ELSEVIER

Sleep Medicine 4 (2003) 195–199

SLEEP MEDICINE

www.elsevier.com/locate/sleep

Review

Measurement properties of the Epworth sleepiness scale[☆]

Michael S. Miletin, Patrick J. Hanly[*]

Division of Respirology, Department of Medicine, Room 6049, St. Michael's Hospital, University of Toronto, 30 Bond Street, Toronto, Ont. M5B 1W8, Canada

Received 4 March 2002; received in revised form 21 October 2002; accepted 8 January 2003

1. Introduction

Obstructive sleep apnea (OSA) has been estimated to affect 2–4% of middle aged adults [1]. Excessive daytime sleepiness is an important criterion both for establishing the severity of OSA and for determining the response to specific treatment such as nasal continuous positive airway pressure (CPAP) [2]. Objective assessment of daytime sleepiness with tests such as the multiple sleep latency test (MSLT) is both time-consuming and costly and is not offered by many sleep disorders clinics [3]. Consequently, a questionnaire that reliably quantifies the severity of daytime sleepiness and that is responsive to changes in daytime sleepiness over time would greatly assist clinicians and researchers in the management and investigation of OSA.

Since its publication in 1991, the Epworth sleepiness scale (ESS) has been used by several groups of investigators to measure daytime sleepiness in patients with known or suspected OSA [4]. Furthermore, the ESS has also been used to track changes in daytime sleepiness during treatment of OSA [5,6]. This usage of the ESS as an evaluative instrument that measures change over time may not be appropriate given the original goals of its development and the consequent design of the questionnaire. Furthermore, the use of the ESS as a diagnostic tool may be premature. The objectives of this review are to determine what the ESS actually measures and whether that measurement truly reflects objective sleepiness, and to determine if the ESS can be used to diagnose pathological sleepiness and follow its response to treatment.

2. What does the ESS measure?

In order to assess the measurement properties of a questionnaire, one must first ask 'how were the items in the questionnaire selected?'

The ESS was intended to differentiate persons with excessive daytime sleepiness from alert individuals by measuring their sleep propensity, which has been described as the net interaction of the waking drive and the sleep drive [7]. Previous work has shown that sleep propensity is dependent upon the situation in which it is measured [8]. With this understanding, the ESS asks respondents how likely they are to doze in eight different daily situations on a 4-point scale.

However, how the items (or situations) that compose the ESS were chosen is poorly described [4].

Two of the eight questions in the ESS were derived from the results of a previous study published only in abstract form. In this study of a population-based survey of adults in New Mexico, respondents were asked to rank a number of pre-selected daily situations on the basis of their relative soporificity [9]. The two items that were ranked 'most sleepy' ('sitting inactive in church' and 'as a passenger in a moving car') in this questionnaire were modified and incorporated into the ESS. The situations described in the remaining six questions were included without explicit criteria for determining soporificity.

The goal of discrimination between alert individuals and those with daytime sleepiness may also explain how items were selected for the ESS. An instrument with good discriminative properties will detect between-subject differences reliably [10]. For example, a scale for the assessment of asthma severity could easily discriminate between healthy individuals and asthmatics if it asks about symptoms such as dyspnea, wheeze, cough, and chest tightness, which most healthy individuals do not commonly experience. However, most humans experience some level of sleepiness during waking hours [11]. If a sleepiness scale were only

[☆] Competing interest: None declared

[*] Corresponding author. Tel.: +1-416-864-6026; fax: +1-416-864-5417.
E-mail address: hanlyp@smh.toronto.on.ca (P.J. Hanly).

doi:10.1016/S1389-9457(03)00031-5

composed of highly soporific scenarios, it would be unable to discriminate between sleepy and alert individuals. Better discrimination could result if the scale were instead made up of only mildly soporific scenarios, but at the expense of rendering the scale too insensitive to differentiate between individuals with mild and severe sleepiness. The scenarios may have been chosen to vary from highly soporific ('lying down to rest in the afternoon when circumstances permit') to minimally soporific ('sitting and talking to someone') as a compromise between these two extremes.

