


Low test–retest reliability of the Epworth Sleepiness Scale within a substantial short time frame

Renáta Rozgonyi¹ | István Dombi² | József Janszky^{1,3} | Norbert Kovács^{1,3} | Béla Faludi¹ 

¹Department of Neurology, Medical School, Clinical Center, University of Pécs, Pécs, Hungary

²Pneumology Department, The Emergency County Hospital of Miercurea Ciuc-Miercurea Ciuc, Romania

³MTA-PTE Clinical Neuroscience MR Research Group, Pécs, Hungary

Correspondence

Béla Faludi, Department of Neurology, Medical School, Clinical Center, University of Pécs, Pécs, 7623, Hungary.
Email: bela.faludi@gmail.com

Abstract

The Epworth Sleepiness Scale (ESS) is a widely used, validated questionnaire for effectively examining patients' sleepiness in a range of different situations. Test–retest reliability is an important aspect of a questionnaire, which, according to only a few studies, was found to be low in the case of the ESS. All these studies applied long intervals between the tests, thereby increasing the possibility of fundamental change in circumstances, which in turn affect the reliability of the test. The aim of the present study was to investigate the test–retest reliability of the ESS in a short time frame to provide stability of the test circumstances. We also compared the originally used and current accepted statistical methods of test–retest evaluation. We examined 100 unselected patients consecutively referred to the sleep laboratory with the ESS questionnaire, using a test–retest paradigm with an interval of 1 h between two ESS tests. The Lin's concordance coefficient was found to be low, whereas the Pearson's correlation revealed good reliability. Our result provides evidence on the poor test–retest reliability of the ESS, despite the examination protocol excluding changes in test circumstances.

KEYWORDS

Epworth Sleepiness Scale, sleepiness, test–retest reliability

1 | INTRODUCTION

Excessive daytime sleepiness (EDS) has a significant influence on everyday life (Hays, 1996; Newman, 2000; Ruggles, 2003; Young, 2004). Many disorders lead to sleepiness and often the underlying condition is serious and life threatening (Chasens, 2009; Ruggles, 2003). The consequences of sleepiness include work-related accidents, motor vehicle accidents and the deterioration of interpersonal and family relationships, etc. (MacLean, 2003). There is a close association

between excessive daytime sleepiness and increased risk of mortality (Hays, 1996; Newman, 2000). The prevalence of sleepiness is higher in some major cardiovascular and cerebrovascular disorders (stroke, acute coronary syndrome, cardiac arrhythmias, etc.) and hypertension. These disorders may be associated with obstructive sleep apnea as a background in which pathology leads to sleepiness.

In addition to examination of the pathological background, the determination and quantification of sleepiness is an important aspect of patient care and follow-up.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Journal of Sleep Research* published by John Wiley & Sons Ltd on behalf of European Sleep Research Society.

There are at least two different methods of determination and characterization of sleepiness in clinical practice: the questionnaire-based (e.g., the Epworth Sleepiness Scale [ESS]) and polysomnography-based (maintenance of wakefulness test and multiple sleep latency test) methods.

The ESS is the most prevalent and validated sleepiness questionnaire and was developed by Johns (Johns, 1992). The scale contains eight topics that represent different day-to-day situations. The likelihood of patients falling asleep in different situations must be effectively evaluated. Patient responses are totalized and a higher score indicates a higher level of sleepiness. Several examinations substantiate the usefulness of the scale in a typical daily routine.

Since 1991, only a few investigations have sought to examine the reliability of the ESS within different patient populations, in different time frames and with different methods.

Stability and similarity of the circumstances during the test and retest periods may increase the reliability of the applied questionnaire. All of the former examinations applied a relatively long inter-investigation time (weeks to months) (Aloe, 1997; Bourke, 2004; Johns, 1991; Knutson, 2006; Lee, 2020; Nguyen, 2002). During these intervals, there are possibilities of rudimentary changes of the test population, which may negatively affect the test results. Recently, Lee and coworkers' (Lee, 2020) results suggested the Epworth Sleepiness Scale is not useful in clinical settings, in the evaluation of a therapeutic intervention or to use (prioritize) during the access to a service.

