JCSM
Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

# Clinical Reproducibility of the Epworth Sleepiness Scale

Anh Tu Duy Nguyen, M.D.[1]; Marc A. Baltzan, M.D., M.Sc.[1,2]; David Small, M.D.[1]; Norman Wolkove, M.D.[1]; Simone Guillon, M.D.[3]; Mark Palayew, M.D.[1]

*[1]Mount Sinai Hospital, Montreal, Canada; [2]Department of Epidemiology and Statistics, McGill University, Montreal, Canada; [3]Snoring Clinic, Pointe Claire, Quebec, Canada*

**Study Objectives:** The Epworth Sleepiness Scale (ESS) is widely used as a subjective measure of sleepiness. To our knowledge, no study has evaluated its reproducibility in the clinical setting.

**Methods:** A retrospective chart review of patients referred to the sleep clinic at Mount Sinai Hospital for evaluation of sleep-disordered breathing from a local private snoring clinic between January 2000 and October 2001 was carried out. Patients were snorers and referred because of suspicion of sleep apnea. Clinical information including results of the ESS scores from the two institutions was analyzed to evaluate reproducibility.

**Results:** There were 142 patients evaluated: 76% men with a mean (SD) age of 44 (11) years, body mass index of 31 (6.1) kg/m$^2$, and apnea-hypopnea index of 41 (34) events per hour. The average time interval between ESS administrations was 71 (92) days. The average ESS score was 11.1 (5.2) at the Snoring Clinic and 11.2 (5.3) at Mount Sinai Hospital. The Bland-Altman plot of the difference against the mean of the ESS score demonstrated a wide scatter of data and variability where 2 SDs ranged 7.8 above and below the mean. A difference between the sequential ESS scores of 5 or more was seen in 23% of the subjects.

**Conclusion:** The ESS score is highly variable when administered sequentially to a clinical population being evaluated for a potential sleep-related breathing disorder.

**Keywords:** Epworth sleepiness scale, reproducibility, sleep apnea, sleepiness

**Citation:** Nguyen ATD; Baltzan M; Small D et al. Clinical reproducibility of the epworth sleepiness scale. *J Clin Sleep Med* 2006;2(2):170-174.

The Epworth Sleepiness Scale (ESS) was developed by Johns as a simple, self-administered questionnaire to assess sleep propensity.[1,2] The subject is asked to rate his or her probability of dozing in each of 8 different situations on a scale of 0 to 3 (0 = no chance of dozing, 1 = slight chance of dozing, 2 = moderate chance of dozing, 3 = high chance of dozing; minimum score = 0, maximum score = 24). Initially published in 1991, the ESS was originally validated in 30 normal control subjects and 150 patients with various sleep disorders.[1] In this population, the ESS was helpful in differentiating normal subjects from those diagnosed with a sleep disorder.

The ESS is widely used in clinical practice and research protocols as a simple rapid assessment of subjective sleepiness. Sleep disorders clinics may prioritize patients for polysomnography based on the ESS results. In addition, the ESS score is frequently used in research studies as a means of quantifying changes in habitual subjective sleep propensity after an intervention.[3,4]

To our knowledge evaluation of the reproducibility of the ESS has been limited to normal subjects who have normal ESS score results.[5-8] We had subjects who had previously completed the ESS at a private snoring clinic repeat the ESS at the time of assessment at the sleep disorders clinic at Mount Sinai Hospital. We hypothesized that a significant variability may exist when the ESS is administered sequentially in subjects being evaluated for a sleep-related breathing disorder.

## METHODS

After ethics approval by the Mount Sinai Hospital Ethics Committee, a retrospective chart review was performed at a private snoring clinic that often referred patients for further evaluation and at the sleep disorders clinic at Mount Sinai Hospital. The Snoring Clinic accepts patients for the management of snoring. Patients are evaluated by a physician and an oromaxillofacial surgeon. Mount Sinai Hospital is a tertiary referral center for sleep disorders with a dedicated polysomnographic laboratory.

