenvalue >1 in the empirical data set is not surprising and does not contradict unidimensionality. Because EFA is based on correlations, and correlations between variables increase when they have similar distributions, items that are endorsed by few (or many) respondents tend to cluster together due to their distributional properties even if they measure the same construct as other items (Nunnally and Bernstein, 1994). Indeed, comparing the factor loadings with the item hierarchy reveals that the items that loaded strongest on the first and second factors are those measuring at the lower and higher ends of the sleepiness continuum, respectively. As suggested by item-total correlations, Rasch and parallel analyses, but in contrast to the interpretation by Marinus *et al.* (2003), the two empirical EFA derived factors are therefore most probably not reflecting multidimensionality. Instead, the two factors appear to represent so called “difficulty factors” and illustrate potential pitfalls associated with item level EFA (Nunnally and Bernstein, 1994).

The hierarchical item structure of the ESS in PD was found very similar to that previously documented in various sleep disorders (Chen *et al.,* 2002; Johns, 2002; Violani *et*

al., 2003). In this study, items 6 and 8 represented the most severe levels of sleep propensity. This is in accordance with previous observations based on rank ordering of items. Although item 8 often has been ranked higher than item 6, the difference between the two across studies has not been significant (Johns, 2002). Similarly, non-significant differences in rank ordering have also been found between items 1 and 4, and items 3 and 7 (Johns, 2002). Therefore, and because ranks (as opposed to Rasch analysis) do not provide linear item location estimates, the ESS item hierarchy found in other patient groups can most probably be considered stable also in PD. This supports the internal construct validity of the ESS in PD.

Violani *et al.* (2003) also used the Rasch model to assess the ESS in an Italian sample with various sleep disorders. While the ordering of items 4 and 7 was reversed in that study compared to what was found here, it is unclear if these differences are significant and/or due to differential item functioning (DIF) between PD and sleep disorders, or between the Swedish and Italian ESS. The presence of DIF means that an item displays different statistical properties and has different meanings across subgroups of respondents (Tennant *et al.,* 2004). Additional studies are necessary to address the potential presence and significance of DIF between various diagnostic and cultural groups.

Two major measurement gaps, at the lower and higher ends of the sleep propensity continuum, were identified. This has implications for measurement precision, which decreases at levels corresponding to such gaps (Wright and Masters, 1982). However, since the ESS mainly is used to detect excessive sleepiness, the gap at the lower end of the continuum is probably of less clinical concern. In agreement with our findings, studies suggest that the ESS may not be able to differentiate between higher levels of sleepiness. For example, Sangal *et al.* (1999) compared the ESS and the maintenance of wakefulness test (MWT) in a group of 522 narcoleptics and found stable ESS scores as the MWT was increasing among more severely affected individuals. Addition of new items that define

situations corresponding to the gap between item 3 and items 6/8, and that extend the range of measurement beyond items 6/8, could thus provide a means of improving measurement precision and, hence, sensitivity within this range. In an attempt to increase the ability to identify PD patients at risk of falling asleep while driving, Hobson *et al.* (2002) added four new items regarding falling asleep in unusual situations (while driving, eating, working, and doing household activities) to the ESS. This improved specificity and the new items had lower endorsement rates than any of the original ESS items, indicating that while they probably extend the range of measurement they would be less likely to fill gaps within the ESS range of measurement. However, further studies using these items together with the ESS are needed to assess the levels of sleepiness that the expanded item set represents. Similarly, strategies to improve the coverage of daytime sleepiness could involve co-administration of the ESS together with other daytime sleepiness items from existing scales and/or new items specifically devised for this purpose.

Using Rasch derived logit scores rather than raw ESS scores has potential to improve measurement of self reported sleep propensity. As opposed to raw scores, logit scores are linear and not confined to ordinal integers. Furthermore, the relationship between raw scores and linear measures takes an *S*-shaped (ogive) form, which has implications when measuring change, particularly at the higher and lower ends of the continuum. For example, the same raw score change (or difference) at the high end and at the mid range of a scale can represent linear changes that differ five-fold in magnitude (Wright, 1997). Logit scores have the potential of removing such measurement bias, which is a particularly attractive feature when assessing individuals (Cella and Chang, 2000). Expanding the ESS item pool at the more severe end and use of Rasch logit scores may thus have the potential to improve its usefulness as an easy and inexpensive clinical tool for individual patient assessment. Such developments appear particularly warranted when considering risk assessments regarding, e.g., falling

asleep while driving, an issue of outmost importance for patient as well as public safety (Hobson *et al.,* 2002; Horne and Reyner, 1999).

