Sleep Medicine Reviews 18 (2014) 321–331

Contents lists available at ScienceDirect

Sleep Medicine Reviews

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

CLINICAL REVIEW

Evaluation of the measurement properties of the Epworth sleepiness scale: A systematic review

Tetyana B. Kendzerska [a,*], Peter M. Smith [b,c,d], Romina Brignardello-Petersen [a,e], Richard S. Leung [f,g], George A. Tomlinson [a,d,h]

[a] Institute of Health Policy, Management and Evaluation, Faculty of Medicine, University of Toronto, 155 College Street, Suite 425, Toronto, ON M5T 3M6, Canada
[b] School of Public Health and Preventive Medicine, Monash University, Victoria, Australia
[c] Institute for Work & Health, Toronto, ON, Canada
[d] Dalla Lana School of Public Health, University of Toronto, ON, Canada
[e] Evidence-Based Dentistry Unit, Faculty of Dentistry, University of Chile, Santiago, Chile
[f] Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[g] Sleep Laboratory, St. Michael's Hospital, Toronto, ON, Canada
[h] Toronto General Research Institute, Toronto General Hospital, Toronto, ON, Canada

ARTICLE INFO

Article history:
Received 7 February 2013
Received in revised form
27 July 2013
Accepted 13 August 2013
Available online 22 August 2013

Keywords:
Daytime sleepiness
The Epworth sleepiness scale
Psychometric properties (or reliability and validity)

SUMMARY

Objective: To examine published evidence on the psychometric properties of the Epworth sleepiness scale (ESS) for describing the level of daytime sleepiness (DS) in adults.
Methods: Articles were located on MEDLINE and EMBASE. Psychometric properties were appraised using the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist.
Results: We found thirty-five studies evaluating psychometric properties of the ESS in adults. Of these, 27 studies examined construct validity, 14 – known-group validity, 8 – internal consistency and 4 – test–retest reliability. Study quality ranged from excellent to poor the majority being fair. Internal consistency by Cronbach's alphas was good (0.73–0.86). There is little available evidence on test–retest reliability. Pooled correlations of the ESS with other constructs varied: from moderate (the maintenance of wakefulness test; $\rho = -0.43$), to weak (the multiple sleep latency test; $\rho = -0.27$, and sleep apnea-related variables; ρ from 0.11 to 0.23). Although ESS scores varied significantly across groups of subjects with known differences in DS, not all differences were clinically important.
Conclusion: There have been relatively few high quality studies on the ESS psychometric properties. The internal consistency of the ESS suggests that this instrument can be recommended for group but not individual-level comparisons. Correlations with other measures of DS were stronger than with sleep apnea-related or general health measures, but still lower than expected. Further studies are required in the areas of test–retest reliability of the ESS.

Introduction

Daytime sleepiness (DS) is a complicated clinical problem, often indicating a serious underlying physiological abnormality [1]. DS is associated with higher mortality [2,3], an increased risk for motor-vehicle crashes [4] and work-related accidents, and a higher prevalence of co-morbid conditions such as diabetes, myocardial infarction, and stroke [5–7].

Accurately estimating an individual's level of DS is important, both to better understand the factors associated with the level of DS and to estimate the health and social consequences of DS. However, the wide spectrum of definitions currently associated with sleepiness complicates its quantification [8]. The *International classification of sleep disorders* – second edition (2005) [9] defines DS as a difficulty in maintaining the alert awake state during the wake phase of the 24-h sleep–wake cycle. DS has been operationalized as drowsiness, as propensity to sleep [10] or by assessing the impact that sleepiness has on various aspects of daily life [8]. The most often used operational definition of sleepiness is the speed, ease or likelihood of falling asleep as opposed to remaining awake and is

* Corresponding author. Tel.: +1 416 669 6759.
E-mail addresses: tetyana.kendzerska@mail.utoronto.ca, kendzerskaya@gmail.com (T.B. Kendzerska).

1087-0792/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.smrv.2013.08.002

Abbreviations

AHI	apnea–hypopnea index
ArI	arousal index
ASP	average sleep propensity
COSMIN	COnsensus-based Standards for the selection of health status Measurement Instruments
CTT	classical test theory
DS	daytime sleepiness
ESS	Epworth sleepiness scale
ICC	intraclass correlation coefficient
IRT	item response theory
ISP	instantaneous sleep propensity
MSLT	multiple sleep latency test
MWT	maintenance of wakefulness test
OSA	obstructive sleep apnea
PSQI	Pittsburgh sleep quality index
SaO_2	oxygen saturation
SSP	situational sleep propensity

Glossary of terms

Instantaneous sleep propensity (ISP) describes a person's sleep propensity over the preceding few minutes in one particular situation, at one particular time (e.g., while reading sitting in a chair in a quiet room today at 10 am).

Situational sleep propensity (SSP) is the person's usual sleepiness in the particular situation. It is formed by combination of multiple ISP values for one situation (e.g., while reading sitting in a chair in a quiet room).

Average sleep propensity (ASP) is a person's general level of sleepiness across a variety of situations commonly encountered in daily life (e.g., reading, watching TV, talking, driving etc). Multiple SSP results for varying situations form the ASP.

Internal consistency is the extent to which the items in an instrument (e.g., the Epworth sleepiness scale (ESS)) all measure the same concept or construct (e.g., daytime sleepiness (DS)).

Cronbach's alpha is a measure of internal consistency. Its value ranges from 0 to 1. Higher scores indicate greater internal consistency.

Unidimensionality an instrument is considered as unidimensional if its items measure a single latent trait or construct (e.g., DS).

Test–retest reliability assesses the amount of absolute change in a measure (e.g., the ESS) when two independent assessments are completed over a period of time when no true change has occurred. The preferred method to assess the test–retest reliability of an instrument is the intraclass correlation.

Intraclass correlation coefficient measures an agreement between two or more raters or evaluation methods on the same set of subjects. Its value ranges from 0 to 1 (1 represents perfect agreement, 0 – no agreement at all).

Construct validity is the process of examining whether the scores on a measure are consistent with the hypothesized relationship the measure should have with other constructs. This can be done by assessing the correlation between one measure and other measures (e.g., correlations between the ESS and maintenance of wakefulness test (MWT) or multiple sleep latency test (MSLT)), or assessing if differences are present in the scores between groups that are known to be different (e.g., difference in ESS scores across groups with different known level of sleepiness).

represented by instantaneous sleep propensity (ISP), situational sleep propensity (SSP) and average sleep propensity (ASP) [11,12]. The ISP describes a person's sleep propensity over the preceding few minutes in one particular situation, at one particular time. Combination of multiple ISP values for one situation forms the SSP, the person's usual sleepiness in that particular situation. In the same way, multiple SSP results for varying situations form the ASP, a person's general level of sleepiness across a variety of situations commonly encountered in daily life. ASP is only related to propensity to sleep and measures only that component of DS which persists from week to week in a given subject. As such, ASP differs from feeling tired, sleepy or drowsy in particular situations and does not measure the impact of sleepiness on aspects of daily life [13].

Different methods have been proposed for measuring sleepiness and they can be classified according to their operational definitions [12,14]. The Epworth sleepiness scale (ESS) is the measure of sleepiness most commonly used in sleep research and clinical settings. A search conducted on July 22nd, 2013 of PubMed articles containing "Epworth sleepiness scale" as a search term returned in total 1868 articles: 163 – in 2010, increasing to 208 articles in 2011 and 238 articles in 2012. By contrast, a search for "Stanford sleepiness scale" returned only 259 articles in total, 15 in 2012. The ESS is the only English language tool available to measure a person's ASP in daily life. This contrasts it with the multiple sleep latency test (MSLT) [15] and the maintenance of wakefulness test (MWT) [16] that measure a person's SSP. Unlike the Karolinska [17] and the Stanford [18] sleepiness scales, the ESS does not measure subjective feelings of drowsiness.

