pii: jc-17-00585 http://dx.doi.org/10.5664/jcsm.7258

JCSM
Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

# The Epworth Sleepiness Scale: Validation of One-Dimensional Factor Structure in a Large Clinical Sample

Brittany R. Lapin, PhD, MPH[1]; James F. Bena, MS[1]; Harneet K. Walia, MD[2]; Douglas E. Moul, MD[2]

*[1]Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio; [2]Center for Sleep Disorders, Cleveland Clinic, Cleveland, Ohio*

**Study Objectives:** The Epworth Sleepiness Scale (ESS) is used by clinicians and researchers to determine level of daytime sleepiness. The number of factors included in the scale has been debated. Our study objective was to clarify the dimensionality of the ESS using a large clinical sample.

**Methods:** A retrospective cohort study included all patients presenting for care in a tertiary care sleep disorders center who answered all items on the ESS from January 8, 2008 to September 28, 2012. Dimensionality was assessed using scree plot, eigenvalues, factor loadings, principal factor analysis, and confirmatory factor analysis. Multigroup confirmatory factor analysis (MGCFA) evaluated dimensionality within 10 subgroups of clinical interest.

**Results:** The mean age of the 10,785 study participants was 50 (± 15) years with 49% female, and 81% white. The one-factor solution explained 63% of the variability in responses with high factor loadings (> .67 for all 8 items). The scree plot identified one factor with eigenvalue > 1. Results of confirmatory factor analysis demonstrated a one-factor solution had acceptable goodness of fit as assessed by root mean square error of approximation of .094 (90% confidence interval: .089–.099). MGCFA confirmed measurement invariance within all 10 demographic and clinical subgroups.

**Conclusions:** Our study confirmed the unidimensionality of the ESS in a large diverse clinical population. Results from this study can be used to justify the interpretation of the ESS within clinical populations, and supports valid comparisons between groups based on the ESS. Future studies are warranted to further understand the items comprising the ESS and potentially eliminate redundant items for increased efficiency in clinical settings.

**Keywords:** dimensionality, Epworth Sleepiness Scale, ESS, multigroup confirmatory factor analysis, patient-reported outcomes, sleepiness factor analysis, unidimensional

**Citation:** Lapin BR, Bena JF, Walia HK, Moul DE. The Epworth Sleepiness Scale: validation of one-dimensional factor structure in a large clinical sample. *J Clin Sleep Med.* 2018;14(8):1293–1301.

**BRIEF SUMMARY**

**Current Knowledge/Study Rationale:** The Epworth Sleepiness Scale (ESS) is the most commonly used measure for clinicians and researchers to measure daytime sleepiness. One of the critical assumptions of measurement theory is unidimensionality, or that a set of items forming a scale all measure one common underlying trait. There has been substantial controversy surrounding ESS's dimensionality.

**Study Impact:** Our study confirmed unidimensionality of the ESS's factor structure through the use of a clinical sample of 10,785 patients and robust statistical methodology. Results from this study can be utilized to justify the interpretation of the ESS within clinical populations, as well as support valid comparisons between groups based on the ESS. Additionally, our findings pave the way for future work using item-response theory models.

## INTRODUCTION

Daytime sleepiness is a prevalent clinical problem resulting in decreased quality of life and increased accidents, and is potentially indicative of underlying physiologic conditions.[1] The propensity to doze, or drowsiness, is the most often used operational definition of sleepiness that sleep clinicians address daily in their care of patients. The Epworth Sleepiness Scale (ESS) is the easiest and most commonly used measure for clinicians and researchers to measure sleepiness.[2] In 1991, Johns constructed the ESS to assess the self-rated likelihood for dozing in commonly encountered everyday situations.[3] The ESS has been shown to have good test-retest reliability, high internal consistency, concurrent validity with objective tests of sleepiness, and discriminate validity compared to other symptom dimensions.[2] It has been translated and tested in a number of languages, and has occasionally required modifications to some of the items.[4–6] Thousands of studies have used the ESS and its practical utility is substantial, allowing clinicians and sleep researchers to compare results across patients as well as within individual patients over time.

One of the critical assumptions of measurement theory is unidimensionality, or that a set of items forming a scale all measure one common underlying trait. This assumption forms the basis for valid calculation of total summed scores, and interpretation of the scores. In order for items on a scale to be summed into a meaningful score for clinicians and researchers to interpret, the scale needs to be unidimensional, meaning all items measure a single construct of daytime sleepiness. If this assumption is not met, a score cannot be compared appropriately between people or samples to evaluate the trait. Dimensionality also has implications for subsequent evaluations using item-response theory (IRT) models, as psychometric theory requires unidimensionality.

Since the ESS was constructed in 1991, there has been substantial controversy surrounding its dimensionality. In several analyses, the unidimensionality of the ESS has been supported[7–15]; however, there are other reports indicating the ESS may have a two-, three-, and even four-dimensional factor structure.[6,16–22,23] Using an overall (unidimensional) score of a multidimensional instrument may result in a loss of information that could reflect an important characteristic of the population. It could also lead to incorrect inferences with consequential outcomes. Across the literature, factor analysis studies of the ESS have been hindered by small sample sizes. The limited sample sizes of these studies leaves open the possibility that the factor structures derived were unstable. A systematic review of 35 studies evaluating psychometric properties of the ESS in adults found very few high-quality studies on the ESS psychometric properties, with the dimensionality of the ESS remaining unsettled.[2] They concluded the IRT model may offer more appropriate methods for scoring and testing the measurement properties of the ESS; however, one assumption of IRT models is that the measure be unidimensional.