The primary goal of the present study is to determine the reliability of the ESS questionnaire by a test-retest method within a short time frame to exclude subjective or objective changes in patients' sleep habits, concomitant medications and diseases.

2 | METHODS

One hundred unselected participants were enrolled into the study. Consequently, the patients referred to the Sleep Laboratory are likely to have suffered from a variety of sleep complaints (insomnia, snoring, sleep disordered breathing, restless leg syndrome, narcolepsy, etc.). We performed the tests in the absence of known diagnosis of the individual patient.

The validated Hungarian translation of the standard eight-item ESS test was utilized (validated by the study group, unpublished, but presented (Kovács, 2013)). The examination was based on test-retest methodology, including short time differences between the two test sessions. All the 100 patients completed the questionnaire by themselves (without supervision) during both sessions in the morning period, between 9:00 and 11:00 AM. According to the original version of the ESS, the participant received written instructions about the test.

Notably, a "blinded" protocol was applied. The patients were informed regarding the primary objective of the study (examination of sleepiness by questionnaire), but were not aware of the repetition of

the same questionnaire following a short period of time (known as a retest condition). The participants were asked only to complete all the questions in the test package within the prescribed sequence.

The design of the experiment was intended to exclude (or minimize) the "carry back" effect of the answers in the first test session. All participants completed the same ESS questionnaire twice with approximately a 1-h interval. Participants were asked to complete non-specific tests (questions independent of the goal and topic of the present study; e.g., instruments measuring quality of life, depression and anxiety) following the first test session to divert attention and minimize the "carry back" effect. Following these "non-specific tests", the participants once again completed the ESS questionnaire.

We chose 1-h intervals between the two assessments because this period is long enough to reduce the "carry-on" effect yet short enough to assess the test-retest reliability. With the possibility to have a rest between the tests we minimized the probability of decreased attention on completion of the questionnaires. Distinctively, subjects were asked to answer the items of the ESS based on their experiences during the prior period, and the responses of the first and second ESS assessments should be based on the same time period and reflect the same experiences. Therefore, we could eliminate variance of sleep quality and daytime sleepiness between the two ESS sessions.

To evaluate the test and retest reliability of the two ESS score datasets, Lin's concordance coefficient and Pearson's correlation were calculated using the SPSS v.22 statistical software package (SPSS, IBM Inc., USA). The two statistical methods were compared on the same ESS datasets.

For Lin's concordance coefficient, McBride (McBride, 2005) proposed the strength-of-agreement criteria, which is poor when and if the value is less than 0.9, moderate between 0.9 and 0.95, substantial if the value is more than 0.95 and almost perfect when and if the value is over 0.99.

The research was reviewed and approved by a local ethics committee (5332/2014, Regional and Institutional Research-Ethical Committee, University of Pécs, Hungary).

3 | RESULTS

Altogether 100 participants were enrolled into the study from an unselected patient population referred to the sleep laboratory of the Department of Neurology, mostly by general practitioners, including other hospital departments. Due to the original concept of the ESS (disease-independent sleepiness determination) we did not define subgroups according to the results of the sleep examinations. The demographic characteristics of the participants include the following. Of the 100 participants, 63 were men and 37 were women. The unbalanced ratio was due to the referral of the patients. The ages of the male patients were between 22 and 77 years (mean, 49.21 years; standard deviation [SD], 13.35) years. The age variation in the female patients was 35 to 79 years of age (mean, 55.98 years; SD, 10.63) years.