For the creation of a questionnaire, most authorities recommend a formal item generation phase that includes input from relevant literature, other health professionals, content experts, and most importantly, patients. A number of different techniques can then be employed to select items from the resultant item pool to create the questionnaire [12, 13]. The development of the ESS lacked significant patient input, and as a result, the questionnaire may fail to reflect clinical realities.

Factor analysis of the ESS suggests that the questionnaire measures only one cohesive factor, presumably sleep propensity [14]. In the same publication, the internal consistency of the questionnaire was assessed using Cronbach's alpha, which is a reasonable statistic to perform on a scale that purports to measure a single attribute [15]. In patients with OSA, alpha was 0.88, suggesting a high level of internal consistency and little redundancy [16]. Significantly, alpha did not increase after deleting any one of the questions from the questionnaire, indicating that no item acts to reduce internal consistency. Several different questions pertaining to the same factor (sleep propensity) are necessary to allow for the expected variability in the daily activities of respondents and maintain good content validity (discussed in greater detail below). The high value of alpha in OSA patients helps to restore some of the confidence that had been lost due to the possible inadequacy of item generation. The ESS manages to measure a single patient attribute. Whether or not this attribute is sleep propensity or even daytime sleepiness requires testing of the validity of the questionnaire.

3. Does the ESS measure objective sleepiness?

Validity testing asks if an instrument truly measures what it purports to measure [15]. In the case of the ESS, which is used primarily to differentiate among respondents and to assess change in level of daytime sleepiness over time, important components of validity are face validity, content validity, and construct validity. Face validity is present when examination of the items in a questionnaire indicates that those items pertain to the attribute that is being measured. High content validity implies that a measure is able to represent a wide range of circumstances that reflect its target behaviour or attribute (sleep propensity in the case of the ESS) [17]. An instrument with a high degree of construct validity will show strong correlations with other measurements of the same attribute or its consequences.

Evaluation of the face validity of the ESS reveals that the focus of the questionnaire is directed towards measuring the propensity to fall asleep in different situations.

Examination of the content validity of the ESS fails to reveal any grossly inappropriate items in the scale. However, due to the lack of a formal item generation phase, the potential for important omissions exists. Given the well-recognized association between OSA and motor vehicle accidents [18], it is surprising that 'while driving a car' was not included as an item in the questionnaire. This would have been distinct from the eighth question in the ESS, 'in a car, while stopped for a few minutes in the traffic'. This question does not make clear whether the respondent is to imagine themselves as the passenger or the driver. Moreover, most motor vehicle accidents take place while the vehicle is in motion, not while stopped at a traffic light. The ESS does not include a question asking patients to describe their likelihood of falling asleep while at work, considering that the time spent at work may comprise the majority of the waking hours of many individuals. Besides assessing behaviour in daily situations, clinicians also pursue other lines of questioning in order to gauge the sleepiness of a patient. Examples include estimation of the number of caffeinated beverages consumed daily, the need for planned daytime naps, and patients' assessments of changes in work performance, memory, and overall level of energy. These issues are not addressed by the items in the ESS.

Assessment of construct validity is dependent upon the existence of relevant literature. The construct measured by the ESS is sleep propensity or daytime sleepiness. A construct can be regarded as a 'mini-theory' which gains credibility (validity) by passing several tests, or hypotheses. These hypotheses ask how the construct may be related to other variables. In the case of the ESS, the variables may be the results of other tests that measure daytime sleepiness (the construct). In 44 patients referred to a sleep disorders clinic, a significant, but moderate correlation (rho $= -0.42$, $P < 0.01$) was found between the ESS and the mean sleep latency measured by the MSLT [19]. A similar report revealed a coefficient of -0.37 ($P = 0.004$) [20]. While statistically significant, these correlations are only moderately strong. Furthermore, Benbadis and coworkers failed to find any correlation between the ESS and sleep latency measured by the MSLT in a retrospective series of 102 patients, 80 of whom had sleep-disordered breathing [21]. After examining data on 237 consecutive patients, Chervin could not demonstrate any meaningful relationship between ESS score and sleep latency, although the ESS score was related to respondents' subjective feeling of sleepiness at the time of testing [22]. The data are conflicting in part because the ESS and the MSLT do not measure exactly the same thing. The MSLT measures sleepiness at the point of testing. On the other hand, the ESS attempts to gauge how sleepy respondents have been at a variety of daily activities in the recent past. In addition, the MSLT suffers from imperfect validity and reliability [8,23] that may further decrease the strength of its correlation with the ESS.