The ESS was developed in 1991 using data from healthy subjects and patients with a variety of sleep disorders to describe "the general level of [DS], as distinct from feelings of sleepiness at a particular time" [19]. The ESS asks people to rate, on a four-point scale, their usual chances of falling asleep in eight different situations, chosen to represent the different levels of "somnificity" that most people encounter as part of their daily lives [19]. Somnificity was defined as "the general characteristic of a posture, activity and situation that reflects its capacity to facilitate sleep-onset in the majority of subjects" [20]. The ESS is inexpensive and easy to administer, complete and score. ESS item-scores are recorded as a number from 0 to 3 written in a single box for each item [19]. The total ESS score is the sum of item-scores and ranges between 0 and 24; the higher the score, the higher the person's level of DS. From the sleep propensity viewpoint, each of the eight ESS item-scores represents a different subjectively-reported SSP [21]. The total score gives a subjectively-reported ASP across the eight ESS situations [20].

Given its widespread use in the field of sleep research it is surprising that there has not been a comprehensive review of the measurement properties of the ESS. While there have been individual papers examining the various aspects of the psychometric properties of the ESS, these studies have not been examined together to evaluate the measurement properties of the ESS. The purpose of this paper is to fill this evidence gap by reviewing the available research examining the measurement properties of the ESS for describing DS in adults. In doing so we hope to provide valuable knowledge for future research projects in this area, in deciding to use the ESS to describe DS.

Methods

Search strategy

A broad search of the literature was performed incorporating both electronic and manual components. Two electronic databases, MEDLINE and EMBASE, were searched. Table 1 displays the terms for psychometric properties used in the search. Finally, we carried out manual searches of the references of all articles deemed relevant.

Selection criteria

Searches were limited to studies in adult populations and English language articles published between 1991 (when the scale was first reported) and June 2012. We included only full text original articles (i.e., not abstracts or reviews), focused on the development or evaluation of the measurement properties of an English version of the ESS. Since the ESS was not developed for a certain population [19], no restrictions were applied to the type of population studied. We excluded studies not primarily designed to assess the psychometric properties of the ESS (e.g., where the ESS was used as one independent variable in a multivariable regression model [22], or where models examined predictors of the ESS).

Measurement properties

Psychometric properties were appraised using the "COnsensus-based Standards for the selection of health status Measurement Instruments" (COSMIN) checklist, an instrument developed to evaluate the methodological quality of studies on measurement properties of health-status-related questionnaires [23] (COSMIN manual July 2011, see http://www.cosmin.nl). The COSMIN checklist was developed to provide a uniform use of terminology and definitions of measurement properties for health-related patient-reported outcomes and is based on an international Delphi study [23,24]. The COSMIN checklist consists of nine boxes (A–I), each containing 5–18 items covering methodological standards against which each measurement property should be assessed. Each item is scored on a four-point rating scale (i.e., "poor", "fair", "good", or "excellent"). An overall score for the methodological quality of a study is determined for each measurement property separately (e.g., if a study focused on both reliability and validity, then it would be given a separate rating for each), by taking the lowest rating of any of the items in a box. As the ESS was designed for a descriptive purpose, under the framework of Kirsher and Guyatt [25], we assessed the measurement properties of the ESS across the domains of reliability and construct validity, described in further detail below. Criterion validity was not assessed because of the absence of a gold standard of DS measurement.

Reliability

The reliability domain consists of internal consistency, reliability and measurement error. Internal consistency is defined as a degree of interrelatedness among the items [23,24]. A prerequisite for internal consistency is evidence of the unidimensionality of a scale [26]. Reliability and measurement error are focused on evaluating the proportion of the total variance in the measurements which is due to 'true' differences between patients, and is commonly assessed using test–retest reliability. Important attributes for the assessment of test–retest reliability under the COSMIN checklist are that the two tests are administrated independently, the underlying concept to be measured is consistent between the measurements, and that the time interval between the administrations is appropriate [27]. The intraclass correlation coefficient (ICC) is the preferred measure of reliability for continuous scores [28].

Validity

We focused on the construct validity domain of the COSMIN checklist, defined as hypothesis testing. Here, hypothesis testing assesses the degree to which the scores of an instrument are consistent with the assumption that the instrument validly measures the construct to be measured. Hypotheses can cover convergent or divergent validity (checking relationships to scores on other instruments) and known-group validity (checking differences between relevant groups). Important attributes of studies assessing validity within the COSMIN checklist are a priori hypotheses about the expected direction and magnitude of the instruments to be compared and an adequate description (and measurement properties) of the comparator instrument [24].

Quality assessment

The methodological quality of the included articles was assessed independently by two reviewers using the COSMIN checklist (TK and PS, or TK and GT, or TK and RB). In case of disagreement between the two reviewers, third reviewer (PS or GT) was consulted to reach consensus.

Table 1
The terms used for the psychometric properties search in MEDLINE and EMBASE.

	Search terms
MEDLINE	reliability.ti,ab./ or Cronbach's alpha.ti,ab./ Cronbach's.ti,ab./ or test–retest.ti,ab./ or variance.ti,ab./ repeatability.ti,ab./ or stability.ti,ab./ or congruence.ti,ab./ or validity.ti,ab./ or "reproducibility of results"/ or Feasibility Studies/
EMBASE	exp EXTERNAL VALIDITY/ or exp VALIDITY/ or exp CONTENT VALIDITY/ or exp FACE VALIDITY/ or exp PREDICTIVE VALIDITY/ or validity.mp. or exp CONSTRUCT VALIDITY/ or exp INTERNAL VALIDITY/ or exp CONSENSUAL VALIDITY/ or exp CRITERION RELATED VALIDITY/ or exp CONCURRENT VALIDITY/ or exp QUALITATIVE VALIDITY/ or exp DISCRIMINANT VALIDITY/ or exp reproducibility/ or exp measurement/ or acceptability.mp.

Guide to search syntax:
exp – Explodes the subject heading to retrieve the search term plus all narrower (more specific) terms (OVID).
/ – All subheadings for a subject heading are included in the search (OVID).
.ti,ab. – Searches in record title and abstract (OVID).
.mp. – Searches in the title, abstract, and subject heading fields (OVID).

Best evidence synthesis: levels of evidence

The assessment of the measurement properties of the ESS was done in two steps. First, the methodological quality of the studies was assessed using the COSMIN checklist. Studies were rated as excellent, good, fair or poor. Second, the consistency of results (levels of evidence) was summarized, taking the methodological quality of different studies into account, based on criteria used previously by the Cochrane Back Review Group [29,30]. Levels of evidence were rated as either strong (consistent findings were observed in multiple studies of "good" methodological quality OR there was one study of "excellent" methodological quality); moderate (consistent findings were observed in multiple studies of "fair" methodological quality OR there was one study of "good" methodological quality); limited (one study of "fair" methodological quality) or unknown (only studies of "poor" methodological quality were found) [30].

To obtain summary estimates for convergent construct validity, we used the generic inverse variance method to pool the Fisher *Z*-transformed correlations [31]. Each pooled correlation and its 95% confidence interval (CI) were back-transformed from the Fisher *Z*-scale. Heterogeneity of results across studies was summarized by the I^2 statistic where higher score means higher heterogeneity; percentages of around 25%, 50%, and 75% indicate low, medium, and high heterogeneity, respectively [32,33].

Results

Literature search

From 462 papers found through the online search, 46 describing or commenting on the psychometric properties of the ESS were selected by scanning their titles and abstracts (Fig. 1). By applying our selection criteria through full text review, 35 primary articles were selected to form the basis of this review: eight studies evaluated internal consistency, four evaluated test–retest reliability, 27 evaluated convergent construct validity, and 14 evaluated known-group validity. The general characteristics of these studies are presented in Table S1.