Recently, the PD-specific SCOPA-SLEEP questionnaire, tapping nighttime sleep problems (5 items) and daytime sleepiness (6 items), was proposed (Marinus *et al.,* 2003). Items of the daytime sleepiness scale are similar to those in the ESS, which was used to validate the new questionnaire. Comparing the performance of the SCOPA-SLEEP (sleepiness subscale) with the ESS reveals few notable differences. However, SCOPA-SLEEP exhibited larger floor and ceiling effects (12.1% and 0.7%, respectively) than those found here. It also appears less sensitive than the ESS, as indicated by failure to detect differences between subsets of patients that were found to differ according to the ESS, and a larger coefficient of variation (inappropriately interpreted as an indication of better sensitivity by the authors) (Marinus *et al.,* 2003). However, comprehensive head-to-head comparisons are needed to assess the relative merits of the two.

In conclusion, this study has provided an evidence base for the use of the ESS in patients with PD by demonstrating good psychometric properties. While additional studies are needed to replicate these results and to assess DIF between different diagnostic and cultural groups, our observations support the psychometric properties and indicate that the hierarchical item structure of the ESS is maintained in PD. However, addition of new items, representing more pronounced sleepiness, and use of Rasch scoring has potential to further enhance the clinical usefulness of the ESS.

Acknowledgements

The authors want to thank all participating patients for their cooperation, B. Eriksson, A. Höglund, I. Knutsson, and J. Reimer for assistance with data collection, and Drs. J. Ahlberg, N. Dizdar, H. Edwall, B. Johnels, J. Lökk, S. Pålhagen, H. Widner, and T. Willows for patient recruitment.

The study was supported by the Swedish Parkinson Foundation, the Swedish Research Council, the World Federation of Neuroscience Nurses, the Skane County Council Research and Development Foundation, the Vårdal Institute, the Department of Nursing at Lund University, and the Inge & Elsa Andersson's Parkinson's disease Research Foundation.

References

Arnulf, I. Excessive daytime sleepiness in parkinsonism. *Sleep Med. Rev*., 2005, 9: 185-200.

Barnette, J. J. ScoreRel CI: an Excel program for computing confidence intervals for commonly used score reliability coefficients. *Educ. Psychol. Meas*., 2005, 65: 980-983.

Bloch, K. E., Schoch, O. D., Zhang, J. N. and Russi, E. W. German version of the Epworth Sleepiness Scale. *Respiration*, 1999, 66: 440-447.

Broman, J. E., Bengtson, H. and Hetta, J. Psychometric properties of a Swedish version of the Epworth Sleepiness Scale. *J. Sleep Res*. 2000, 9 Suppl 1: 27.

Buysse, D. J., Reynolds, C. F., 3rd, Monk, T. H., Berman, S. R. and Kupfer, D. J. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res*., 1989, 28: 193-213.

Cella, D. and Chang, C. H. A discussion of item response theory and its applications in health status assessment. *Med. Care*, 2000, 38: II66-72.

Chen, N. H., Johns, M. W., Li, H. Y., Chu, C. C., Liang, S. C., Shu, Y. H., Chuang, M. L. and Wang, P. C. Validation of a Chinese version of the Epworth sleepiness scale. *Qual. Life Res*., 2002, 11: 817-821.

Fahn S, Elton RL, members of the UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. In: S. Fahn, C. D. Marsden, D. B. Calne and M. Goldstein (Eds), *Recent Developments in Parkinson's Disease, Vol. 2*. MacMillan Healthcare Information, Florham Park, 1987: 153-163.

Folstein, M. F., Folstein, S. E. and McHugh, P. R. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res*., 1975, 12: 189-198.

Hagell, P., Höglund, A., Reimer, J., Eriksson, B., Knutsson, I., Widner, H. and Cella, D. Measuring fatigue in Parkinson's disease: a psychometric study of two brief generic fatigue questionnaires. *J. Pain Symptom Manage*., in press.

Hobson, D. E., Lang, A. E., Martin, W. R., Razmy, A., Rivest, J, and Fleming, J. Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease: a survey by the Canadian Movement Disorders Group. *JAMA*, 2002, 287: 455-463.