Another important assumption of measurement theory is invariance, meaning person factors such as age, sex, and disease severity do not affect the way the instrument is answered. Some studies have found variation between subgroups suggesting items may not measure a single latent construct within all respondents. Differences in internal reliability of the ESS have been found by race, age, and sex.[24,25] Differences in factor structure have been reported between nonclinical samples such as students and healthy community controls, and clinical samples with sleep disorders.[7,20]

The current study aimed to clarify the dimensionality of the ESS's factor structure through the use of a clinical sample of more than 10,000 patients. Our primary objective was to assess dimensionality of ESS using a large generalizable sample of patients seen at a sleep disorders center. Our secondary aim was to evaluate measurement invariance across various clinical subgroups to confirm dimensionality within all patients. Through validating the structure of the ESS, researchers and clinicians will be able to ensure the appropriate methodology is applied to the ESS, allowing for accurate conclusions based on scores as well as paving the way for future work using the IRT model.

## METHODS

### Study Design

We performed a retrospective cohort study of all patients presenting for care in the Cleveland Clinic Sleep Disorders Center for the first time between January 8, 2008 and September 28, 2012 who completed questionnaires as part of their initial clinical contact. As part of routine care, patient- and clinician-reported scales are collected through the Knowledge Program® (KP), an electronic platform for the systematic collection of patient-reported information.[26] Patient-reported outcome measurements are administered on tablets at the time of their clinic visit or through the electronic health record patient portal (MyChart, Epic Systems, Verona, Wisconsin, United States) prior to their appointment. The data were linked to the electronic medical record, as well as to the polysomnographic results, as available. This study was approved by the Cleveland Clinic Institutional Review Board. All adult patients age 18 years and older who completed at least one questionnaire were included in the study cohort.

### Clinical and Patient-Reported Data

Demographics and clinical characteristics were obtained from the electronic health record. Approximate household income was estimated based on the 2010 census data by household ZIP code. Clinical and sleep-related comorbid diagnoses were collected from ICD-9 codes. The apnea-hypopnea index (AHI) was determined using American Academy of Sleep Medicine scoring rules in patients who underwent polysomnographic testing, and data were obtained from a clinically maintained Polysmith-based database.

The Epworth Sleepiness Scale (ESS) was among the questionnaires that the patients were asked to complete through the KP. The ESS is an eight-item self-reported questionnaire that requires patients to rate their likelihood of falling asleep, or dozing, in eight different situations on a four-point scale from 0 ("would never doze") to 3 ("high chance of dozing"). The questions have different levels of soporific difficulty, or likelihood to be endorsed, dependent on where a patient falls along the continuum. Items 1 (sitting and reading), 2 (watching TV), and 5 (lying down to rest in the afternoon) are representative of the most soporific situations, where items 6 (sitting and talking to someone) and 8 (in a car, while stopped for a few minutes in the traffic) represent the least soporific, and items 3 (sitting in a public place), 4 (as a passenger in a car), and 7 (sitting quietly lunch) are intermediate. The level of agreement to each question was summed to provide a total score, from 0 to 24, per patient, which is capable of distinguishing individuals over the full spectrum of daytime sleepiness.

Additional questionnaires were completed through the KP portal. All the questionnaires in this study calculated respective global scores by addition of the item scores. The Insomnia Severity Index (ISI) is a seven-item scale validated for measuring insomnia severity, with each item scored 0–3.[27] An ISI score $\geq$ 15 indicates clinical insomnia. The Patient Health Questionnaire-9 (PHQ-9) depression screen is a nine-item scale validated for measuring depression severity in primary care settings, each item scaled 0–3.[28] A PHQ-9 score $\geq$ 20 indicates severe depression. The Fatigue Severity Scale (FSS) is a 9-item tool validated for assessing fatigue with each item graded on a scale of 1 (disagree) to 7 (agree) with a score $\geq$ 36 considered abnormal fatigue.[29]

To examine dimensionality within different subgroups of patients, the following subgroups were chosen by experts (DM, HW) based on clinical relevance: sex, age 70 years or older, race, income based on median split (< $53,944), AHI $\geq$ 30, diagnosis groups (obstructive sleep apnea (OSA) only and insomnia only), ISI $\geq$ 15, and FSS $\geq$ 36.

### Statistical Methods

Patient characteristics were compared for nonresponse biases between the subset of patients included in the study sample

(complete responses to all items comprising the ESS) and patients who completed part of or none of the ESS. Categorical variables were compared using chi-square test and continuous variables were compared using t-test or Mann-Whitney *U* test (nonparametric), as appropriate. A two-step process was utilized to assess evidence of dimensionality. The first step identified an optimal factor structure using principal axis factor analysis and the second step confirmed the structure using confirmatory factor analysis (CFA). To assess dimensionality through both steps, the total study sample was divided into two random samples: Sample 1 was randomly selected for the exploratory analysis and Sample 2 was assigned as the confirmatory sample. Descriptive summaries of the ESS items and total score were tabulated by sample. The association between ordinal items was assessed using polychoric correlation coefficients. Polychoric correlations assume that the ordinal responses reflect a normally distributed measure that has been cut to derive the responses. Additionally, Spearman correlation coefficients with 95% confidence intervals were calculated to evaluate the relationship among items and the total score.

Adequacy of the Sample 1 data for factor analysis was evaluated using the Kaiser-Meyer-Olkin (KMO) test for sampling adequacy and Bartlett test of sphericity. A KMO value greater than .6 indicates factor analysis is appropriate, and a significant Bartlett test ($P < .05$) indicates the correlation matrix is significantly different from the identity matrix, which would be indicative of poor conditions to fit a factor analysis.[30] Factor loadings greater than .32 were considered sufficient, whereas items with factor loadings of .32 or greater on more than one factor were considered cross-loading.[31] A scree plot of eigenvalues and parallel analysis was used to suggest the appropriate number of factors to include. Principal axis factor analysis, based on polychoric correlations, was performed using oblique (oblimin) rotation for two or more factors. In our analysis, when three factors were considered, the principal axis extraction method failed to converge, so a minimum residual factor extraction method was used instead.