Quality assessment

Among studies evaluating internal consistency, two were rated as "excellent", two as "good", one as "fair" and three as "poor" (Table 2). The most common methodological flaw was a failure to check the unidimensionality of the ESS scale, with no reference made to another study that checked this property. Among studies evaluating test–retest reliability, one study was rated as "good", one as "fair" and two as "poor". The most common flaws in these studies were uncertainty over whether the test and retest populations were comparable due to the long time interval between assessments or because the test conditions had changed. Even though two studies reported an ICC [34,35], the model and formula for the ICC used were not stated.

Among studies evaluating convergent construct validity six studies were rated as "good", 19 studies – as "fair" and two studies – as "poor". Among studies receiving a "fair" score, eight studies had "fair" on only one evaluation aspect, with the other parts of the study rated as "good" or "excellent". The common flaws in these studies were vague or absent a priori hypotheses and lack of information provided on measurement properties of the comparator instruments.

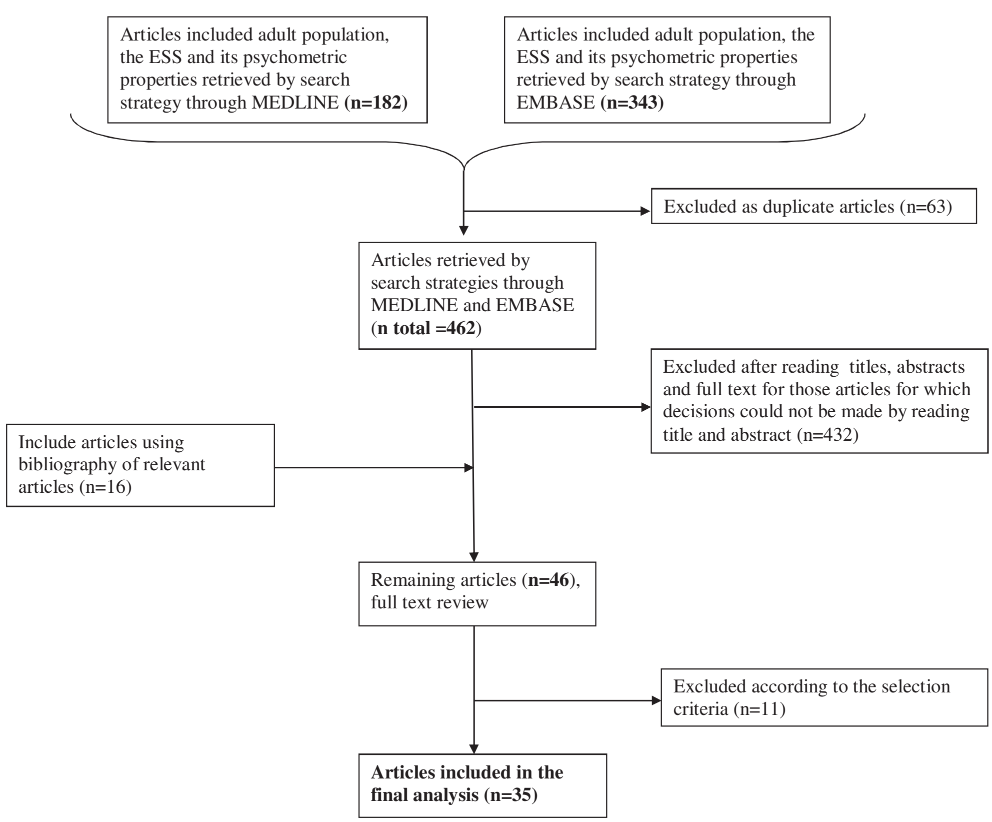

Fig. 1. Flow diagram for identification of relevant studies (Ovid MEDLINE(R) 1990 to June Week 4 2012; EMBASE 1990–2012 Week 27).

Table 2
Methodological quality of included studies assessed by the COSMIN checklist.

#	Study, reference number	Reliability		Hypotheses testing	
		Internal consistency	Test–retest	Convergent validity	Known-group validity
1.	Banks et al., 2004 [47]	–	–	Good	Fair[c]
2.	Beaudreau et al., 2012 [38]	Poor[b]	–	Fair[c]	Good
3.	Benbadis et al., 1999 [79]	–	–	Fair[c]	–
4.	Bennett et al., 1998 [80]	–	–	Fair	–
5.	Bennett et al., 1999 [56]	–	–	Fair[c]	–
6.	Briones et al., 1996 [55]	–	–	Fair[c]	–
7.	Bourke et al., 2004 [34]	Poor	Poor	Poor	–
8.	Buysse et al., 2008 [57]	–	–	Good	–
9.	Chervin et al., 1997 [51]	–	–	Fair[a,c]	Good
10.	Chervin and Aldrich, 1999 [50]	–	–	Fair[c]	Fair[c]
11.	Gottlieb et al., 1999 [81]	–	–	–	Fair[c]
12.	Gottlieb et al., 2000 [82]	–	–	–	Fair[c]
13.	Hesselbacher et al., 2012 [59]	–	–	Fair[c]	Good
14.	Johns, 1991 [19]	–	–	Poor[a]/Fair	Fair
15.	Johns, 1992 [13]	Excellent	Good	–	–
16.	Johns, 1993 [83]	–	–	Fair	Fair
17.	Johns, 1994 [21]	Good/Fair[d]	–	Fair	
18.	Kezirian et al., 2007 [60]	–	–	–	Good
19.	Kezirian et al., 2009 [61]	–	–	–	Good
20.	Kingshott et al., 1995 [54]	–	–	Fair	–
21.	Kingshott et al., 1998 [40]	Fair	–	–	–
22.	Kingshott, Engleman et al., 1998 [46]	–	–	Good	–
23.	Knutson et al., 2006 [35]	–	Poor	–	–
24.	Martin et al., 1997 [84]	–	–	Fair	–
25.	Nguyen et al., 2006 [36]	Good	Fair	–	–
26.	Olson et al., 1998 [49]	–	–	Good	–
27.	Osman et al., 1999 [85]	–	–	Fair	–
28.	Sangal et al., 1999 [48]	–	–	Fair	–
29.	Skibitsky et al., 2012 [58]	–	–	Fair	–
30.	Smith et al., 2008 [37]	Excellent	–	Fair	–
31.	Spira et al., 2012 [39]	Poor[b]	–	Fair[c]	Good
32.	Stavitsky et al., 2010 [86]	–	–	Fair	Fair
33.	Sunwoo et al., 2012 [72]	–	–	Fair	–
34.	Weaver et al., 2004 [52]	–	–	Good	Fair[c]
35.	Weaver et al., 2005 [53]	–	–	Good	–

[a] For the multiple sleep latency test (MSLT).
[b] Received "poor" for one parameter only (such as "Was the unidimensionality of the scale checked?"), others parameters were "good" or "excellent".
[c] Received "fair" for one parameter only (such as i) hypotheses vague or not formulated but possible to deduce what was expected; or ii) only some information on measurement properties of the comparator instrument(s) in any study population provided), others were "good" or "excellent".
[d] Received "fair" for the subsample of 44 patients.

Levels of evidence

Reliability: internal consistency

Internal consistency of the ESS was supported by strong evidence (two studies of excellent and two of good quality) [13,21,36,37]. In these studies Cronbach's alphas ranged from 0.73 in sample of university students [13] to 0.88 in a sample of consecutive patients with various sleep disorders [13,21]. The Cronbach's alpha in the studies that were omitted because they were rated "poor" for one parameter only ("Was the unidimensionality of the scale checked?"), otherwise parameters were "good" or "excellent", was 0.76 [38] and 0.70 [39], and are consistent with the values in the studies reported. Applying the item response theory (IRT) approach in an obstructive sleep apnea (OSA) population, a value of 0.79 for Rho (a reliability coefficient in Mokken scale analyses) has been reported, marking this scale as reliable [40]. However, the quality of this study was rated as fair [40] and internal consistency was calculated on only the four items found to adhere to the IRT model (#1 ("reading"), #2 ("watching TV"), #6 ("talking"), #7 ("after a lunch")).