The ESS was administered with similar methodology in both centers. The patients had no intervention between completion of the ESS done at each clinic, other than being advised to lose weight and/or being prescribed nasal steroids if symptoms of nasal congestion were present.

Nocturnal polysomnography was performed on all patients. Sleep staging was performed using standard electroencephalographic leads (C4, C3, O1, O2). Respiratory monitoring consisted of airflow measured with nasal pressure cannulae, thoracic and abdominal movements measured by inductive plethysmography, snoring measured by cervical microphone, and arterial oxyhemoglobin saturation measured by finger pulse oximetry. All signals were acquired on a digital data management system (Sandman 6.1, Kanata, Ontario, Canada). The apnea-hypopnea index (AHI) was calculated from the total number of apneas (total cessation of breathing for > 10 seconds) and the total number of hypopneas

**Disclosure Statement**
This was not an industry supported study. Drs. Palayew, Nguyen, Baltzan, Small, Wolkove, and Guillon have indicated no financial conflicts of interest.

**Submitted for publication July 28, 2005**
**Accepted for publication November 15, 2005**
Address correspondence to: Mark Palayew M.D., Respiratory Division, Room G 203, SMBD- Jewish General Hospital, 3755 Cote Ste Catherine RD., Montreal, Quebec, Canada; H3T 1E2; Tel: (514) 340-7900; Fax: (514) 340-7555; E-mail: mpalayew@pne.jgh.mcgill.ca

Downloaded from jcsm.aasm.org by 190.2.133.230 on March 25, 2022. For personal use only. No other uses without permission.
Copyright 2022 American Academy of Sleep Medicine. All rights reserved.

(decrease in airflow by at least 50% for > 10 seconds, or decrease in airflow by at least 30% for > 10 seconds, associated with oxygen desaturation of at least 3% or terminated by an unequivocal electroencephalographic cortical arousal) throughout the night divided by the total hours of electroencephalographic sleep.

Patients were selected based on consecutive referrals from the Snoring Clinic to Mount Sinai Hospital between January 2000 and October 2001.

The following information was extracted during the retrospective chart review: subject demographics, total ESS score from each center, scores for each of the ESS questions, the time interval between the ESS completion at the 2 centers, and the results of polysomnography for each patient.

Data Analysis

To evaluate the reproducibility of the ESS, we assessed group means and correlations and individual differences in the Epworth scores. Comparisons between the group means obtained at the Snoring Clinic and Mount Sinai Hospital were performed using a 2-tailed Student t test for paired samples. The differences between sequential Epworth scores at both clinics were calculated. Pearson correlation coefficients were calculated to evaluate the correlation of individual scores at both clinics. Repeatability was assessed using the Bland-Altman plot,[9,10] as well as calculation of the percentage of patients with discrepancies in the Epworth score. Factor analysis was used to assess the dominant and other potential factors measured in our population by the ESS. Internal consistency of the ESS was evaluated using Cronbach α.[11] To evaluate the 8 individual questions of the ESS, each question was separately analyzed with calculations of Pearson correlation coefficients, mean discrepancy between the first and second answer, and calculation of the percentage of patients with discrepancies in their answers of 1 or more. Significant differences in mean discrepancy across the 8 questions were tested with one-way analysis of variance. Regression analysis was used to evaluate the effect of age by 10-year categories, obstructive sleep apnea diagnosis (AHI of 10 or more), AHI quartiles, sex, and time interval between measurements (continuous and by quartiles) on the differences between sequential Epworth scores. Statistical analysis was performed using SAS software (SAS Institute, Cary, NC).