The apnea-hypopnea index and the respiratory disturbance index refer to the number of apneas and hypopneas per hour of sleep and are commonly used to define the presence and the severity of obstructive sleep apnea [24]. Data from the Sleep Heart Health Study indicate that ESS scores increase significantly with increases in the number of apneas and hypopneas per hour in a study population that was selected independent of the diagnosis of sleep apnea [25]. Johns has shown a significant negative correlation between ESS score and the minimum oxygen saturation during sleep [7], but this relationship has not been confirmed in other studies [20].

Another group of variables are those that measure other constructs or attributes in the same patients. For example, one may hypothesize that daytime sleepiness should be correlated with health-related quality of life. Bennett et al recently studied the relationship of overall health-related quality of life to daytime sleepiness measured by the ESS. Health-related quality of life was measured using the SF-36 questionnaire in 51 subjects referred to a sleep disorders clinic who also completed the ESS. The ESS was correlated negatively with both the Physical Component Score ($r = -0.43$) and the Mental Component Score ($r = -0.51$) ($P < 0.01$ for both). The energy/vitality dimension of the SF-36 was that most strongly correlated with the ESS score ($r = -0.47$, $P < 0.001$). These findings indicate that the ESS correlates at least modestly and in the expected direction with related domains of a well-validated quality of life instrument [26,27].

Clinical experiments provide opportunities to test longitudinal construct validity, an important property for an evaluative instrument [10]. Here, changes in the ESS are related to changes in an external measure over time. Engleman et al. showed that ESS scores decreased from 15 ± 6 to 7 ± 5 ($P < 0.001$) after successfully treating patients with OSA with nasal CPAP [28]. The authors also hypothesized that the rate of self-reported traffic accidents would also decrease after CPAP treatment. In concert with the decrease in ESS scores, patients did report a significantly decreased rate of traffic accidents after initiating CPAP treatment. A general population study in the UK has since found that the ESS score is significantly and moderately correlated with the likelihood of falling asleep while driving a car [29]. These results indicate that the ESS does have longitudinal construct validity.

The ESS may reflect objective daytime sleepiness, as defined by some validated outcome measurements, but the data is conflicting. There is little data to support it as a measurement of sleep propensity, perhaps owing to a lack of consensus over the definition of this concept.

4. Should the ESS be used to diagnose pathologic sleepiness and assess the response to treatment?

Adequate questionnaire validity is not sufficient to recommend its use as a diagnostic tool. The instrument must also be shown to be reliable. Reliability refers to the ability of an instrument to provide similar results when administered on repeated occasions to respondents whose measured attributes have remained stable. A high degree of reliability minimizes background ‘noise’ and allows for any changes recorded over time to be confidently interpreted as reflecting a true change in the respondent, not measurement variability [30].

The major source of measurement variability in the ESS is the patient who completes the questionnaire. Patient variability is influenced by the spectrum of disease within a tested group, recall bias, actual clinical change over time, and the testing conditions [31]. The ESS attempts to circumvent the latter factor by asking patients to estimate their sleep propensity during the past few weeks and not at the time of testing.

In order for the ESS to merit use as an evaluative instrument, it must be shown to have adequate test-retest reliability. Although the ESS has never undergone reliability testing in patients with OSA, it has been evaluated in healthy medical students [14]. However, the methods used reveal some methodological and analytic flaws. The ESS was completed by 104 medical students during a regularly scheduled class, and was repeated 5 months later. Seventeen students did not complete the second iteration. On the first occasion the mean score was 7.4 ± 3.9 (standard deviation, SD), and on the second it was 7.6 ± 3.8 (SD). Reliability was assessed using Pearson’s product-moment correlation. The result indicated a strong correlation, with $r = 0.822$ ($P < 0.001$). However, correlation is not an adequate estimate of reliability. Scores that do not agree on retesting could nonetheless still be correlated. No indication of the degree of similarity between scores is provided by the correlation coefficient, which may therefore be misleading [15]. A better statistic would have been the intraclass correlation coefficient (ICC), which increases as within-person variability decreases and between-person variability increases. No subsequent studies have been performed to redress the lack of adequate reliability testing of the ESS.