Unidimensionality is required for the internal consistency statistic to be interpretable [26]. There was strong evidence that not all items of the ESS belong to one underlying dimension. Two studies (one rated as good and one as excellent) [13,21] found one underlying dimension of DS. However, the normalized factor loadings for items #6 ("talking"), #8 ("in traffic") and #5 ("lying down") were low (0.25 [13] and 0.26 [21] for #6, 0.37 [13] for #8 and 0.08 [21] for #5) in a sample of students. Only values with an absolute value greater than 0.4 can be seen to reflect loading on a particular dimension [41]. Another good quality study [36] reported three underlying factors for the ESS. One study of excellent quality examined unidimensionality using confirmatory factor analysis and found that a one factor model with all items in the ESS produced sub-optimal model fit as assessed by multiple fit indices [37]. While the authors did not test for the presence of a second factor for the ESS model in their analyses, they did report that model fit improved after the removal of items #6 and #8.

Reliability: test–retest

Moderate evidence was found in support of test–retest reliability (Table S2). Only one of the four studies was given a quality rating of "good". This study examined the test–retest reliability of the ESS in 87 medical students [13] and reported a Pearson correlation coefficient of 0.82 between two administrations of the ESS over a time period of five months.

Construct validity

Fig. 2 summarizes the evidence for convergent construct validity, and Table 3 the evidence for known-group validity. We have used published guidelines [42] that interpret absolute values of correlations above 0.70 as strong, between 0.3 and 0.7 as moderate, and those less than 0.3 as weak. These values correspond to shared variances of approximately 50%, 10%–50% and less than 10% respectively. Some specific details of results on validity are presented below.

Given the lack of a priori hypotheses across studies, we developed our own a priori hypotheses with which to evaluate the levels of evidence across the various studies examining construct validity. These studies included comparisons with other measures of DS, measures of obstructive sleep apnea (OSA), and general health status measures. The ESS, MWT and MSLT measure DS from the viewpoint of sleep propensity; however, the ESS measures sleep propensity among a range of situations while MWT and MSLT only measure situational sleep propensity (SSP). A subject's sleep propensity in any one situation is not always closely related to that in a different situation [12,21], and both of these laboratory-based tests can only measure SSP in highly artificial situations that may not reflect real-life experience [43]. As such, we would only expect moderate correlations (0.3–0.7) between the ESS and each of these two measures. The MSLT measures how quickly one falls asleep when asked to do so (lying down in a quiet and dark room) [44], while the MWT measures a person' ability to stay awake (while seated in a quiet and dark room). Because the ESS is based on the assumption that subjects, when exposed to situations of varying soporificity (e.g., "watching TV" or "in traffic") are usually trying to stay awake, not fall asleep, we would expect the ESS to be more highly correlated with the MWT than the MSLT. A moderate correlation can be also expected between the ESS and the severity of OSA, as OSA is known to cause excessive DS [45]. Measures of OSA included the apnea–hypopnea index (AHI), oxygen saturation (SaO_2), and the arousal index (ArI). However, we would expect this correlation to be weaker than for the MSLT and MWT, as these are direct measures of DS. A weak correlation can be expected between the ESS and quality of life related measures such as SF-36 and Pittsburgh sleep quality index (PSQI).

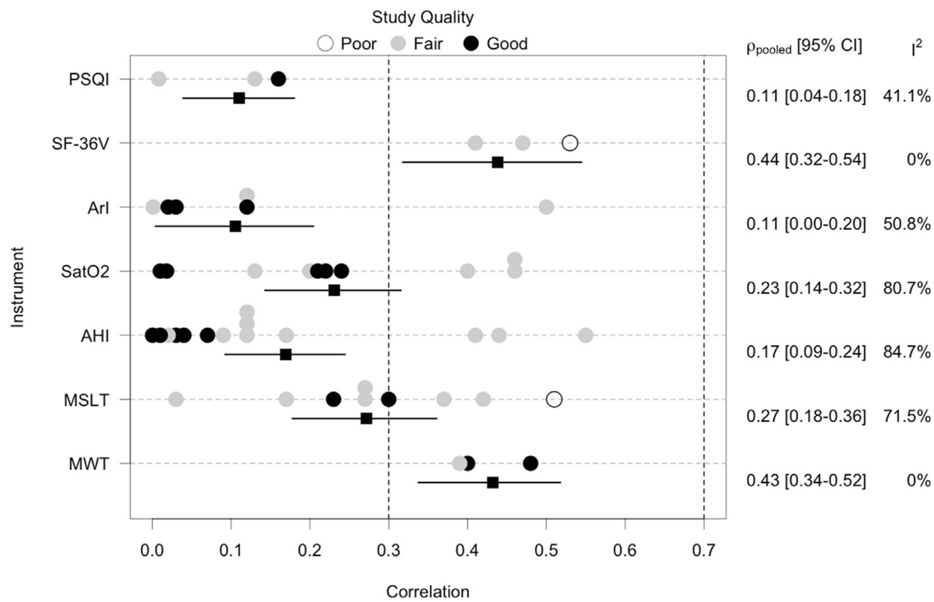

Fig. 2. Convergent and divergent construct validity of the Epworth sleepiness scale (ESS) (all correlation coefficients are presented as absolute values). AHI – the apnea–hypopnea index; ArI – the arousals index; MSLT – the multiple sleep latency test; MWT – the maintenance of wakefulness test; SaO_2 – oxygen saturation; SF-36V – SF-36, vitality subscale; PSQI – Pittsburgh sleep quality index.

The convergent construct validity

A moderate level of evidence was found to support a moderate association between the ESS and the MWT. Spearman rank correlation coefficients of −0.48 [46] and −0.40 [47] were reported in two good quality studies. In addition, one study of fair quality reported a Spearman rank correlation coefficient of −0.39 [48]. The pooled correlation was −0.43 (95%CI: −0.52 to −0.34) with no heterogeneity (I^2 of 0%).

A moderate level of evidence was also found for a weak association between the ESS and the MSLT. Two good quality studies reported Spearman rank correlation coefficients of −0.30 [49] and −0.23 [46]. The correlation coefficients in the studies of fair quality ranged from −0.03 [50] to −0.42 [21]. The pooled correlation was −0.27 (95%CI: −0.36 to −0.18) with I^2 of 71.5%.

There was strong evidence of no or weak association between the ESS and severity of OSA as expressed by AHI. Five good [46,49,51–53] quality studies reported correlation coefficients ranging from 0.0002 [49] to 0.07 [46]. The correlation coefficients in the studies of fair quality ranged from 0.02 [54] to 0.55 [19]. The pooled correlation was 0.17 (95%CI: 0.09–0.24) with I^2 of 84.7%.

The evidence for an association with the SF-36 vitality subscale was limited, coming from two studies of fair quality reporting moderate correlations of −0.41 [55] and −0.47 [56] and one additional study of poor quality [34]. Moderate evidence was observed for a weak association with the PSQI. Pearson correlation coefficients were 0.16 [57] in a good quality study and 0.008 [58] and 0.13 [39] in two fair quality studies.

Known-group validity

The evidence for known-group validity was strong, comprising six studies of good quality [38,39,51,59–61] and eight studies of fair quality. These studies examined differences in the ESS scores between normal subjects and patients suffering from a variety of sleep disorders known to be associated with excessive DS (Table 3). In five studies of good quality the difference in the ESS scores between normal controls and subjects with sleep disorders ranged from 0.5 (AHI ≥ 18.8 vs AHI < 7.9) [61] to 11.6 (narcolepsy with cataplexy vs subjects without hypersomnolence) [51].

Discussion

The comprehensive literature search identified 35 studies that evaluated psychometric properties of the ESS in an adult population. The bulk of these studies examined construct validity; eight evaluated internal consistency and only four examined test–retest reliability. The study quality ranged from excellent to poor, with the majority being fair. We discuss the results below under the domains of reliability and construct validity.