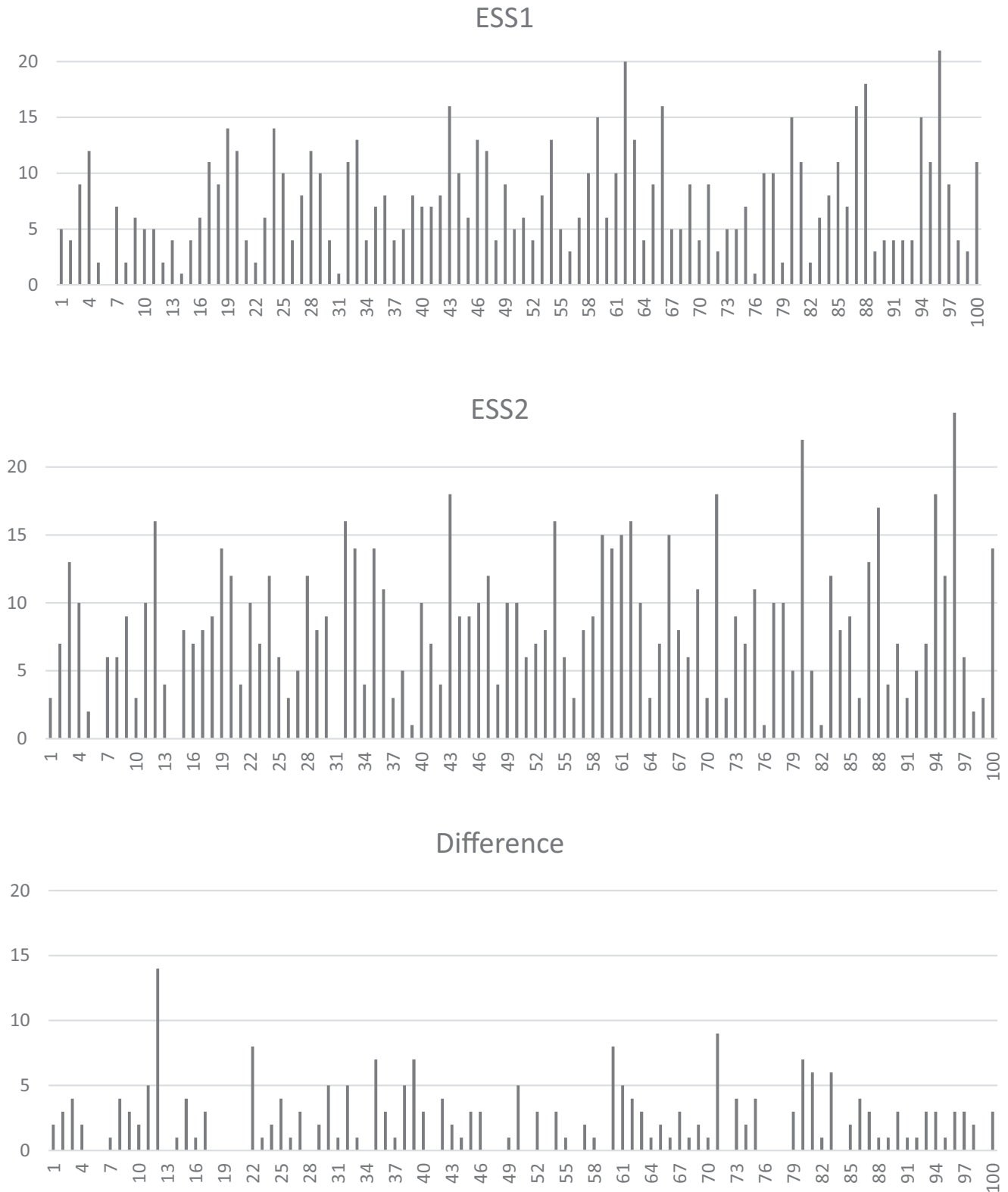


FIGURE 1 The values of test and retest of the ESS of the 100 patients. ESS1: discrete values of the first ESS test of the 100 patients. ESS2: discrete values of the second ESS test (retest) of the 100 patients. Difference: difference between ESS1 and ESS2