The emergent factor structure was tested in Sample 2 using CFA. Model goodness-of-fit was assessed using the Comparative Fit Index (CFI) and root mean square error of approximation (RMSEA), with values ≥ .90 and < .10 indicating adequate model fit, respectively.[32] After confirming unidimensionality of ESS, internal consistency was evaluated via ordinal alpha. Ordinal alpha more accurately estimates reliability for data on a Likert scale as compared to the more widely used Cronbach alpha.[33]

Multigroup confirmatory factor analysis (MGCFA) analyses were conducted to explore measurement invariance within specific subgroups of patients, as previously defined, within Sample 2. Measurement invariance is established when items on a questionnaire measure identical constructs across different groups. Measurement invariance is composed of configural, metric, and scalar invariance, and was tested by comparing three models with increasingly stringent equality constraints.[34,35] To test configural invariance, which determined if the ESS had the same number of factors in each subgroup and the same pattern of parameters, separate models were constructed for each subgroup, imposing a one-factor structure and allowing model parameters to be freely estimated. A one-factor structure was appropriate for each subgroup, and group was included in a baseline model as a covariate and model fit statistics were tabulated as measures of configural invariance. Next, metric invariance was tested to determine whether the subgroups had equal factor loadings. This was assessed by restricting the parameters, or factor loadings, in the baseline model to be equivalent across groups. Last, assuming that metric invariance was satisfied, scalar invariance determined whether subgroups had similar intercepts. Model intercepts and factor loadings were set to be equal across groups in the baseline model. The nested models were compared using the change of CFI. A change in CFI ≤ .01 was considered acceptable to establish measurement invariance.[36] Model parameters were estimated using the weighted least squares means and variance adjusted estimator for ordinal indicators. As model fit was the primary outcome and was evaluated via multiple criteria, no adjustments for multiplicity were made.

Analyses were performed using SAS version 9.4 statistical software (SAS Inc., Cary, North Carolina, United States) and R software version 3.2.4,[37] with functions from the psych[38] package used to perform the factor analysis, the nFactors[39] package to fit the scree plot, and the lavaan[40] package for conducting MGCFA.

## RESULTS

A total of 12,047 adult patients were eligible to answer the initial tablet survey, of which 10,785 provided a complete set of ESS item responses. The patient characteristics for the study are provided in **Table 1**. The mean age of the study cohort was 49.6 (± 15.0) years, with 49.1% female, and 80.5% white. The most prevalent comorbidities included hypertension (37.6%), depression (19.7%), and diabetes (19.4%). CPAP use was indicated in 2,213 patients (20.7%). Tests for potential sampling bias suggested a higher percentage of female patients than male in those who did not complete all items compared to patients who completed all ESS items (62.4% versus 49.1%, respectively, $P < .01$). Statistically significant differences in age and comorbidities were indicated between patients who completed all items and those who did not, although very few of the differences are clinically relevant. Rates of sleep-related comorbidities, including sleep apnea, insomnia, and restless legs syndrome (RLS), however, were substantially lower in excluded patients. Patients in the study cohort were randomized to either Sample 1 (n = 5392, 48.8% female, mean age 49.4 ± 15.0) or Sample 2 (n = 5393, 49.3% female, mean age 49.7 ± 15.0). All study characteristics were similar between the two samples (**Table S1** in the supplemental material).

**Table 2** shows the descriptive summaries of the ESS questions and total score by sample, with a mean ESS of 9.4 (± 5.7) for Sample 1. Questions 6 (sitting and talking to someone) and 8 (in traffic) have large floor effects, with most patients indicating they have no chance of dozing (72.2%, 74.2%, respectively, in Sample 1). Question 5 (lying down in the afternoon) resulted in the fewest number of participants indicating they had no chance of dozing (8.4% in Sample 1).

**Table 1**—Patient characteristics of ESS study cohort and excluded patients, n = 12,047.

| | Patient Cohort | Excluded Patients | *P* |
|---|---|---|---|
| Total number of patients, n (%) | 10,785 (89.5) | 1,262 (10.5) | |
| Female Sex, n (%) | 5,292 (49.1) | 787 (62.4) | < .001 |
| Age (years), mean ± SD | 49.6 ± 15.0 | 51.2 ± 16.6 | < .001 |
| Race, n (%) | | | .44 |
| White | 8,205 (80.5) | 981 (82.1) | |
| Black | 1,758 (17.3) | 190 (15.9) | |
| Other | 224 (2.2) | 24 (2.0) | |
| Household income (×$1k), median (q1, q3) | 53.9 (42.6, 66.4) | 53.9 (44.0, 67.3) | .14 |
| Comorbidities, n (%) | | | |
| Diabetes | 2,095 (19.4) | 173 (13.7) | < 0.001 |
| Depression | 2,119 (19.7) | 222 (17.6) | .09 |
| Hypertension | 4,054 (37.6) | 424 (33.7) | .007 |
| Sleep-related comorbidities, n (%) | | | |
| Sleep apnea | 8,160 (76.7) | 451 (36.2) | < .001 |
| Insomnia | 3,685 (34.6) | 249 (20.0) | < .001 |
| Restless legs | 1,800 (16.9) | 93 (7.5) | < .001 |
| CPAP use | 2,213 (20.7) | 54 (22.8) | .42 |
| Polysomnographic sleep study | 6,529 (60.5) | 358 (28.4) | < .001 |
| AHI*, median (q1, q3) | 15.7 (6.9, 28.2) | 15.6 (7.9, 26.5) | .70 |
| Patient-reported measurements, mean ± SD | | | |
| Insomnia severity scale | 17.4 ± 5.6 | 18.9 ± 4.2 | .23 |
| Fatigue severity scale | 42.6 ± 14.8 | 44.2 ± 13.7 | .27 |
| PHQ-9 depression score | 9.0 ± 6.4 | 6.7 ± 6.0 | < .001 |

Study cohort included patients who completed all items on the ESS. Excluded patients are those who did not complete all eight items on the ESS. *P* values obtained by chi-square test, *t* test, or Mann-Whitney *U* test, as appropriate. * = 6,887 patients underwent a polysomnographic sleep study and had an AHI measurement. AHI = apnea-hypopnea index, ESS = Epworth Sleepiness Scale, SD = standard deviation, q = quartile.