Reliability

Internal consistency

Cronbach's alphas ranged from 0.7 to 0.9, values indicating adequate internal consistency for within- or between-group comparisons; however, Cronbach's alpha of 0.9 and higher is recommended if a scale is used for comparison of individual scores [36,62,63]. We noted that Cronbach's alphas were generally lower among non-clinical samples (students, community samples) and white women and higher for clinical samples (patients with sleep disorders known associated with higher level of DS) and black women [13,21,38].

Studies examining unidimensionality of the ESS raised concerns over whether item #6 ("talking") and #8 ("in traffic") belong to one dimension. In studies using factor analysis, two [13,21] reported low factor loadings for these items while a third [37] reported that good fit statistics could only be obtained after the removal of these two items. These results may indicate possibility that these items are less-than-optimal measures of ASP in general.

We propose three possible explanations for the results on internal consistency and unidimensionality: i) not all items are good measures of the construct of DS; ii) if different items represent different levels of somnificity, classical test theory (CTT) does not take this into consideration; or iii) different subgroups of

Table 3
Evidence of known-group validity of the Epworth sleepiness scale (ESS).

#	Study	Group comparison	Mean ESS (SD) or median (IQR)	p Value
1.	Banks et al., 2004 [47]	Mild to moderate OSA ($n = 110$)	10.7 (4.5)	<0.0001[a]
		Control subjects ($n = 41$)	5.44 (3.0)	
2.	Beaudreau et al., 2012 [38]	No reported diagnosis ($n = 2688$)	5 (3-8)	
	Each sleep disorder	Insomnia ($n = 108$)	4 (2-7)	$p = 0.01$
	compared to no diagnosis	Restless legs ($n = 129$)	5 (3-9)	$p = 0.24$
		Sleep apnea ($n = 26$)	6.5 (3-10)	$p = 0.22$
3.	Chervin et al., 1997 [51]	OSAS ($n = 23$)	13.0 (6.8)	<0.05
		UARS ($n = 16$)	15.6 (4.2)	
		Narcolepsy with cataplexy ($n = 9$)	18.6 (2.6)	
		Without hypersomnolence ($n = 15$)	7 (NS)	
4.	Chervin et al., 1999 [50]	How often do you have a major problem with sleepiness during the daytime?	NS (Fig. 2 in [50])	<0.05
		Responses of never (PS = 1), seldom (PS = 2), occasionally (PS = 3), often (PS = 4), almost always (PS = 5)		
5.	Gottlieb et al., 1999 [81]	RDI < 5 ($n = 998$)	7.2 (4.3)	<0.001
		5 ≤ RDI < 15 ($n = 524$)	7.8 (4.4)	
		15 ≤ RDI < 30 ($n = 211$)	8.3 (4.6)	
		RDI ≥ 30 ($n = 191$)	9.3 (4.9)	
6.	Gottlieb et al., 2000 [82]	RDI < 1.5 ($n = 1585$)	7.1 (4.2)	<0.001
		1.5–<5 ($n = 1498$)	7.5 (NS)	
		5–<15 ($n = 1672$)	7.8 (NS)	
		≥15 ($n = 1022$)	8.8 (4.8)	
		≥30 ($n = 354$)	9.7 (4.9)	
7.	Hesselbacher et al., 2012 [59]	OSA (RDI ≥15 events/h) (n = NS)	11.3 (5.7)	<0.0001
		Non OSA patients (n = NS)	9.8 (5.4)	
8.	Johns, 1991 [19]	Normal controls ($n = 30$)	5.9 (2.2)	<0.0001
	Sleep disorders vs normal subjects	Primary snorers ($n = 32$)	6.5 (3.0)	
		OSA ($n = 55$)	11.7 (4.6)	
		Narcolepsy ($n = 13$)	17.5 (3.5)	
		Idiopathic hypersomnia ($n = 14$)	17.9 (3.1)	
		Insomnia ($n = 18$)	2.2 (2.0)	
		PLMD ($n = 18$)	9.2 (4.0)	
	OSA severity	Primary snorers ($n = 32$)	6.5 (3.0)	<0.05
		Mild (>5–15) ($n = 22$)	9.5 (3.3)	
		Moderate OSA (>15–30) ($n = 20$)	11.5 (4.2)	
		Severe OSA (>30) ($n = 13$)	16.0 (4.4)	
9.	Johns, 1993 [83]	Primary snorers ($n = 108$)	8.0 (3.5)	<0.001
		Mild OSA (RDI: 5–24.9) ($n = 105$)	11.0 (4.2)	
		Moderate OSA (25–49.9) ($n = 41$)	13.0 (4.7)	
		Severe OSA (≥50) ($n = 19$)	16.2 (3.3)	
10.	Kezirian et al., 2007 [60]	AHI < 7.06 ($n = 153$)	5.2 (3.3)	0.0026
		AHI: 7.06–17 ($n = 154$)	5.6 (3.6)	
		AHI > 14 ($n = 154$)	6.6 (3.9)	
11.	Kezirian et al., 2009 [61]	AHI < 7.90 ($n = 948$)	5.9 (3.6)	0.005
		AHI: 7.90–<18.89 ($n = 949$)	6.0 (3.6)	
		AHI ≥ 18.89 ($n = 952$)	6.4 (3.8)	
12.	Spira et al., 2012 [39]	No diagnosis ($n = 2652$)	6.0 (3.6)	
	Each sleep disorder	Insomnia ($n = 47$)	5.7 (3.7)	0.54
	compared to no diagnosis	Restless legs ($n = 65$)	7.5 (4.7)	0.015
		Periodic leg movements ($n = 54$)	7.1 (4.3)	0.063
		Sleep apnea ($n = 189$)	7.3 (4.1)	<0.001
		Narcolepsy ($n = 11$)	11.5 (6.1)	0.015
13.	Stavitsky et al., 2010 [86]	Healthy subjects ($n = 14$)	6.3 (3.5)	<0.002
		Parkinson disease ($n = 30$)	11.5 (4.6)	
14.	Weaver et al., 2004 [52]	Mild (5–15)	12 (5)	Non-significant
		Moderate (15.1–30)	13 (6)	(p-value not
		Severe (>30)	13 (6)	specified)

Studies highlighted in gray are those of good quality.
AHI – apnea–hypopnea index; IQR - interquartile range; NS - not stated; OSA – obstructive sleep apnea; OSAS – obstructive sleep apnea syndrome; PLMD – periodic limb movement disorder; PS – the subjective problem of sleepiness as defined by Chervin et al. [50]; RDI – respiratory disturbance index; SD - standard deviation; UARS – upper airway resistance syndrome.
[a] Calculated based on mean, standard deviation (SD) and sample size provided in article.

respondents interpret items differently (differential item functioning).

The developer of the ESS chose items to represent situations of a “widely different soporific nature” [13]. This scale has been found to represent from four to six levels of somnificity [12,20,21,40]. Item #5 (“lying down”) has most often been found to be the most sleep-inducing situation [12,20,21]. The least sleep-inducing situations were items #6 (“talking”) and #8 (“in traffic”) [12,20,21,40]. Redundancy was most often reported for two pairs of items, #6 and #8, and #3 (“in a public place”) and #7 (“after a lunch”) [12,20]. This redundancy can inflate summary score interpretation amongst sleepy patients because the same level of somnificity is counted twice. But redundant items representing low somnificity, such as items #6 and #8, will have no impact on scores among less sleepy respondents because those patients will report a low likelihood of falling asleep in these situations. Others have shared our opinion that the sum of all eight items is not likely to be the best index of ASP [37].