From a psychometric perspective, the high degree of internal consistency of the ESS measured by Cronbach’s alpha statistic has adequately established the instrument’s reliability [16]. Since each question measures the same thing (sleepiness), the completion of the ESS is akin to multiple test-retesting of the questionnaire. The intraclass correlation coefficient can be shown to be equivalent to Cronbach’s alpha, which was 0.73 in the students and 0.88 in 54 patients with OSA described in the same report [14]. However, there would be greater confidence in the reliability of the ESS if the ICC had been calculated in a group of OSA patients with a broad spectrum of disease severity. Consequently, the lack of convincing reliability data in patients with OSA limits the interpretation of the responsiveness of the ESS.

Responsiveness refers to the ability of a measurement instrument to detect change over time. The ratio of the change (the 'signal') to the error resulting from measuring stable subjects repeatedly (the 'noise') provides an index of responsiveness [10]. Examination of the ESS reveals that the majority of its questions should not adversely affect its responsiveness, which is critical if it is to be used to assess change in daytime sleepiness following treatment of OSA. However, one may take issue with one question, specifically sleepiness while 'lying down to rest in the afternoon when circumstances permit'. The scenarios comprising the ESS were chosen according to their supposed soporificity. In the development of the ESS, this scenario was considered to be the most soporific, that is, most individuals, including those without excessive daytime sleepiness, would indicate at least a slight chance of dozing off in that situation [4]. If this is true, it is unlikely to help the ESS serve as an evaluative instrument, since the answers of patients with OSA to this question are unlikely to change with treatment. If healthy individuals doze off in this situation, then patients with OSA are likely to do so as well. Even if CPAP restores these patients to 'normal', they are still likely to doze off, and the aggregated ESS score may not have the resolution necessary to detect whether a decrease in the likelihood of dozing off has occurred.

No formal assessment of the responsiveness of the ESS has yet been performed. Clinical trials of CPAP for the treatment of OSA have measured ESS scores pre- and post treatment [5,6]. However, none of these trials have measured ESS scores at repeated intervals in patients prior to the initiation of therapy. Therefore, the variability of scores in stable patients with OSA is unknown. In the sample of 87 medical students who were administered the ESS twice at an interval of 5 months, there was a mean difference in scores of 0.2, with a standard deviation of 2.3 [14]. This represents a variation in measurement of 21%. Ballester et al. reported a difference of 0.8 in mean ESS scores over a 3-month period in OSA patients who had been randomized to receive conservative treatment (i.e. without CPAP) [5]. This result may have been biased by the fact that it was derived from patients already in a clinical trial, who might behave differently from those studied without the benefit of the same close monitoring and follow-up.

The lack of reliability testing in patients with significant daytime sleepiness (and/or OSA) combined with the conflicting validity data described in the previous section suggests that the ESS should not currently be used to diagnose pathological sleepiness. Furthermore, the paucity of relevant data poses significant limitations to the use of the ESS for the determination of therapeutic efficacy (e.g. with CPAP in the case of OSA).

5. Conclusions

In summary, the ESS seems to measure only one factor and has a high degree of internal consistency. Validity testing does not unequivocally support the use of the ESS as a measure of daytime sleepiness or sleep propensity. However, ESS scores do relate to important clinical outcomes such as road traffic accidents and health-related quality of life. Serious reservations exist regarding the reliability and responsiveness of the ESS, given the lack of test-retest data in patients with OSA. Due to the limitations posed by the lack of adequate construct validity and responsiveness data, use of the ESS as a diagnostic tool and as a clinical outcome measure may not be justified at the present time. Future research should explore these issues with methodologic and statistical rigour.

References

[1] Young T, Palta M, Dempsey J, et al. The occurrence of sleep disordered breathing among middle-aged adults. N Engl J Med 1993; 328:1230–5.