**Table 2**—Descriptive summaries for ESS questions by sample.

| Chance of Dozing | Sample 1, n = 5,392 | | | | Sample 2, n = 5,393 | | | |
|---|---|---|---|---|---|---|---|---|
| | None (0) | Slight (1) | Moderate (2) | High (3) | None (0) | Slight (1) | Moderate (2) | High (3) |
| ESS1: Sitting and reading | 929 (17.2) | 1,838 (34.1) | 1,353 (25.1) | 1,272 (23.6) | 967 (17.9) | 1,774 (32.9) | 1,439 (26.7) | 1,213 (22.5) |
| ESS2: Watching TV | 776 (14.4) | 1,889 (35.0) | 1,545 (28.7) | 1,182 (21.9) | 736 (13.6) | 1,963 (36.4) | 1,540 (28.6) | 1,154 (21.4) |
| ESS3: Sitting inactive in a public place (eg, theater or meeting) | 2,126 (39.4) | 1,776 (32.9) | 957 (17.8) | 533 (9.9) | 2,165 (40.1) | 1,793 (33.3) | 932 (17.3) | 503 (9.3) |
| ESS4: As a passenger in a car for an hour without a break | 1,525 (28.3) | 1,600 (29.7) | 1,021 (18.9) | 1,246 (23.1) | 1,577 (29.2) | 1,611 (29.9) | 1,080 (20.0) | 1,125 (20.9) |
| ESS5: Lying down to rest in the afternoon when circumstances permit | 453 (8.4) | 1,160 (21.5) | 1,229 (22.8) | 2,550 (47.3) | 447 (8.3) | 1,235 (22.9) | 1,238 (23.0) | 2,473 (45.8) |
| ESS6: Sitting and talking to someone | 3,893 (72.2) | 1,045 (19.4) | 353 (6.5) | 101 (1.9) | 3,883 (72.0) | 1,078 (20.0) | 340 (6.3) | 92 (1.7) |
| ESS7: Sitting quietly after lunch without alcohol | 1,832 (34.0) | 1,794 (33.3) | 1,058 (19.6) | 708 (13.1) | 1,888 (35.0) | 1,792 (33.2) | 1,009 (18.7) | 704 (13.1) |
| ESS8: In a car, while stopped for a few minutes in traffic | 4,004 (74.2) | 975 (18.1) | 284 (5.3) | 129 (2.4) | 4,036 (74.8) | 950 (17.6) | 284 (5.3) | 123 (2.3) |
| **ESS Total:** Mean (SD); Median (IQR); Range | 9.4 (5.7); 9 (5, 13); 0–24 | | | | 9.3 (5.7); 9 (5, 13); 0–24 | | | |

Count and percentage of patient response to eight questions comprising ESS. ESS Total score descriptive statistics included. ESS = Epworth Sleepiness Scale, IQR = interquartile range, SD = standard deviation.

Items were all significantly correlated with one another ($P < .001$ for all). Question 1 (sitting and reading) was most highly correlated with questions 2 (watching TV) and 3 (sitting inactive in a public place) (Sample 1 $r = .74, .70$, respectively;

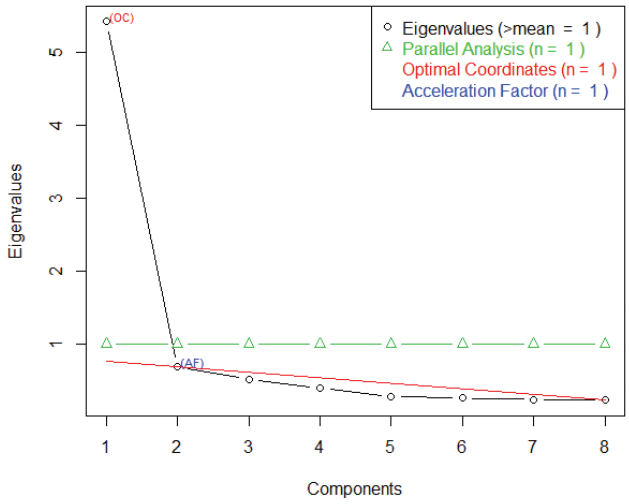
Sample 2 $r$ = .73, .70). Question 3 was highly correlated with questions 6 (sitting and talking to someone), 7 (sitting quietly after lunch), and 8 (stopped for traffic) (Sample 1 $r$ = .75, .71, .71; Sample 2 $r$ = .76, .70, .68). The lowest correlations were demonstrated between question 5 (lying down to rest in the afternoon) and questions 6 (talking) and 8 (Sample 1 $r$ = .46, .43; Sample 2 $r$ = .47, .43). **Table 3** shows correlations among the questions and the total score. All questions were significantly correlated with the total score, with question 8 (traffic) having the lowest correlation (Sample 1 $r$ = .60; Sample 2 $r$ = .59).

Sample 1 data were adequate for factor analysis as confirmed by a KMO statistic of .908 and a significant Bartlett test of sphericity ($P$ < .001). **Figure 1** provides a scree plot of the eigenvalues. The first and second eigenvalues were 5.43 and 0.685, respectively, indicating one factor was most appropriate for the ESS.

**Table 4** shows results of the one, two-, and three-factor solutions in Sample 1. With one or two factors, the principal axis and minimum residual factor extraction method yielded similar results. In the one-factor solution, all questions had loadings between .67 and .86, and the one-factor solution explained 63.4% of the variability in responses. For the two-factor rotation, the two factors explain 68.4% of the variability in responses. Factor 1 loaded heavily on the sleepiness in public, while talking, and in traffic questions, whereas factor 2 had greater loadings on reading, television, resting, and after lunch questions. The passenger question cross-loaded on factors one and two. The principal axis solution would not converge with three factors, so a minimum residual extraction method was used instead. The three factors found explained 75.4% of the variability in responses.