The ability of items to measure different levels of somnificity has important theoretical implications that are not taken into consideration in classical reliability and factor analyses models [64]. When items on a scale can be ordered, it is termed 'cumulative' [65]. The ESS may have a cumulative structure. For example, sleepy subjects who report a high chance of dozing in a low soporific situation such as "talking" should also report a high chance of dozing in the more soporific situations, such as "lying down" and "sitting quietly after lunch", resulting in high concordance between items. Similarly, non-sleepy respondents who report a low chance of dozing in a highly soporific situation, "lying down", should also report a low or even lower chance of dozing in less soporific situations such as "talking". Between these extremes, subjects who report a high likelihood of falling asleep in a highly soporific situation such as "lying down" will not necessarily report falling asleep in a situation with a low soporific nature such as "talking"; the result will be reduced concordance between items measuring these different levels of somnificity. Consequently, Cronbach's alphas can vary between subgroups with different levels of DS and the items may not appear to measure a single latent construct [66].

The item descriptions leave room for subjects to make widely different interpretations. For example, the description "in a car, while stopped in the traffic" does not distinguish between being a passenger or a driver. For some items (e.g., "watching TV", "sitting inactive in a public place (e.g., a theater or a meeting)") patients could interpret the location, body position and their level of interest in the situation in different ways. To address this concern, one study [67] compared the performance of the ESS and a new set of questions that included descriptions of a respondent's position, location, and interest in the activity. Below an ESS of 16, the new items and the ESS items behaved differently. Above a total ESS score of 16, sleepiness dominated wakefulness regardless of an activity's location or level of interest. The net result is that among less sleepy subjects different interpretations of location, position and level of interest could affect what type of somnificity the item represents, undermining validity by leading to a situation where patients with identical true levels of ASP are assigned different total scores.

Interpretation of items could be affected also by the universality of the situations they represent as well as the request for subjects to "work out" how the situations would have affected them if they have not done those things recently. For example, "as a passenger" and "in traffic" refer to being in a car, and along with "lying down" may not occur frequently enough in some subjects to allow them to make a valid assessment [40,60]. Furthermore, subjects may watch TV more frequently than they read; as a result, they may recall falling asleep watching TV and report it as being sleep-inducing for that reason alone. Likewise, sleepy subjects who do not read may choose "would never doze" for that question. On the other hand, less sleepy individuals who read a lot and never fall asleep will also choose "would never doze", so two different levels of SSP may lead to the same item score.

The Sleep Heart Health Study found evidence that men and women report DS differently [68] Despite the fact that women reported feeling sleepy and unrested more often than men, they were less likely to have an ESS score above 10. The authors suggested that using the ESS to detect subjective sleepiness is more likely to identify men with sleepiness than women.

Test–retest reliability

Despite its wide use, there is a paucity of good quality evidence for the test–retest reliability of the ESS. In the one good quality study a test–retest correlation of 0.82 was reported [13]; however, there are questions about generalizing this finding in students to the middle-aged population who are most commonly referred to sleep clinics with DS. Furthermore, this study did not use the preferred reliability statistic, the ICC. In addition this study measured the stability of the ESS over a five-month period (the original administration was two months after the start of the academic year, and the retest was two months after the winter vacation). Given the test–retest study requires stability in the dimension being measured between the test and the retest, one might question whether this is the case over a five-month time period. It is possible that a small part of this population may have had changes in their DS over this time period, and as a result the test–retest reliability statistic in the study may be underestimated.

As such, we recommend, further research is required to establish the test–retest of the ESS over time periods when the underlying construct of ASP has not changed and for middle-aged population.

Construct validity

Convergent construct validity

Based on moderate levels of evidence, the association between the ESS and the MWT was moderate (ρ pooled $= -0.43$), and the association between the ESS and the MSLT was weak (ρ pooled $= -0.27$). There was strong evidence that the association between the ESS and OSA related variables (AHI, SaO_2, ArI) was weak, with ρ pooled ranging from 0.11 to 0.23. Although these groups of correlations were in the expected order they tended to be lower than expected based on the a priori hypotheses developed by the research team. This may reflect i) ambiguity regarding the construct of DS measured by the ESS or ii) the high heterogeneity between included studies.

Poor agreement between the various measures of sleepiness, both subjective and objective, may reflect the multidimensional nature of sleepiness [69–72]. Recent research [72] has reported weak correlations among objective measures of sleepiness (the MSLT, psychomotor vigilance test and divided attention driving task), and between the ESS and each of these objective measures of sleepiness. The authors of this paper concluded that a comprehensive evaluation of sleepiness may require multiple measures [72].

In addition, the MSLT and the MWT were reported by some researchers as inaccurate measures of SSP [43]. Not only are they measuring sleepiness in a single situation, but this situation (being in a lab with electrodes attached to the head) is artificial and would never occur in the patient's daily life. It is possible that the smaller than expected correlations of these objective measures to the ESS relates more to the deficiencies of the MSLT and MWT than it does to the ESS.

The weak correlation between the ESS and AHI may be due to the AHI itself having a poor correlation with DS [73,74]. Although the AHI is the most commonly used accepted measure of OSA severity [45], the correlation between AHI and other measures of sleepiness, such as MSLT, is low to non-existent [74,75]. It is unclear whether this stems from OSA not causing DS in many subjects, or in the use of the AHI as a measure of OSA severity. But none of the other physiologic abnormalities associated with OSA (SaO_2, ArI) had noteworthy correlations with the ESS either.

Variation in correlations across studies might also be explained by diversity in target populations, clinical and methodological heterogeneity (see Table S1), and diversity in the definition of comparator instruments. This variation was largest for comparisons between the ESS and AHI and SaO_2 ($I^2 = 84.7\%$ and 80.7%, respectively) that mostly can be explained by diversity in their definitions.

Known-group validity

We found strong evidence supporting differences in ESS scores across groups of subjects with known differences in DS. However, in three [59–61] of five good studies, even though the differences were statistically significant, they were at most 1.5 points, which may not be a clinically important difference [76]. Studies comparing normal subjects and subjects with narcolepsy [19,39,51], each reported clinically meaningful differences. We recommend that future studies examining known-group validity of the ESS make an a priori specification of expected mean differences across groups.

In addition to more high quality studies assessing the reliability and validity of the ESS, other areas of future research on the measurement properties of the ESS may also focus on more general measurement issues such as potential for regression to the mean in ESS scores and floor and ceiling effects on individual ESS items.

Limitations

We acknowledge that the COSMIN approach may set too high standards for achieving good ratings on some of criteria [77]. The COSMIN methods were developed in an attempt to improve future research on measurement, and to challenge readers to be critical while interpreting results. To address this limitation, we present results from both good and fair quality studies. In addition, we want to emphasize that poor quality of study should not be interpreted as reflecting that the ESS is of poor quality, only that limited inference towards the overall measurement quality of the instrument can be gained from such studies.

Finally, although it may be seen as a limitation that we studied only the English-language version of the ESS, we feel this was the right approach. Translation of the ESS can affect interpretation of items [78] and results from one version of the ESS will not necessarily generalize to another.

Conclusion

Although the ESS is widely used in sleep research and clinical settings, overall it has only modest measurement properties and there have been relatively few high quality studies on its psychometric properties.

The internal consistency of the ESS suggests that this instrument can be used for group level comparisons, but caution is recommended if using the ESS for individual level comparison. Questions remain about the unidimensionality of the ESS scale, particularly for items that may occur infrequently or that represent situations where there is a very low probability of falling asleep. There is limited evidence on the test–retest reliability of the ESS. The Item Response Theory model may offer more appropriate methods for scoring and testing the measurement properties of the ESS.

We found larger correlations of the ESS with other measures of DS (the MWT and the MSLT) than with less closely related constructs (OSA and general health measures) but all correlation coefficients were lower than expected. Known-group construct validity was established but future studies should state whether differences in scores between groups are in line with clinical expectations, rather than relying on whether differences between groups are statistically significant. In summary, despite twenty years of use, the ESS requires further studies of high methodological quality to assess its measurement properties. In particular, studies are required in the area of test–retest reliability.