The descriptive statistical parameters of the test session include the following: mean, 7.62; SD, 4.48; median, 7 for all participants (7.24, 4.51 and 6 for men, 8.27, 4.29 and 8 for women). For the

retest, the mean was 8.49, the SD was 5.00 and the median 8 for all participants (8.40, 4.77 and 8 for men, 8.65, 5.30 and 9 for women). The differences in the ESS scores of the test and retest sessions

were calculated with consideration of all the participants. The mean of the difference in the test and retest results was 2.44, the SD was 2.40 and the median was 2 (2.46, 2.65 and 2 for men, and 2.41, 1.88 and 2 for women). Decreases in the second ESS values were seen in 34 cases, increased ESS values were found in 43 cases, and they remained unchanged in 23 cases. The differences in ESS values of the two test sessions were higher than 2 in 42 cases (men, 26; women, 18). See Table S1 for further patient characteristics!

Figure 1 shows the discrete values of the test and retest sessions of the 100 participants.

Figure 2 demonstrates the Bland-Altman plot based on the ESS scores of the test and retest sessions. The Bland-Altman plot (Bland, 1998) is a visual method using a graph to compare two measurements, in which the differences between the two measurements are plotted against the averages of the two measures.

The test-retest reliability was examined by Lin's concordance coefficient. The first (test) and second (retest) ESS scores of the 100 participants were used for calculation. The means and variances of the ESS test were 7.62 and 20.09, respectively. The means and variances of the retest ESS scores were 8.49 and 25.02, respectively. The Rc value for Lin's concordance coefficient was 0.748 (the correlation is poor if the value is less than 0.9). Additionally, we calculated Pearson's correlation and report the value was 0.76 (there is good correlation if the value is between 0.5 and 1). A statistician has verified this analysis.

4 | DISCUSSION

Reliability is a basic requirement for every questionnaire and must be tested in various circumstances. To determine the effective

reliability, different domains can be examined, as well as internal consistency (Cronbach, 1951) and test-retest reliability (Guttman, 1944). The internal consistency reflects the interrelatedness of the items, whereas test-retest reliability focuses on evaluating the proportion of the total variance in the measurements acquired during two independent tests (consistency of a measure across time). The appropriate statistical model for the internal consistency is the Cronbach's alpha (Cronbach, 1951) and for the test-retest reliability is the intraclass correlation coefficient (Bartko, 1966; Feldt, 1965) or Lin's concordance coefficient (Lin, 1989).

Former examinations proved the usefulness of the ESS but revealed poor reliability in some circumstances. The test-retest reliability examination provided controversial results.

The main goal of the present study was to determine the test-retest reliability of the ESS questionnaire utilizing an extremely short time difference (1 h) between the two test sessions, including a variety of different statistical methods (Lin's concordance coefficient and Pearson's correlation). Using Lin's concordance coefficient, we demonstrated a low Rc value, representing poor reliability in our paradigm in a relatively large sample.

As far as the authors are aware, our work is the first of this type of test-retest validation study. In particular, the additional novelty of our method is the "blinded" protocol (the participants were not aware of the repetition of the ESS questionnaire) and the non-different questions between the two ESS tests. Both of them are applied to minimize the "carry back" effect on the second ESS test.