We conducted confirmatory factor analysis procedures with Sample 2. Results of the CFA substantiated a one-factor solution had acceptable goodness of fit as assessed by the CFI of .983 and RMSEA of .094 (90% confidence interval: .089–.099) (**Table S1**). After unidimensionality was established in Sample 2 patients, internal consistency reliability was assessed using ordinal alpha coefficient, as well as the effect of dropping individual questions on the alpha measure (**Table 5**). The overall ordinal alpha was very high (.93), and removing any of the

**Figure 1**—Scree plot and parallel analysis for the 8 items in the ESS in Sample 1, n = 5,392.




The parallel analysis criterion identifies where observed eigenvalues fall below random chance. The optimal coordinate (OC) method identifies the number of eigenvalues based on regression, subject to a minimum value of 1, whereas acceleration factor (AF) is based on where changes in eigenvalues slow, subject to the same eigenvalue minimum of 1, as dictated by Kaiser rule. ESS = Epworth Sleepiness Scale.

**Table 3**—Spearman correlation coefficients between each item and the total score by sample.

| Items | Sample 1, n = 5,392 | Sample 2, n = 5,393 |
|---|---|---|
| ESS1: Reading | .822 (.813, .830) | .821 (.812, .829) |
| ESS2: TV | .775 (.764, .785) | .771 (.760, .781) |
| ESS3: Public | .808 (.798, .817) | .803 (.793, .812) |
| ESS4: Passenger | .783 (.772, .793) | .779 (.768, .790) |
| ESS5: Resting | .709 (.696, .722) | .700 (.687, .714) |
| ESS6: Talking | .647 (.631, .662) | .646 (.630, .661) |
| ESS7: Afternoon | .814 (.805, .823) | .809 (.800, .818) |
| ESS8: Traffic | .601 (.584, .618) | .590 (.572, .607) |

Spearman correlation coefficient with 95% confidence interval between item and total ESS score. ESS = Epworth Sleepiness Scale.

**Table 4**—Results from one-, two-, and three-factor principal axis extraction solution in Sample 1.

| | One-Factor Solution | Two-Factor Solution | | Three-Factor Solution | | |
|---|---|---|---|---|---|---|
| Question | Factor 1 | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 3 |
| Variance Explained | 63.4% | 38.1% | 30.3% | 37.6% | 21.3% | 16.5% |
| ESS1: Reading | .832 | .293 | .598 | .171 | .676 | .073 |
| ESS2: TV | .769 | .251 | .573 | −.024 | .879 | .007 |
| ESS3: Public | .865 | .693 | .221 | .668 | .228 | .022 |
| ESS4: Passenger | .780 | .475 | .351 | .613 | .033 | .218 |
| ESS5: Resting | .673 | −.114 | .866 | −.002 | .008 | .995 |
| ESS6: Talking | .818 | .839 | .029 | .771 | .138 | −.053 |
| ESS7: Afternoon | .844 | .369 | .531 | .505 | .172 | .269 |
| ESS8: Traffic | .772 | .903 | −.080 | .944 | −.106 | −.022 |

One-factor solution conducted with principal axis extraction solution with orthogonal rotation, two-factor solution with principal axis extraction solution with oblique rotation, and three-factor solution with minimum residual extraction solution with oblique rotation. ESS = Epworth Sleepiness Scale.

questions resulted in a drop in the internal consistency of the tool, indicating all items contributed to the construct.

Multigroup confirmatory factor analysis was conducted based on comparison between multiple subgroups of interest as determined *a priori*. To determine measurement invariance across subgroups in Sample 2, the configural invariance of the one-factor model was examined first within each subgroup of interest (**Table S2** in the supplemental material). AHIs and Insomnia Severity Scales were available in a subset of patients (n = 3,265 and n = 1,042, respectively), with the one-factor solution demonstrating acceptable goodness of fit. Although the RMSEA was higher than acceptable, the 90% confidence interval did include a range lower than the cutoff criteria of .10, for all but the subset of patients with PHQ-9 scores 20+. The one-factor solution resulted in acceptable goodness of fit consistently within each subgroup, and item factor loadings were above threshold within all groups (data not shown).

**Table 5**—Ordinal alpha scores overall and by dropping individual questions in Sample 2.

| Factor | Ordinal Alpha |
|---|---|
| All | .931 |
| ESS1: Reading | .918 |
| ESS2: TV | 922 |
| ESS3: Public | .916 |
| ESS4: Passenger | .921 |
| ESS5: Resting | .930 |
| ESS6: Talking | .919 |
| ESS7: Afternoon | .917 |
| ESS8: Traffic | .923 |

Top row presents overall ordinal alpha score for all eight items. Rows present ordinal alpha for the remaining seven items, after excluding the listed item. ESS = Epworth Sleepiness Scale.

Given the values of RMSEA and CFI, the one-factor solution was supported in all subgroups, allowing for further assessment of metric and scalar invariance. Model fit statistics were first calculated in the baseline model by including the subgroup as a covariate in the configural invariance model (**Table 6**). In the metric invariance model, loadings were set to be equivalent across groups, and the values of RMSEA and CFI all indicated excellent fit. The change in CFI between the configural and nested metric models were all below .01, indicating no differences in psychological meaning across groups. Last, scalar invariance was tested by setting both loadings and intercepts equal across groups. RMSEA and CFI values demonstrated excellent fit, and the change in CFI between metric and nested scalar models were less than .01 for all subgroups. These results suggest that there are no systematic response biases across different groups of patients.