Practice points

A systematic review of the measurement properties of the Epworth sleepiness scale revealed the following:

1) Although the ESS is widely used in sleep research and clinical settings, there have been relatively few high quality studies on its psychometric properties.
2) The internal consistency of the ESS (Cronbach's alphas ranged from 0.7 to 0.9) suggests that this instrument can be used for group level comparisons, but caution is recommended if using the ESS for individual level comparison.
3) Questions remain about the unidimensionality of the ESS scale; particularly, whether items that may occur infrequently or that represent situations where there is a very low probability of falling asleep (e.g., item #6 ("talking") and #8 ("in traffic")) belong to one dimension.
4) There is limited evidence on the test–retest reliability of the ESS.
5) Larger correlations of the ESS with other measures of daytime sleepiness (the MWT (ρ pooled $= -0.43$) and the MSLT (ρ pooled $= -0.27$)) than with less closely related constructs (severity of obstructive sleep apnea and general health measures, ρ pooled ranging from 0.11 to 0.23) were found, but all correlation coefficients were lower than expected.
6) Known-group construct validity was established; however, the differences across groups of comparison may not be clinically important even though they were statistically significant.

Research agenda

1) Despite twenty years of use, the ESS requires further studies of high methodological quality to assess its measurement properties.
   - In particular, studies are required in the area of test–retest reliability with justified appropriate time interval between measures;
   - Future studies examining known-group validity of the ESS should make an a priori specification of expected mean differences across groups of comparison and state whether differences in scores between groups are in line with clinical expectations, rather than relying on whether differences between groups are statistically significant;
   - The Item Response Theory model may offer more appropriate methods for scoring and testing the measurement properties of the ESS.

Acknowledgments

The first author, Dr. Tetyana Kendzerska is supported by 2011/2012 Ontario Graduate Scholarship, 2011/2012 Hunter Graduate

Scholarship (the University of Toronto) and 2012/2013 Doctoral Research Award from the Canadian Institutes of Health Research. Dr. Peter Smith is supported by a Discovery Early Career Research Award from the Australian Research Council.

Conflict of interest statement

Our study had no external funding source. The authors have no conflict of interest to declare pertaining to this review.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.smrv.2013.08.002.

References

[1] Ruggles K, Hausman N. Evaluation of excessive daytime sleepiness. WMJ 2003;102:21–4.
[2] Hays JC, Blazer DG, Foley DJ. Risk of napping: excessive daytime sleepiness and mortality in an older community population. J Am Geriatr Soc 1996;44:693–8.
[3] Newman AB, Spiekerman CF, Enright P, Lefkowitz D, Manolio T, Reynolds CF, et al. Daytime sleepiness predicts mortality and cardiovascular disease in older adults. The Cardiovascular Health Study Research Group. J Am Geriatr Soc 2000;48:115–23.
[4] MacLean AW, Davies DR, Thiele K. The hazards and prevention of driving while sleepy. Sleep Med Rev 2003;7:507–21.
[5] Chasens ER, Sereika SM, Burke LE. Daytime sleepiness and functional outcomes in older adults with diabetes. Diabetes Educ 2009;35:455–64.
[6] Koutsourelakis I, Perraki E, Economou NT, Dimitrokalli P, Vagiakis E, Roussos C, et al. Predictors of residual sleepiness in adequately treated obstructive sleep apnoea patients. Eur Respir J 2009;34:687–93.
[7] Stroe AF, Roth T, Jefferson C, Hudgel DW, Roehrs T, Moss K, et al. Comparative levels of excessive daytime sleepiness in common medical disorders. Sleep Med 2010;11:890–6.
[8] Young TB. Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. J Clin Psychiatry 2004;65:12–6.
[9] American Academy of Sleep Medicine. International classification of sleep disorders: diagnostic and coding manual, second edition (ICSD-2); 2005.
[10] Johns MW. The subjective measurement of excessive daytime sleepiness. In: Pandi-Perumal SR, Verster JC, Monti JM, Lader M, Langer SZ, editors. Sleep disorders: diagnosis and therapeutics. London: Informa Healthcare; 2008. p. 643–57.
[11] Johns M. A sleep physiologist's view of the drowsy driver. Transp Res Part F Traffic Psychol Behav 2000;3:241–9.
[12] Johns W. A new perspective on sleepiness. Sleep Biol Rhythm 2010;8:170–9.
*[13] Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. Sleep 1992;15:376–81.
[14] Weaver TE, Laizner AM, Evans LK, Maislin G, Chugh DK, Lyon K, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. Sleep 1997;20:835–43.
[15] Carskadon MA, Dement WC. The multiple sleep latency test: what does it measure? Sleep 1982;5:S67–72.
[16] Mitler MM, Gujavarty KS, Browman CP. Maintenance of wakefulness test: a polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. Electroencephalogr Clin Neurophysiol 1982;53:658–61.
[17] Akerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Int J Neurosci 1990;52:29–37.
[18] Hoddes E, Dement W, Zarcone V. The development and use of the Stanford sleepiness scale. Psychophysiology 1972;9:150.
[19] Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–5.
[20] Johns MW. Sleep propensity varies with behaviour and the situation in which it is measured: the concept of somnificity. J Sleep Res 2002;11:61–7.
*[21] Johns MW. Sleepiness in different situations measured by the Epworth Sleepiness Scale. Sleep 1994;17:703–10.
[22] Winne F. Distortions of construct validity in multiple regression analysis. Can J Behav Sci 1983;15:187–202.
[23] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.
[24] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737–45.
[25] Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27–36.
[26] Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78:98–104.
[27] Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River, N.J.: Prentice Hall Health; 2000.
[28] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford; New York: Oxford University Press; 2008.
[29] van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine 2003;28:1290–9.
[30] Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2012;21:659–70.
[31] Cooper HM. Research synthesis and meta-analysis: a step-by-step approach. 4th ed. Thousand Oaks, Calif: Sage Publications; 2010.
[32] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[33] Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 2006;11:193–206.
[34] Bourke SC, McColl E, Shaw PJ, Gibson GJ. Validation of quality of life instruments in ALS. Amyotroph Lateral Scler Other Motor Neuron Disord 2004;5:55–60.
[35] Knutson KL, Rathouz PJ, Yan LL, Liu K, Lauderdale DS. Stability of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Questionnaires over 1 year in early middle-aged adults: the CARDIA study. Sleep 2006;29:1503–6.
[36] Nguyen ATD, Baltzan MA, Small D, Wolkove N, Guillon S, Palayew M. Clinical reproducibility of the Epworth Sleepiness Scale. J Clin Sleep Med 2006;2:170–4.
*[37] Smith SS, Oei TPS, Douglas JA, Brown I, Jorgensen G, Andrews J. Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnoea. Sleep Med 2008;9:739–44.
[38] Beaudreau SA, Spira AP, Stewart A, Kezirian EJ, Lui LY, Ensrud K, et al. Validation of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Scale in older black and white women. Sleep Med 2012;13:36–42.
*[39] Spira AP, Beaudreau SA, Stone KL, Kezirian EJ, Lui L-Y, Redline S, et al. Reliability and validity of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Scale in older men. J Gerontol A Biol Sci Med Sci 2012;67:433–9.
[40] Kingshott R, Douglas N, Deary I. Mokken scaling of the Epworth Sleepiness Scale items in patients with the sleep apnoea/hypopnoea syndrome. J Sleep Res 1998;7:293–4.
[41] Stevens J. Applied multivariate statistics for the social sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1992.
[42] McHorney CA, Ware Jr JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.
[43] Sullivan SS, Kushida CA. Multiple sleep latency test and maintenance of wakefulness test. Chest 2008;134:854–61.
[44] Richardson GS, Carskadon MA, Flagg W, Van den Hoed J, Dement WC, Mitler MM. Excessive daytime sleepiness in man: multiple sleep latency measurement in narcoleptic and control subjects. Electroencephalogr Clin Neurophysiol 1978;45:621–7.
[45] Epstein LJ, Kristo D, Strollo Jr PJ, Friedman N, Malhotra A, Patil SP, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J Clin Sleep Med 2009;5:263–76.
[46] Kingshott RN, Engleman HM, Deary IJ, Douglas NJ. Does arousal frequency predict daytime function? Eur Respir J 1998;12:1264–70.
*[47] Banks S, Barnes M, Tarquinio N, Pierce RJ, Lack LC, McEvoy RD. Factors associated with maintenance of wakefulness test mean sleep latency in patients with mild to moderate obstructive sleep apnoea and normal subjects. J Sleep Res 2004;13:71–8.
[48] Sangal RB, Sangal JM, Belisle C. Subjective and objective indices of sleepiness (ESS and MWT) are not equally useful in patients with sleep apnea. Clin Electroencephalogr 1999;30:73–5.
[49] Olson LG, Cole MF, Ambrogetti A. Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms. J Sleep Res 1998;7:248–53.
*[50] Chervin RD, Aldrich MS. The Epworth Sleepiness Scale may not reflect objective measures of sleepiness or sleep apnea. Neurology 1999;52:125–31.
[51] Chervin RD, Aldrich MS, Pickett R, Guilleminault C. Comparison of the results of the Epworth Sleepiness Scale and the Multiple Sleep Latency Test. J Psychosom Res 1997;42:145–55.
[52] Weaver EM, Kapur V, Yueh B. Polysomnography vs self-reported measures in patients with sleep apnea. Arch Otolaryngol Head Neck Surg 2004;130:453–8.
*[53] Weaver EM, Woodson BT, Steward DL. Polysomnography indexes are discordant with quality of life, symptoms, and reaction times in sleep apnea patients. Otolaryngol Head Neck Surg 2005;132:255–62.
[54] Kingshott RN, Sime PJ, Engleman HM, Douglas NJ. Self assessment of daytime sleepiness: patient versus partner. Thorax 1995;50:994–5.