## RESULTS

All patients identified through this review were referred for suspicion of obstructive sleep apnea. A total of 159 charts were identified and evaluated of which 142 were retained for analysis. The reasons for exclusion of 17 patients were as follows: there was no ESS score available for 2 patients in the Mount Sinai Hospital chart; there was no ESS score available for 10 patients in the Snoring Clinic chart; 5 patients had documentation of an ESS score in both charts, but the answers to the individual questions were either incomplete or not available (2 at Mount Sinai Hospital, 3 at the Snoring Clinic). The characteristics of these 17 individuals were not considered to be significantly different from the 142 patients retained for analysis. The mean (SD) age of these 17 patients was 50.0 (9.4) years, 11 (65%) of whom were men, with a mean body mass index of 30.1 (5.2) kg/m$^2$ and an AHI of 35.3 (SD 37.3, range 4 to 135) events per hour.

During the 21-month time interval, 142 consecutive patients were evaluated and had complete data available for assessment. Seventy-six percent were men, with a mean (SD) age of 44 (11) years and mean AHI of 41.0 (SD 34.4, range 1 to 153) events per hour. An AHI < 10 was found in 16% of patients who completed polysomnography. The average time interval between ESS completion was 71 (92) days. There was no statistically significant difference between the group mean ESS score of 11.1 (5.2) at the Snoring Clinic and of 11.2 (5.3) at Mount Sinai Hospital (p = .89). The Pearson correlation coefficient was 0.73 (p < .001).

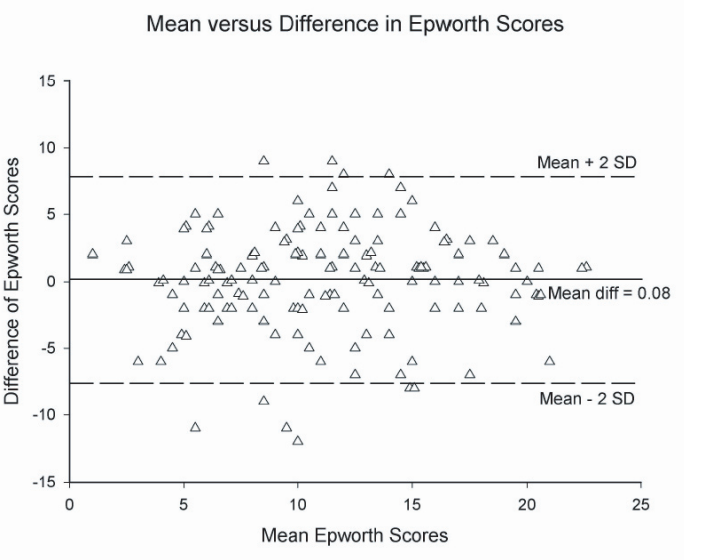


**Figure 1**—Graph of the mean versus the difference of the Epworth score. The Bland and Altman plot is a graphical method of assessing agreement between 2 measurements, in this case the 2 Epworth scores for each individual. Each triangle represents an individual data point.

The Bland-Altman plot, a graphical representation of the ESS score distribution with the mean of the ESS score plotted against the ESS score difference (Figure 1), was used to assess repeatability.[9,10] Visual inspection of the plot revealed a wide scatter of values around 0, with -7.8 to 7.8 as the range of ESS score difference within 2 SD above and below the mean.

An ESS score difference of 3 or more occurred in 41% of patients, an ESS score difference of 5 or more was present in 23% of patients, and an ESS score difference of 7 or more was seen in 10% of subjects (Table 1).

Analysis of the individual questions of the ESS showed that the mean discrepancy varied significantly across the 8 questions of the ESS score, with questions 6 (Sitting and talking to someone) and 8 (In a car, while stopped for a few minutes in traffic) being the least variable, and question 7 (Sitting quietly after a lunch without alcohol) being the most variable (Table 2).

Regression analysis did not demonstrate an effect of age, OSA diagnosis, AHI (Figure 2), sex, and time interval between measurements on the differences between sequential Epworth scores (all p values >.20).