## DISCUSSION

The current study provides robust evidence for the ESS having a one-dimensional structure explaining 63% of total variance in our sample population of English-speaking clinical patients. Alternatively, the two-factor solution explained 68% of total variance, whereas the three-factor solution could not be analytically derived. The most popular heuristic to determine the optimal number of factors is based on the eigenvalue-greater-than-one rule and the scree plot. In the current study, both of these indicate one factor adequately explained the correlations among items. The current study confirms a number of prior studies in other populations and languages that have found a one-dimensional structure.[7–15] In studies of patients with diagnosed or suspected sleep disorders, the variance explained by one factor ranged from 40% to 53%.[10,13,41] In two studies concluding one factor within community samples, the variance explained was higher, at 55% and 56%.[9,11] The larger

**Table 6**—Summary of measurement invariance from multigroup confirmatory factor analysis in Sample 2.

| | Configural Invariance Model | | | Metric Invariance Model Loadings Equivalent Across Groups | | | | Scalar Invariance Model Loadings and Intercepts Equivalent | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $X^2$ ($df$ = 40) | RMSEA | CFI | $X^2$ ($df$ = 47) | RMSEA | CFI | Δ CFI | $X^2$ ($df$ = 62) | RMSEA | CFI | Δ CFI |
| Sex: males versus females | 996 | .094 | .984 | 844 | .079 | .986 | .002 | 1173 | .082 | .981 | .005 |
| Age: < 70 versus 70+ | 976 | .093 | .984 | 744 | .074 | .988 | .004 | 953 | .073 | .984 | .004 |
| Race: black versus white | 985 | .097 | .982 | 723 | .076 | .987 | .005 | 1106 | .082 | .980 | .007 |
| Income: < $53,944 versus ≥ $53,944 | 985 | .095 | .983 | 779 | .077 | .987 | .004 | 998 | .076 | .983 | .004 |
| Depression: PHQ-9 score 20+ versus < 20 | 938 | .094 | .982 | 627 | .070 | .988 | .006 | 909 | .074 | .983 | .005 |
| Fatigue: FSS score 36+ versus < 36 | 973 | .094 | .981 | 753 | .076 | .986 | .005 | 934 | .073 | .982 | .004 |
| Comorbidities: sleep apnea versus none | 972 | .094 | .984 | 829 | .079 | .987 | .003 | 1165 | .082 | .981 | .006 |
| Comorbidities: insomnia versus none | 956 | .093 | .983 | 641 | .069 | .989 | .006 | 827 | .068 | .986 | .003 |
| Insomnia: ISI 15+ versus < 15 | 250 | .100 | .980 | 202 | .080 | .986 | .006 | 240 | .074 | .983 | .003 |
| AHI 30+ versus < 30 | 575 | .091 | .983 | 412 | .069 | .989 | .006 | 534 | .068 | .985 | .004 |

Model parameters estimated using weighted least squares means and variance adjusted estimator. Configural invariance fit statistics are from the baseline model with loadings freely estimated. Metric invariance fit statistics are from the model where factor loadings are invariant. Δ CFI compares configural model to metric model. Scalar invariance fit statistics are from the model with loadings and means set to equivalent between groups. Δ CFI compares metric model to scalar model. AHI = apnea-hypopnea index, CFI = comparative fit index, FSS = Fatigue Severity Scale, ISI = Insomnia Severity Index, PHQ-9 = Patient Health Questionnaire-9 Depression Screen, RMSEA = root mean square error of approximation.

variance explained in our study may be due to the increased sample size, as prior studies reporting variance had fewer than 300 participants.

In our study, factor 1 loaded heavily on items 3 (sleepiness in public), 6 (while talking), and 8 (in traffic questions), whereas factor 2 had greater loadings on the more soporific situations of reading, television, resting, and after lunch (items 1, 2, 5, and 7, respectively). A number of studies concluding a two-factor structure have also found a separation based on severity of the items.[12,16,18,19,42] A study of 8,481 undergraduate students in 4 countries,[42] a study within 337 pregnant women,[19] and a study of 843 truck drivers[16] each found 2 factors with eigenvalues > 1, with items 6 and 8 comprising the second factor. Smith et al. evaluated 759 patients attending a sleep disorders clinic with a clinical diagnosis of OSA and directly compared a model with all 8 items to a model with 6 items (excluding items 6 and 8), and found the model with 6 items had improved model fit.[17] These studies argued the ESS has two factors: one that measures sleepiness in socially acceptable situations and another in socially unacceptable situations. Hagell et al. also extracted 2 eigenvalues > 1.0 based on severity yet appropriately concluded that does not contradict unidimensionality.[12] Exploratory factor analysis is based on correlations, and items having similar distributions or that are endorsed by fewer responses will have higher correlations. As the items most commonly factored out in other studies, items 6 and 8 represent those with higher difficulty to endorse and illustrate a drawback of item-level factor analysis, rather than representing another construct. Additionally, many of these prior studies were hindered by small sample sizes, which may result in responses on the ESS items having nonuniform coverage of severity across the severity spectrum. It would be understandable that smaller sample studies may have derived factors more as a matter of item clustering than as a matter of actual multidimensionality, with factor analysis unable to differentiate between the two.

Other studies finding a two- to four-factor structure for the ESS concluded the difference in dimensionality was due to the study sample. Internal reliability of the ESS has been found to be lower among nonclinical samples such as students and healthy community controls, and higher in clinical samples with sleep disorders,[2,7] the 1992 study by Johns assessing the validity of ESS within a clinical sample of patients with OSA and within a group of medical students found a higher internal reliability for the patients as compared to students (Cronbach alpha = .88 versus .73)[7]. Our study also found internal reliability was higher for patients with OSA compared to patients without OSA (ordinal alpha .88 versus .76). This too may be due to a differentiation based on the severity continuum, with fewer students and community members endorsing items 6 and 8, causing less variance with those item-scores. Our study, however, concluded measurement invariance by sleep disorder status, including patients with and without OSA, insomnia, and across AHI levels. Across differing levels of sleep disorders, our study demonstrated equivalence in how patients perceive sleepiness and endorse the items across the severity continuum of sleepiness.

Differences in the internal reliability have also been shown by age group, sex, and race. Two studies of older community members (age 65 years or older) showed adequate internal consistency in the total sample (Cronbach alpha = .70 for men [n = 3,059],[25] and .76 for women [n = 2,968][24]) with a lower corrected item-total correlation for item 8 (in traffic) seen in white women but not black women.[24] The authors concluded this may be because black women have higher levels of sleepiness as compared to white women. Another study found a lack of measurement invariance between the median age cutoff of 40 years and concluded that from middle age on, people were more aware of their sleepiness.[8] In contrast, our study found measurement invariance across all age groups, sex, race, as well as for differing income levels, depression, and fatigue status. Our results indicate sleepiness, as measured by the ESS, has similar psychological meaning across all demographic subgroups. Through establishing measurement invariance, observed mean differences can be attributed to differences in the construct of sleepiness rather than differences in how the subgroups responded to the ESS items, or possible differences in the levels of sleepiness the subgroups experienced.