* The most important references are denoted by an asterisk.

[55] Briones B, Adams N, Strauss M, Rosenberg C, Whalen C, Carskadon M, et al. Relationship between sleepiness and general health status. Sleep 1996;19:583–8.
[56] Bennett LS, Barbour C, Langford B, Stradling JR, Davies RJ. Health status in obstructive sleep apnea: relationship with sleep fragmentation and daytime sleepiness, and effects of continuous positive airway pressure treatment. Am J Respir Crit Care Med 1999;159:1884–90.
*[57] Buysse DJ, Hall ML, Strollo PJ, Kamarck TW, Owens J, Lee L, et al. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample [Erratum appears in J Clin Sleep Med. 2010 Feb 15;6(1):table of contents]. J Clin Sleep Med 2008;4:563–71.
[58] Skibitsky M, Edelen MO, Martin JL, Harker J, Alessi C, Saliba D. Can standardized sleep questionnaires be used to identify excessive daytime sleeping in older post-acute rehabilitation patients? J Am Med Dir Assoc 2012;13:127–35.
*[59] Hesselbacher S, Subramanian S, Allen J, Surani S. Body mass index, gender, and ethnic variations alter the clinical implications of the Epworth Sleepiness Scale in patients with suspected obstructive sleep apnea. Open Respir Med J 2012;6:20–7.
[60] Kezirian EJ, Harrison SL, Ancoli-Israel S, Redline S, Ensrud K, Claman DM, et al. Behavioral correlates of sleep-disordered breathing in older women. Sleep 2007;30:1181–8.
*[61] Kezirian EJ, Harrison SL, Ancoli-Israel S, Redline S, Ensrud K, Goldberg AN, et al. Behavioral correlates of sleep-disordered breathing in older men. Sleep 2009;32:253–61.
[62] Bland JM, Altman DG. Cronbach's alpha. BMJ 1997;314:572.
[63] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293–307.
[64] Baker FB, Kim S-H. Item response theory: parameter estimation techniques. 2nd ed. New York: M. Dekker; 2004.
[65] Schuur Wv. Ordinal item response theory: Mokken scale analysis. Los Angeles: Sage; 2011.
[66] Schuur WHv. Mokken scale analysis: between the Guttman scale and parametric item response theory. Polit Anal 2003;11:139–63.
[67] Sharafkhaneh A, Hirshkowitz M. Contextual factors and perceived self-reported sleepiness: a preliminary report. Sleep Med 2003;4:327–31.
[68] Baldwin CM, Kapur VK, Holberg CJ, Rosen C, Nieto FJ. Associations between gender and measures of daytime somnolence in the Sleep Heart Health Study. Sleep 2004;27:305–11.
[69] Cluydts R, De Valck E, Verstraeten E, Theys P. Daytime sleepiness and its evaluation. Sleep Med Rev 2002;6:83–96.
[70] Johns M. Rethinking the assessment of sleepiness. Sleep Med Rev 1998;2:3–15.
[71] Kim H, Young T. Subjective daytime sleepiness: dimensions and correlates in the general population. Sleep 2005;28:625–34.
[72] Sunwoo BY, Jackson N, Maislin G, Gurubhagavatula I, George CF, Pack AI. Reliability of a single objective measure in assessing sleepiness. Sleep 2012;35:149–58.
[73] Gondim LM, Matumoto LM, Melo Junior MA, Bittencourt S, Ribeiro UJ. Comparative study between clinical history and polysomnogram in the obstructive sleep apnea/hypopnea syndrome. Braz J Otorhinolaryngol 2007;73:733–7.
[74] Chervin RD, Guilleminault C. Assessment of sleepiness in clinical practice. Nat Med 1995;1:1252–3.
[75] Walter TJ, Foldvary N, Mascha E, Dinner D, Golish J. Comparison of Epworth Sleepiness Scale scores by patients with obstructive sleep apnea and their bed partners. Sleep Med 2002;3:29–32.
[76] Mulgrew AT, Lawati NA, Ayas NT, Fox N, Hamilton P, Cortes L, et al. Residual sleep apnea on polysomnography after 3 months of CPAP therapy: clinical implications, predictors and patterns. Sleep Med 2010;11:119–25.
[77] Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. BMC Med Res Methodol 2011;11:152 [author reply].
[78] Takegami M, Suzukamo Y, Wakita T, Noguchi H, Chin K, Kadotani H, et al. Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. Sleep Med 2009;10:556–65.
[79] Benbadis SR, Mascha E, Perry MC, Wolgamuth BR, Smolley LA, Dinner DS. Association between the Epworth sleepiness scale and the multiple sleep latency test in a clinical population. Ann Intern Med 1999;130:289–92.
[80] Bennett LS, Langford BA, Stradling JR, Davies RJ. Sleep fragmentation indices as predictors of daytime sleepiness and nCPAP response in obstructive sleep apnea. Am J Respir Crit Care Med 1998;158:778–86.
[81] Gottlieb DJ, Whitney CW, Bonekat WH, Iber C, James GD, Lebowitz M, et al. Relation of sleepiness to respiratory disturbance index: the Sleep Heart Health Study. Am J Respir Crit Care Med 1999;159:502–7.
[82] Gottlieb DJ, Yao Q, Redline S, Ali T, Mahowald MW. Does snoring predict sleepiness independently of apnea and hypopnea frequency? Am J Respir Crit Care Med 2000;162:1512–7.
[83] Johns MW. Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth Sleepiness Scale. Chest 1993;103:30–6.
[84] Martin SE, Engleman HM, Kingshott RN, Douglas NJ. Microarousals in patients with sleep apnoea/hypopnoea syndrome. J Sleep Res 1997;6:276–80.
[85] Osman EZ, Osborne J, Hill PD, Lee BW. The Epworth Sleepiness Scale: can it be used for sleep apnoea screening among snorers? Clin Otolaryngol Allied Sci 1999;24:239–41.
[86] Stavitsky K, Saurman JL, McNamara P, Cronin-Golomb A. Sleep in Parkinson's disease: a comparison of actigraphy and subjective measures. Parkinsonism Relat Disord 2010;16:280–3.