**Table 1**—Discrepancies Between ESS Scores in Studies That Measured Test-Retest Reliability

| Discrepancy in ESS Score | Study Author | | | |
|---|---|---|---|---|
| | Bloch[5] | Chung[6] | Johns[8] | Current |
| ≥ 2 | NA | 46 | 48 | 61 |
| ≥ 3 | 26 | 27 | 18 | 41 |
| ≥ 5 | NA | 4 | 3 | 23 |
| ≥ 7 | NA | NA | NA | 10 |

Data are presented as percentages. ESS refers to Epworth Sleepiness Scale.

Downloaded from jcsm.aasm.org by 190.2.133.230 on March 25, 2022. For personal use only. No other uses without permission.
Copyright 2022 American Academy of Sleep Medicine. All rights reserved.

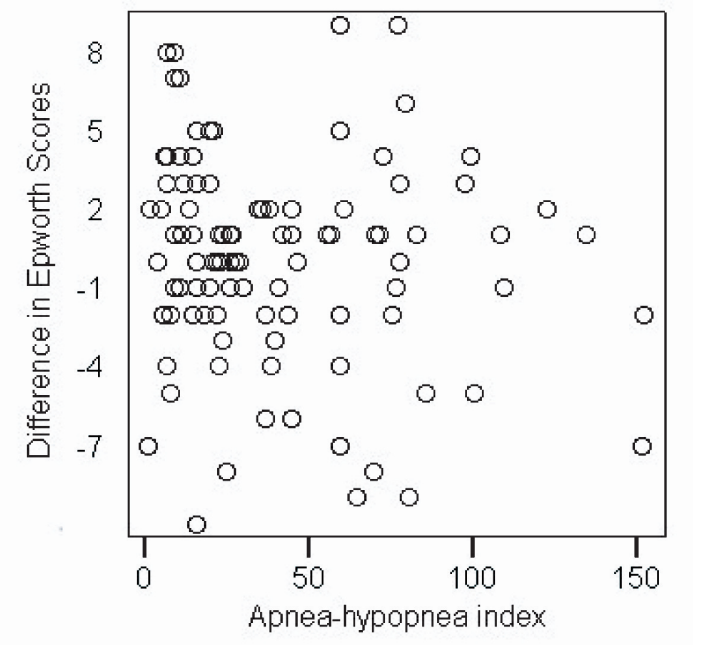


**Figure 2**—Difference in Epworth Scores vs Apnea Hypopnea Index. Each open circle represents an individual data point.

Internal consistency of the ESS in our clinical population was evaluated using Cronbach α statistics. The response of the clinical questionnaire was assessed separately for the Snoring Clinic and for Mount Sinai Hospital. Cronbach α was 0.78 and 0.82, respectively.

Factor analysis confirmed that sleepiness was the dominant factor measured in our population (eigenvalue = 6.2); 2 other factors had eigenvalues of 1.7 and 1.4, respectively. These 2 minor significant factors were not clearly discernable from the major factor.

## DISCUSSION

The ESS is a simple, self administered questionnaire that evaluates subjective sleepiness.[1] Clinically, it is widely used as a means of assessing sleepiness in sleep disorders clinics. Research protocols often use sequential ESS scores to assess the response to an intervention.[3,4] In this study, a large variability in ESS scores on sequential testing was found in subjects referred to a tertiary care sleep disorders clinic for possible sleep-disordered breathing.

Johns evaluated the reproducibility of the ESS by having 87 healthy medical students complete the scale on two occasions, with a 5-month interval.[8] In Johns' study, the mean initial ESS score was 7.4 (3.9), the mean ESS score 5 months later was 7.6 (3.8), and the mean difference was 0.2 (2.3). The Pearson correlation coefficient was 0.82 (highly significant). The difference between the ESS score on sequential testing was ≥ 5 in only 3% of the subjects. On the basis of this analysis, Johns concluded that the ESS was a reliable method for measuring persistent daytime sleepiness in adults. The correlation between the 2 individual measurements of the ESS in our study was also relatively high (0.73). The correlation coefficient, however, is a measure of the strength of a relation between 2 variables and not the agreement between them. Correlation coefficients are not able to assess reproducibility.[9]