Our study attempted to clarify the controversy surrounding dimensionality of the ESS through robust statistical methodology and the evaluation of many criteria for establishing dimensionality. Studies concluding multi-dimensionality have cited either the eigenvalues or Cronbach alpha in support of factor structures; however, eigenvalues often result in overfactoring or underfactoring and have been criticized for their subjectivity, and internal consistency is more influenced by the number of items than homogeneity.[43] Model misfit is another commonly cited issue when concluding multidimensionality over unidimensionality, with studies reporting two or three factors based on the recommended cutoff values for model fit indices such as RMSEA and CFI. These cutoffs, however, are subjective and there is no standardized agreement on thresholds. Although our study models also resulted in poorer fit than is commonly reported in factor analyses, typically RMSEA < .08 or < .10, our two-factor model resulted in similar fit statistics. Because unidimensionality is not an absolute but an issue of degree, all of these criteria should be weighed when determining dimensionality.

Given the retrospective nature of our study design, there are some limitations to this research. All patients seen in the sleep disorders center are provided a tablet to complete patient-reported outcomes, including the ESS. To assess potential sampling biases, demographics and general clinical characteristics were compared between patients who completed the ESS versus those who did not. Female sex, fewer comorbidities, and less severe sleep-related comorbidities were found as factors that may have influenced the likelihood of obtaining a set of complete ESS item responses. Reasons for this may be aversion to or inability to use the tablet technology to collect the responses. The current study may thus have potential flaws including some nonresponse biases, and it cannot address measurement differences that may exist for the ESS when deployed in other languages or cultural settings. Additionally, CFA is prone to confirmation bias as it supports the hypothesized research model. We attempted to protect against this bias through utilizing multiple criteria to determine the factor structure, as well as compared results from the two- and

three-factor models. Results of CFA are often only generalizable to the study sample; however, we utilized MGCFA within 10 subgroups to verify the findings. Despite these potential limitations, our study is the largest of its kind to assess the dimensionality of ESS within a clinical patient population.

To the authors' knowledge, this is the first study with a primary focus on confirming the dimensionality of the ESS. Given the vast amount of literature with contradictory findings, our study applies a methodological rigor to demonstrate the unidimensionality of the ESS. We conclude the variability in dimensionality reported in prior studies could be because of the population heterogeneity and severity factoring. We attempted to address this variability within our current study by investigating dimensionality within 10 subgroups of clinical interest. We concluded the ESS measured the same construct across demographic and clinical characteristics. Future research should focus on using IRT methods to derive the ESS's item characteristics, and confirm the individual items do not have differential item-response biases within specific subgroups. In clinical practice, it would be reasonable to ask about the degree of the tendency to doze only for items specifically targeting pathological sleepiness. In the current scoring convention, a global ESS sum score of greater than 10 indicates potential pathological sleepiness, but this convention requires the respondent to answer also the lower severity items, which are superfluous to the clinical task of assessing for pathological sleepiness. It may be clinically sufficient only to ask about ESS items that would report pathological levels of the tendency to doze. To support such an approach psychometrically for asking fewer, but higher-information items, the ESS items would need to be characterized formally for IRT characteristics. Our study findings support the use of these analyses given the ESS is unidimensional in its scale structure.

In conclusion, our study confirmed the unidimensionality of the ESS in a large diverse clinical population. Whether within a clinical population of patients with severe sleep apnea, or a population of young healthy respondents, our study confirmed the ESS will provide an accurate measure of the construct of daytime sleepiness that is interpreted similarly across patients. Results from this study can be used to justify the interpretation of the ESS within clinical populations, as well as support valid comparisons between groups based on the ESS. Now that unidimensionality is confirmed, future studies using IRT models are warranted to further understand the items comprising the ESS and potentially eliminate redundant items for increased efficiency in clinical settings.

## ABBREVIATIONS

AHI, apnea-hypopnea index
CFA, confirmatory factor analysis
CFI, comparative fit index
ESS, Epworth Sleepiness Scale
FSS, Fatigue Severity Scale
IRT, item-response theory
ISI, Insomnia Severity Index
KMO, Kaiser-Meyer-Olkin test
MGCFA, Multi-Group Confirmatory Factor Analysis
PHQ-9, Patient Health Questionnaire-9 depression screening
RLS, restless legs syndrome
RMSEA, root mean square error of approximation