Hoehn, M. M. and Yahr, M. D. Parkinsonism: onset, progression and mortality. *Neurology,* 1967, 17: 427-442.

Horne, J. and Reyner, L. Vehicle accidents related to sleep: a review. *Occup. Environ. Med*., 1999, 56: 289-294.

Howard, K. I. and Forehand, G. A. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educ. Psychol. Meas*., 1962, 22: 731-735.

Johns, M. W. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*, 1991, 14: 540-545.

Johns, M. W. Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep*, 1992, 15: 376-381.

Johns, M. W. Sleepiness in different situations measured by the Epworth Sleepiness Scale. *Sleep*, 1994, 17: 703-710.

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

Linacre, J. M. Optimizing rating scale category effectiveness. *J. Appl. Meas*., 2002, 3: 85-106.

Marinus, J., Visser, M., van Hilten, J. J., Lammers, G. J. and Stiggelbout, A. M. Assessment of sleep and sleepiness in Parkinson disease. *Sleep*, 2003, 26: 1049-1054.

McHorney, C. A. and Tarlov, A. R. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual. Life Res*., 1995, 4: 293-307.

McHorney, C. A., Ware, J. E., Jr., Lu, J. F. and Sherbourne, C. D. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med. Care*., 1994, 32: 40-66.

Nunnally, J. C. and Bernstein, I. H. *Psychometric Theory, 3rd edition*. McGraw-Hill, Inc., New York, 1994.

O'Connor, B. P. SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test. *Behav. Res. Methods Instrum. Comput*., 2000, 32: 396-402.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut, Copenhagen, 1960 (Reprinted: University of Chicago Press, Chicago, 1980).

Reimer, J., Grabowski, M., Lindvall, O. and Hagell, P. Use and interpretation of on/off diaries in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry*, 2004, 75: 396-400.

Rye, D. B. The two faces of Eve: dopamine's modulation of wakefulness and sleep. *Neurology*, 2004, 63: S2-7.

Sangal, R. B., Mitler, M. M. and Sangal, J. M. 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*., 1999, 110: 2131-2135.

Streiner, D. L. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. Pers. Assess*., 2003, 80: 99-103.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., Lawton, G., Simone, A., Carter, J., Lundgren-Nilsson, A., Tripolski, M., Ring, H., Biering-Sorensen, F., Marincek, C., Burger, H. and Phillips, S. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential

item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med. Care*, 2004, 42: I37-48.

Thompson, B. and Daniel, L. G. Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines. *Educ. Psychol. Meas*., 1996, 56: 197-208.

Violani, C., Lucidi, F., Robusto, E., Devoto, A., Zucconi, M. and Ferini Strambi, L. The assessment of daytime sleep propensity: a comparison between the Epworth Sleepiness Scale and a newly developed Resistance to Sleepiness Scale. *Clin. Neurophysiol*., 2003, 114: 1027-33.

Wright, B. D. Fundamental measurement for outcome evaluation. *Physical Medicine and Rehabilitation: State of the Art Reviews*, 1997, 11: 261-288.

Wright, B. D. and Linacre, J. M. Reasonable mean-square fit values. *Rasch Measurement Transactions*, 1994, 8: 370.

Wright, B. D. and Masters, G. N. *Rating Scale Analysis*. MESA Press, Chicago, 1982.

Table 1 Sample characteristics (n=118)

Gender (men / women)	64 (54%) / 54 (46%) [a]
Age (years)	63.9 (9.6; 41-82) [b]
Time since PD diagnosis (years)	8.4 (5.7; 0.4-30.7) [b]
Daily anti-PD medication [d]	780 (518-1110; 0-5580) [c]
Hoehn & Yahr stage of PD (during "on") [e]	II (II-III; I-IV) [c]
Hoehn & Yahr stage of PD (during "off") [e]	III (II-III; I-V) [c]
Schwab & England ADL score (during "on") [f]	90 (90-100; 40-100) [c]
Schwab & England ADL score (during "off") [f]	90 (80-90; 10-100) [c]
UPDRS motor score (during "on") [g]	17 (10.5-27; 1-50) [c]
MMSE score [h]	29 (28-30; 23-30) [c]
PSQI total score [i]	7 (4-10; 1-18) [c]

[a] n (%).

[b] Mean (standard deviation; min-max).