**Table 2**—Scores on Individual Epworth Sleepiness Scale Questions

| Question | Correlation | Discrepancy[a] | Discrepancy > 1, % |
|---|---|---|---|
| 1 | 0.51 | 0.63 ± 0.76 | 11 |
| 2 | 0.63 | 0.53 ± 0.63 | 7 |
| 3 | 0.50 | 0.68 ± 0.59 | 12 |
| 4 | 0.67 | 0.59 ± 0.73 | 10 |
| 5 | 0.50 | 0.53 ± 0.82 | 10 |
| 6 | 0.48 | 0.40 ± 0.60 | 6 |
| 7 | 0.38 | 0.85 ± 0.84 | 18 |
| 8 | 0.68 | 0.39 ± 0.61 | 6 |

To evaluate repeatability of the ESS, we used the Bland-Altman plot (Figure 1). The Bland-Altman plot demonstrated a wide scatter of values for ESS score differences within the same subject on sequential testing, with a ± 2 SD range of nearly -8 to 8. This demonstrates a large intrasubject variability in the ESS scores. The mean difference in the ESS score is a means of quantifying bias (the greater the difference from 0, the stronger the bias). The mean difference in the ESS score is 0.08 (p = .89), suggesting that no significant bias is present to explain the variability of the ESS score. Furthermore, there is no clear pattern to the data points, with no increase in scatter with increase in mean ESS score. This implies that the variability is not related to the magnitude of the ESS score.

As an alternative method of analyzing the test-retest variability of the ESS, we evaluated the percentage of subjects who had a difference of at least 2, 3, 5, and 7 between ESS scores. This method took advantage of data reported in the previous studies evaluating the reproducibility of the ESS, thus enabling comparisons between studies (Table 1). In our study, 41% of the subjects had difference in the ESS score of 3 or more, and 23% had a change in the ESS score of 5 or more. This is a greater discrepancy than anticipated from 3 previous studies[5,6,8] that evaluated test-retest variability. The initial study by Johns reported a change in ESS score of 3 or more in 18% of healthy subjects and of 5 or more in 3% of healthy subjects.[8] Chung et al, in evaluating test-retest characteristics of the Chinese version of the ESS, noted a change in the ESS score of 3 or more in 27% and of 5 or more in 4% of subjects studied.[6] Similarly, a study of the German version of the ESS, in which only data for a change in the ESS score of 3 or more was available, found that 26% of subjects had a change in the ESS score of 3 or more.[5]

There are 3 major differences between the previous studies and this study. First, previous studies that examined the reproducibility of the ESS were limited to normal subjects,[5-8] whereas our assessment of test-retest variability was performed in a sleep-clinic population. The subjects in the other studies consisted of hospital staff members in a German study,[5] nonmedical hospital staff or friends of the author in a Chinese study,[6] and medical students in Johns' repeatability study.[8] Second, our current study of 142 subjects is, to our knowledge, the largest group in which test-retest variability of the ESS has been evaluated. Third, this study evaluated an unselected clinical population of consecutive patients seen within a defined time period.

A potential bias toward subjects with milder obstructive sleep apnea may have developed because subjects were initially evaluated at the Snoring Clinic for evaluation of snoring. Sleep-related breathing abnormalities were generally severe, as assessed by polysomnography (mean AHI of 41 events per hour), however, so our population is similar to other respiratory sleep clinic populations.[3,4]

We suggest 3 possible reasons for the discrepancy between

Downloaded from jcsm.aasm.org by 190.2.133.230 on March 25, 2022. For personal use only. No other uses without permission.
Copyright 2022 American Academy of Sleep Medicine. All rights reserved.

the results of our study and previous studies. First, by evaluating test-retest variability in a clinical population with sleep-related breathing disorders, we have a more valid assessment of the performance characteristics of the ESS than by extrapolating from data obtained in student or staff populations. Healthy population samples would be expected to have normal ESS scores; one would expect low variability on this basis alone. The variability in the ESS score may only become apparent when subjects with a wider range of ESS scores, demonstrating varying degrees of subjective sleepiness from sleep-related breathing disorders, are evaluated. Second, by evaluating a larger group of consecutive subjects, we may be able to detect variability that may have been missed with smaller selected populations. Third, although our study has the potential limitation of being retrospective in nature, the selection of subjects who were unaware of the study suggests that our results are more clinically applicable than the previously cited validation studies.