## REFERENCES


1. Ruggles K, Hausman N. Evaluation of excessive daytime sleepiness. *WMJ.* 2003;102(1):21–24.
2. Kendzerska TB, Smith PM, Brignardello-Petersen R, Leung RS, Tomlinson GA. Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review. *Sleep Med Rev.* 2014;18(4):321–331.
3. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep.* 1991;14(6):540–545.
4. Zhang JN, Peng B, Zhao TT, Xiang M, Fu W, Peng Y. Modification of the Epworth Sleepiness Scale in Central China. *Qual Life Res.* 2011;20(10):1721–1726.
5. Bajpai G, Shukla G, Pandey RM, et al. Validation of a modified Hindi version of the Epworth Sleepiness Scale among a North Indian population. *Ann Indian Acad Neurol.* 2016;19(4):499–504.
6. Rosales-Mayor E, Rey de Castro J, Huayanay L, Zagaceta K. Validation and modification of the Epworth Sleepiness Scale in Peruvian population. *Sleep Breath.* 2012;16(1):59–69.
7. Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep.* 1992;15(4):376–381.
8. Martinez D, Breitenbach TC, Lumertz MS, et al. Repeating administration of Epworth Sleepiness Scale is clinically useful. *Sleep Breath.* 2011;15(4):763–773.
9. Neu D, Mairesse O, Hoffmann G, et al. Do 'sleepy' and 'tired' go together? Rasch analysis of the relationships between sleepiness, fatigue and nonrestorative sleep complaints in a nonclinical population sample. *Neuroepidemiology.* 2010;35(1):1–11.
10. Sargento P, Perea V, Ladera V, Lopes P, Oliveira J. The Epworth Sleepiness Scale in Portuguese adults: from classical measurement theory to Rasch model analysis. *Sleep Breath.* 2015;19(2):693–701.
11. Pilcher JJ, Pury CL, Muth ER. Assessing subjective daytime sleepiness: an internal state versus behavior approach. *Behav Med.* 2003;29(2):60–67.
12. Hagell P, Broman JE. Measurement properties and hierarchical item structure of the Epworth Sleepiness Scale in Parkinson's disease. *J Sleep Res.* 2007;16(1):102–109.
13. Izci B, Ardic S, Firat H, Sahin A, Altinors M, Karacan I. Reliability and validity studies of the Turkish version of the Epworth Sleepiness Scale. *Sleep Breath.* 2008;12(2):161–168.
14. Riachy M, Juvelikian G, Sleilaty G, Bazarbachi T, Khayat G, Mouradides C. [Validation of the Arabic Version of the Epworth Sleepiness Scale: Multicentre study]. *Rev Mal Respir.* 2012;29(5):697–704.
15. Mills RJ, Koufali M, Sharma A, Tennant A, Young CA. Is the Epworth sleepiness scale suitable for use in stroke? *Top Stroke Rehabil.* 2013;20(6):493–499.
16. Heaton K, Anderson D. A psychometric analysis of the Epworth Sleepiness Scale. *J Nurs Meas.* 2007;15(3):177–188.
17. Smith SS, Oei TP, 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(7):739–744.
18. Violani C, Lucidi F, Robusto E, Devoto A, Zucconi M, Ferini Strambi L. The assessment of daytime sleep propensity: a comparison between the Epworth Sleepiness Scale and a newly developed Resistance to Sleepiness Scale. *Clin Neurophysiol.* 2003;114(6):1027–1033.
19. Baumgartel KL, Terhorst L, Conley YP, Roberts JM. Psychometric evaluation of the Epworth sleepiness scale in an obstetric population. *Sleep Med.* 2013;14(1):116–121.
20. Olaithe M, Skinner TC, Clarke J, Eastwood P, Bucks RS. Can we get more from the Epworth Sleepiness Scale (ESS) than just a single score? A confirmatory factor analysis of the ESS. *Sleep Breath.* 2013;17(2):763–769.

21. Sadeghniiat Haghighi K, Montazeri A, Khajeh Mehrizi A, et al. The Epworth Sleepiness Scale: translation and validation study of the Iranian version. *Sleep Breath.* 2013;17(1):419–426.
22. Nguyen AT, Baltzan MA, Small D, Wolkove N, Guillon S, Palayew M. Clinical reproducibility of the Epworth Sleepiness Scale. *J Clin Sleep Med.* 2006;2(2):170–174.
23. Peng LL, Li JR, Sun JJ, et al. [Reliability and validity of the simplified Chinese version of Epworth sleepiness scale]. *Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi.* 2011;46(1):44–49.
24. Beaudreau SA, Spira AP, Stewart A, et al. Validation of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Scale in older black and white women. *Sleep Med.* 2012;13(1):36–42.
25. Spira AP, Beaudreau SA, Stone KL, 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(4):433–439.
26. Katzan I, Speck M, Dopler C, et al. The Knowledge Program: an innovative, comprehensive electronic data capture system and warehouse. *AMIA Annu Symp Proc.* 2011;2011:683–692.
27. Bastien CH, Vallieres A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med.* 2001;2(4):297–307.
28. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16(9):606–613.
29. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch Neurol.* 1989;46(10):1121–1123.
30. Kaiser H. An index of factorial simplicity. *Psychometrika.* 1974;39(1):31–36.
31. Costello AB, Osborne JW. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation.* 2005;10(7):173–178.
32. Kline R. *Principles and Practice of Structural Equation Modeling.* 3rd ed. New York: Guilford Press; 2011.
33. Gadermann AM, Guhn M, Zumbo BD. Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation.* 2012;17(3):1–13.
34. Hirschfeld G, von Brachel R. Multiple-Group confirmatory factor analysis in R- A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation.* 2014;19(7):1–12.
35. Xu H, Tracey TJG. Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *The European Journal of Counselling Psychology.* 2017;6(1).
36. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling.* 2002;9:233–255.
37. R Core Team. R Foundation for Statistical Computing, Vienna, Austria. The R Project for Statistical Computing. http://www.R-project.org/. Published 2013. Accessed June 12, 2018.
38. Revelle W. psych: Procedures for Personality and Psychological Research, Northwestern University. Evanston, Illinois, USA, Version 1.4.8. http://CRAN.R-project.org/package=psych. Published 2014. Accessed June 12, 2018.
39. Raiche G. nFactors: an R package for parallel analysis and non graphical solutions to the Cattell scree test. R package version 2.3.3. http://CRAN.R-project.org/package=nFactors. Published 2010. Accessed June 12, 2018.
40. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. *J Stat Softw.* 2012;48(2):1–36.
41. 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(4):293–294.
42. Gelaye B, Lohsoonthorn V, Lertmeharit S, et al. Construct validity and factor structure of the pittsburgh sleep quality index and epworth sleepiness scale in a multi-national study of African, South East Asian and South American college students. *PLoS One.* 2014;9(12):e116383.
43. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods.* 1999;4(3):272–299.


## SUBMISSION & CORRESPONDENCE INFORMATION

**Submitted for publication November 7, 2017**
**Submitted in final revised form March 27, 2018**
**Accepted for publication March 29, 2018**
Address correspondence to: Brittany R. Lapin, PhD, MPH, 9500 Euclid Avenue, JJ3-603, Cleveland, Ohio 44195; Tel: (216) 445-5729; Email: LapinB@ccf.org; ORCID: 0000-0002-4314-2282

## DISCLOSURE STATEMENT

All authors have seen and approved the manuscript. The authors report no conflicts of interest.