[c] Median (interquartile range; min-max).

[d] Expressed as total levodopa equivalent dose: 100 levodopa equivalents = 100 mg standard levodopa = 133 mg controlled-release levodopa = 10 mg bromocriptine = 5 mg ropinirole = 1 mg pramipexole = 1 mg cabergoline = 2 mg apomorphine. For patients who received a COMT-inhibitor, the sum of standard levodopa and 0.75 times the dose of controlled-release levodopa were multiplied by 1.3 (Reimer *et al.,* 2004).

[e] Range, I-V (I = mild unilateral disease; II = Bilateral disease without postural impairment; III = Bilateral disease with postural impairment, moderate disability; IV = Severe disability, still able to walk and stand unassisted; V = Confined to bed or wheelchair unless aided).

[f] Range, 0-100 (100 = normal ADL functioning).

[g] Range, 0-108 (0 = no signs of parkinsonism).

[h] Range, 0-30 (30 = normal cognition).

[i] Range, 0-21 (0 = no sleep quality impairments).

PD, Parkinson's disease; ADL, activities of daily living; UPDRS, Unified Parkinson's Disease Rating Scale; MMSE, Mini-Mental State Exam; PSQI, Pittsburgh Sleep Quality Index.

Table 2 Descriptive and psychometric statistics for the ESS (n=118)

Score mean (SD)	9.6 (5.0) [a]
Score median (IQR)	10 (6-13) [a]
Data quality	
Missing item responses (%)	0.95 (0-2.5) [b]
Computable scale scores (%)	96.6
Reliability	
Cronbach's alpha (95% CI) [c]	0.84 (0.79-0.88)
Floor/ceiling effects (%) [d]	1.7 / 0
Construct validity [e]	
No problems staying awake during past month (n=81)	8 (5-12)
Problems staying awake during past month (n=30)	14 (7-16) [f]

[a] Score range, 0-24 (24 = more sleep propensity).
[b] For whole questionnaire (range for each item).
[c] Should preferably be around 0.8 and not below 0.70.
[d] Should not exceed 15%.
[e] ESS scores compared between patients indicating that they had not experienced any problems staying awake during daily activities during the past month (score 0 on PSQI item 8) and those indicating that they had. Among 113 responders to PSQI item 8, total ESS scores were available for 111.
[f] $P<0.0001$ (Mann-Whitney U-test).

ESS, Epworth Sleepiness Scale; SD, standard deviation; IQR, inter-quartile range; CI, confidence interval; PSQI, Pittsburgh Sleep Quality Index.

Table 3 Rasch item statistics of the ESS (n=118)

Item no. [a]	Item content	Logit (SE) [b]	INFIT [c]
6	Sitting and talking to someone	2.27 (0.19)	0.83
8	In a car, while stopped for a few minutes in the traffic	2.19 (0.19)	1.19
3	Sitting inactively in a public place	0.24 (0.14)	0.85
7	Sitting quietly after lunch without alcohol	-0.01 (0.14)	0.88
4	As a passenger in a car for 1 h without a break	-0.05 (0.14)	1.00
1	Sitting and reading	-0.64 (0.14)	0.72
2	Watching TV	-1.47 (0.14)	1.29
5	Lying down to rest in the afternoon when circumstances permit	-2.54 (0.16)	1.29

[a] Listed in order of location along the latent continuum.

[b] Linear Rasch (rating scale model) calibrated item logit measures. Each item logit measure represents that item's location (propensity to induce sleep) on the latent continuum. Mean item logit measure is 0.

[c] Mean-square (MNSQ) information-weighted goodness-of-fit (INFIT) to the Rasch model. Values ≤1.4 support unidimensionality.

ESS, Epworth Sleepiness Scale; logit, log-odds units; SE, standard error; INFIT, information-weighted goodness-of-fit statistics.

Legends

Fig. 1.

Plot of ESS item loadings on the two first factors extracted by exploratory principle component factor analyses with orthogonal (varimax) rotation. Kaiser-Meyer-Olkin measure of sampling adequacy: 0.84; Bartlett's test of sphericity: χ^2, 322.94, P<0.0001. Numbered dots represent items (for item contents, see Table 3).

Fig. 2.