[2] Strollo Jr PJ, Rogers RM. Obstructive sleep apnea. N Engl J Med 1996;334:99–104.

[3] Carskadon M, Dement WC, Mitler MM, Roth T, et al. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. Sleep 1986;9:519–24.

[4] Johns MW. A new method for measuring daytime sleepiness. Sleep 1991;14:540–5.

[5] Ballester E, Badia JR, Hernandez L, et al. Evidence of the effectiveness of continuous positive airway pressure in the treatment of sleep apnea/hyponea syndrome. Am J Respir Crit Care Med 1999; 159:495–501.

[6] Engleman HM, Kingshott RN, Wraith PK, et al. Randomized placebo-controlled crossover trial of continuous positive airway pressure for mild sleep apnea/hypopnea syndrome. Am J Respir Crit Care Med 1999;159:461–7.

[7] Johns MW. Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth sleepiness scale. Chest 1993;103:30–6.

[8] Sangal RB, Thomas L, Mitler MM. Maintenance of wakefulness test and multiple sleep latency test: measurement of different abilities in patients with sleep disorders. Chest 1992;101:898–902.

[9] Schmidt-Nowara WW, Wiggins CL, Walsh JK, et al. Prevalence of sleepiness in an adult population. Sleep Res 1989;18:302.

[10] Kirshner B, Guyatt GH. A methodologic framework for assessing health indices. J Chron Dis 1985;38:27–36.

[11] Briones B, Adams N, Strauss M, et al. Sleepiness and health: relationship between sleepiness and general health status. Sleep 1996; 19:583–8.

[12] Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ 1986;134:889–95.

[13] Juniper EF, Guyatt GH, Epstein RS, et al. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax 1992;47:76–83.

[14] Johns MW. Reliability and factor analysis of the Epworth sleepiness score. Sleep 1992;15:376–81.

[15] Streiner DL, Norman GR. Health Measurement Scales: a practical guide to their development and use, 2nd ed. Oxford: Oxford University Press; 2000.

[16] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

[17] Feinstein AR. The theory and evaluation of sensibility. In: Feinstein

AR, editor. Clinimetrics. Westford, MA: Murray Printing Company; 1987.

[18] Young T, Blustein J, Finn L, et al. Sleep-disordered breathing and motor vehicle accidents in a population-based sample of employed adults. Sleep 1997;20:608–13.

[19] Johns MW. Sleepiness in different situations measured by the Epworth sleepiness scale. Sleep 1994;17:703–10.

[20] Chervin RD, Aldrich MS, Pickett R, Guilleminault C. Comparison of the results of the Epworth sleepiness scale and the multiple sleep latency test. J Psych Res 1997;42:145–55.

[21] Benbadis SR, Mascha E, Perry MC, Wolgamuth B, et al. Association between the Epworth sleepiness scale and the multiple sleep latency test in a clinical population. Ann Int Med 1999;130:289–92.

[22] Chervin RD, Aldrich MS. The Epworth sleepiness scale may not reflect objective measures of sleepiness or sleep apnea. Neurology 1999;52:125–31.

[23] Roth T, Hartse K, Zorick F, Conway W. Multiple naps and the evaluation of daytime sleepiness in patients with upper airway sleep apnea. Sleep 1980;3:425–39.

[24] Strohl KP, Redline S. Recognition of obstructive sleep apnea. Am J Respir Crit Care Med 1996;154:279–89.

[25] Gottlieb DJ, Whitney CW, Bonekat WH, et al. Relation of sleepiness to respiratory disturbance index. Am J Respir Crit Care Med 1999; 159:502–7.

[26] Bennett LS, Barbour C, Langford B, Stradling JR, et al. 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.

[27] McHorney C, Ware JE, Raczek A. 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.

[28] Engleman HM, Asgari-Jirhandeh N, McLeod AL, et al. Self-reported use of CPAP and benefits of CPAP therapy. Chest 1996;109:1470–6.

[29] Maycock G. Sleepiness and driving: the experience of UK care drivers. Accid Anal Prev 1997;29:453–62.

[30] Guyatt GH, Kirschner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45: 1347–51.

[31] Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol 1992;45:1201–18.