The basis for the discrepancy between the ESS scores remains to be defined, yet certain factors can be excluded as significant contributors. The only clinical intervention after completion of the ESS at the Snoring Clinic was instruction for overweight subjects to lose weight and the occasional prescription of nasal steroids. Fewer than 2% of subjects lost 10 pounds or more. None of the subjects were treated for obstructive sleep apnea with continuous positive airway pressure, oral appliances, or surgery prior to completing the ESS at Mount Sinai Hospital. The fact that the group means differed by only 0.08 is an argument against any important systematic intervention changing the clinical status of patients during the waiting period to be seen at Mount Sinai Hospital. Knowledge of the first measurement affecting the second would be expected to increase the difference between the mean ESS score from 0, which was not found in our study. In addition, regression analysis did not reveal any correlation between the length of time between testing, age, sex, or AHI severity and the difference between the 2 Epworth scores.

To determine whether there was a particular question or questions from the ESS that led to the poor test-retest characteristics of the ESS, we conducted an analysis of correlation and discrepancy for each of the 8 questions (Table 2). The most variability was seen in Question 7 ("Sitting quietly after lunch without alcohol"), which often scored 1 or 2. The least variability was seen in Questions 6 and 8 ("Sitting and talking to someone" and "In a car, while stopped for a few minutes in traffic"), which often scored 0. We believe that the low variability observed in Questions 6 and 8 is due to a floor effect exemplified by the frequent 0 responses.

In an attempt to characterize the nature of the problems with reproducibility in our clinical population, we analyzed the ESS using factor analysis, which confirmed the presence of a single dominant factor with a high eigenvalue of 6.2. This factor represents sleepiness. The finding of 2 coincident minor factors suggests that the ESS may be unintentionally measuring aspects other than sleepiness, such as boredom or inattentiveness. We also assessed internal consistency using Cronbach α, which is a measure of how well items in a scale correlate with one another. The Cronbach α was 0.78 in the Snoring Clinic and 0.82 at Mount Sinai Hospital. The Cronbach α statistic is a measure of internal consistency within a scale using a calculation based on the number of items in the scale and the total variance of the total score, as well as the sum of the variants of the individual question scores. Clinical usefulness for comparison of repeated measures within individuals is considered to require a score of more than 0.90; scores below 0.90 but above 0.70 may be adequate to compare repeated measures for calculations within groups or for comparison of means between groups, such as is often done in clinical trials. Our calculated Cronbach statistics for the Epworth Scale of 0.78 and 0.82 are within the expected range for a scale to be appropriate for use in comparison of means when groups are considered. These values for the statistic support our findings that the Epworth scale does not perform with sufficient consistency to be clinically useful for repeated measures within the same individual.

Sleepiness, and more particularly subjective sleepiness, is difficult to quantify. This symptom, however, is often the chief complaint in a sleep medicine practice. The ESS was developed in an attempt to provide a clinical tool to quantify subjective sleepiness. While we did not attempt to assess content or construct validity of the ESS, we demonstrated that, in terms of scales, it has important problems with internal consistency and test-retest repeatability, both important elements in validating the ESS.[13] Johns has demonstrated that the ESS correlates only weakly with sleep latency during multiple sleep latency testing.[14] Furthermore, in studies of patients with obstructive sleep apnea, the ESS score has been correlated weakly with sleep-onset latency measured by various methods and has been correlated only weakly if at all with AHI and oximetry.[15-23]

In conclusion, when used in the clinical setting of evaluating subjects with potential sleep-related breathing disorders, the ESS frequently has large discrepancies when repeated over time in the same untreated individual. We caution against using the ESS as the sole tool for risk stratification of patients referred for possible sleep apnea or for response following treatment interventions. Further work is required to evaluate the reproducibility of the ESS in the clinical setting. If the problems with reproducibility are confirmed, we suggest a reevaluation of the ESS to better understand the features that limit its reproducibility and to permit further development of questionnaires with better test-retest characteristics.