Rating scale functioning of the ESS. Curves show the probability of observing each category (y-axis) relative to the location on the latent sleep propensity trait (x-axis). Rating scale category thresholds (i.e., the intersection points where there are 50/50 probabilities of endorsing adjacent categories; indicated by arrows in the graph) should be ordered in an expected manner, and response categories should emerge as an even succession of "hills" (i.e., being more and more probable) from less to more along the underlying continuum (x-axis). Rating scale categories denote "would never doze" (0), "slight chance of dozing" (1), "moderate chance of dozing" (2), and "high chance of dozing" (3). ESS, Epworth Sleepiness Scale.

Fig. 3.

Illustration of the locations of ESS rating scale category thresholds and items, as well as the sample along the underlying sleep propensity continuum in Parkinson's disease (n=118). All locations were estimated according to the Rasch rating scale model along the same linear logit (log-odds units) metric (y-axis), with the mean item location set at 0 logits (Wright & Masters, 1982). Each item is indicated by its number and abbreviated content (for full item

contents and exact item locations, see Table 3). Items located at the lower end are easily endorsed (representing relatively low degrees of sleepiness), whereas items at the higher end represent high degrees of sleepiness. Similarly, people at the lower end of the sample distribution report less sleepiness than those at the higher end. Rating scale category threshold plots are locations of the points where there are 50/50 probabilities of endorsing adjacent item response categories and represent the range and coverage of measurement of the ESS. Areas not represented by rating scale category thresholds are measured with less precision. Arrows indicate gaps in measurement coverage exceeding 0.5 logits (dashed arrows) and 1.0 logits (solid arrows). Horizontal lines represent percentiles of the sample distribution according to the logit metric (y-axis).

Fig. 1

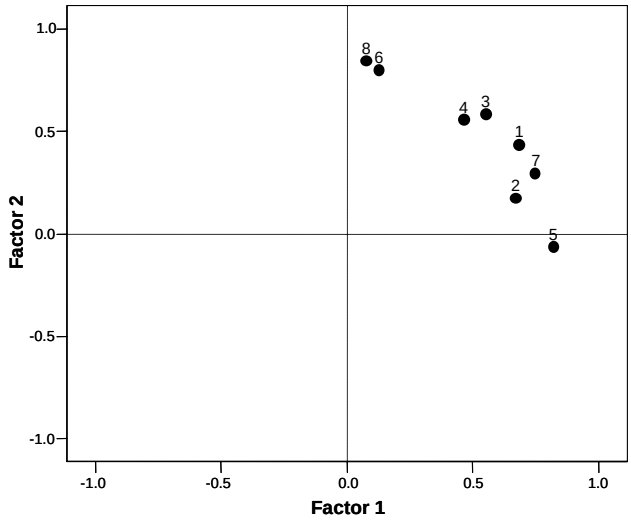

Fig. 2

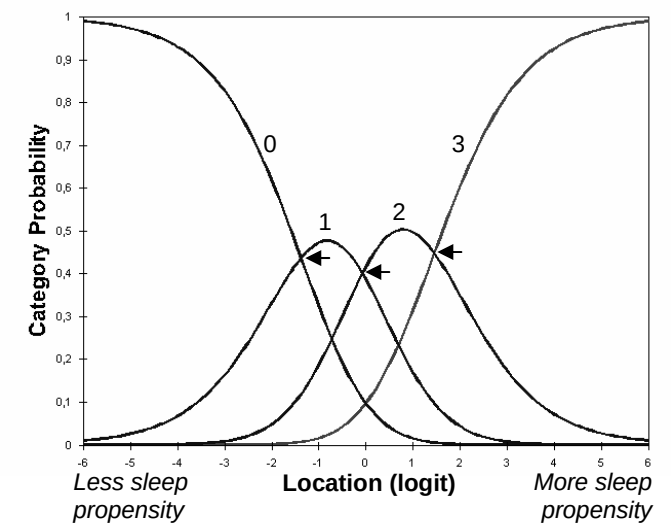
Category Probability
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
1
2
3
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Less sleep propensity
Location (logit)
More sleep propensity

Fig. 3


More sleep propensity
Less sleep propensity
Location (logit)
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
Rating Scale Category Thresholds
Item Locations
Sample Distribution
(6) Talking
(8) In stopped car
(3) Inactive in public
(4) Car ride
(7) Sitting after lunch
(1) Reading
(2) Watching TV
(5) Lying down
90th percentile
75th percentile
50th percentile
25th percentile
10th percentile
0
10
20
30
40
50
n