## REFERENCES


1. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540-5.
2. Johns MW. Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth Sleepiness Scale. Chest 1993;103:30-6.
3. Jenkinson C, Davies RJ, Mullins R, et al. Comparison of therapeutic and subtherapeutic nasal continuous positive airway pressure for obstructive sleep apnoea: a randomised prospective parallel trial. Lancet 1999;353:2100-5.
4. 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.
5. Bloch KE, Schoch OD, Zhang JN, et al. German version of the Epworth Sleepiness Scale. Respiration 1999;66:440-7.
6. Chung KF. Use of the ESS in Chinese patients with obstructive sleep apnea and normal hospital employees. J Psychosom Res 2000;49:367-72.
7. Izquierdo-Vicario Y, Ramos-Platon MJ, Conesa-Peraleja D, et al. Epworth Sleepiness Scale in a sample of the Spanish population. Sleep 1997;20:676-7.
8. Johns MW. Reliability and factor analysis of the ESS. Sleep 1992;15:376-81.
9. Bland JM, Altman DG. Statistical methods for assessing agreement


Downloaded from jcsm.aasm.org by 190.2.133.230 on March 25, 2022. For personal use only. No other uses without permission.
Copyright 2022 American Academy of Sleep Medicine. All rights reserved.

between two methods of clinical measurement. Lancet 1986;1:307-10.
10. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346:1085-7.
11. Mitler MM, Miller JC. Methods of testing for sleepiness. Behav Med 1996;21:171-83.
12. Bland JM, Altman DG. Cronbach's alpha. BMJ 1997;314:572.
13. Bland JM, Altman DG. Statistics notes: validating scales and indexes. BMJ 2002;324:606-7.
14. Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth Sleepiness Scale: failure of the MSLT as a gold standard. J Sleep Res 2000;9:5-11.
15. Benbadis SR, Mascha E, Perry MC, et al. Association between the ESS and the multiple sleep latency test in a clinical population. Ann Intern Med 1999;130:289-92.
16. Chervin RD, Aldrich MS, Pickett R, et al. Comparison of the results of the ESS and the Multiple Sleep Latency Test. J Psychosom Res 1997;42:145-55.
17. Punjabi NM, Bandeen-Roche K, Young T. Predictors of objective sleep tendency in the general population. Sleep 2003;26:678-83.
18. Walter TJ, Foldvary N, Mascha E, et al. Comparison of ESS scores by patients with obstructive sleep apnea and their bed partners. Sleep Med. 2002;3:29-32.
19. Leng PH, Low SY, Hsu A, et al. The clinical predictors of sleepiness correlated with the multiple sleep latency test in an Asian Singapore population. Sleep 2003;26:878-81.
20. Banks S, Barnes M, Tarquinio N, et al. 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.
21. Chervin RD, Aldrich MS, Pickett R, et al. Comparison of the results of the ESS and the Multiple Sleep Latency Test. J Psychosom Res 1997;42:145-55.
22. Chervin RD, Aldrich MS. The ESS may not reflect objective measures of sleepiness or sleep apnea. Neurology 1999;52:125-31.
23. Chen NH, Johns MW, Li HY, et al. Validation of a Chinese version of the Epworth sleepiness scale. Qual Life Res 2002;11:817-21.

Downloaded from jcsm.aasm.org by 190.2.133.230 on March 25, 2022. For personal use only. No other uses without permission.
Copyright 2022 American Academy of Sleep Medicine. All rights reserved.