Former studies using test-retest methods (Bourke, 2004; Johns, 1991; Knutson, 2006; Lee, 2020; Nguyen, 2002) applied different intervals between the two test sessions. The interval varied from 70 days to 6 months. The drawback of these protocols is the potential changes in circumstances between the two test periods,

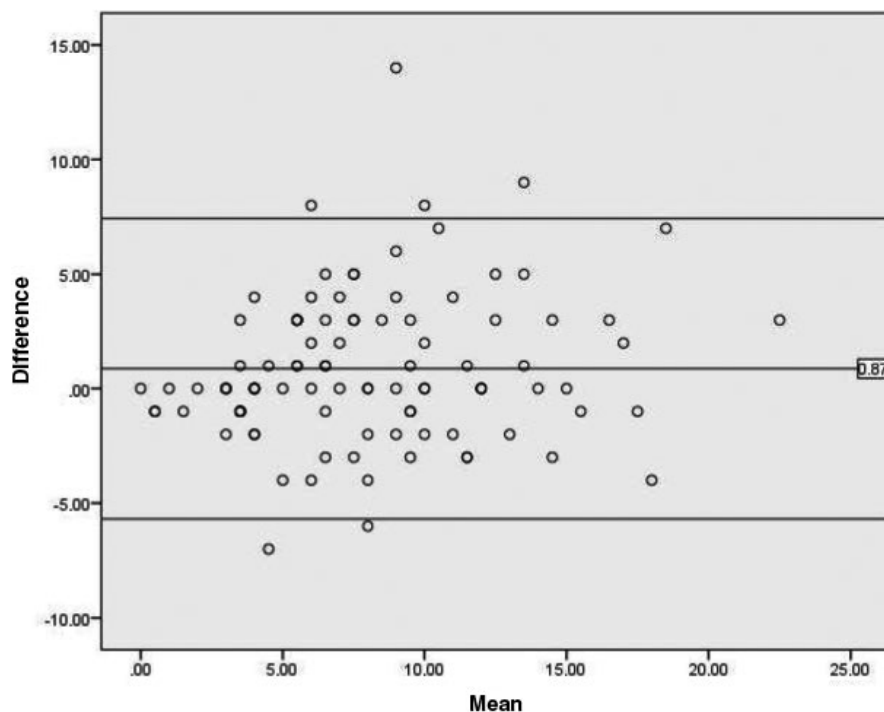


FIGURE 2 Bland-Altman plot of the two ESS tests. Horizontal lines represent the mean difference (middle line) and the limits of agreement (upper and lower lines), which are defined as the mean difference plus and minus 1.96 times the standard deviation of the differences. The values outside the ± 1.96 standard deviation represent the extreme in differences

which could affect the result: changes in the patient's comorbidities, sleep habits and consequential sleepiness, drowsiness, prescription medication, etc.

The results of these examinations revealed primarily moderate evidence in relation to the test-retest reliability of the ESS questionnaire. Only one study demonstrated good reliability with the given methodology (Johns, 1991). Johns tested 87 healthy medical students and repeated the test 5 months later. Correlations of the two score sets were then calculated and a high correlation was found. In contrast, other examinations failed to prove these ideal results over subsequent decades (Bourke, 2004; Knutson, 2006; Nguyen, 2006; Lee, 2020). Notably, the results varied from poor to fair.

On the basis of the applied methodology, it can be supposed that the poor test-retest reliability is the consequence of fundamental changes in conditions during the two test sessions.

The applied protocol (repetition of the ESS questionnaire after 1 h and completion of an independent questionnaire between the two ESS sessions) probably did not influence negatively the patients' attention and, consequently, the reliability of the study. In our experiment, the completion of the different subsidiary questionnaires between the two ESS sessions did not require much time and did not require a long (1 h) period of sustained attention, leaving an opportunity to have a rest. This protocol can provide more accurate control over the experimental session than the former delayed test-retest situations where the circumstances of the two sessions are more variable and too unstable. The similarity and stability of circumstances in our method are underlined by the result of Pearson's correlation, which revealed good reliability, whereas there is low or moderate reliability in the majority of the delayed non-stationary types of studies (Aloe, 1997; Bourke, 2004; Knutson, 2006; Nguyen, 2002).

Another possible explanation of the poor reproducibility of the ESS is the highly subjective interpretation of the items as reported by the patients. One study found fundamental differences between the guided (administration of ESS by nurse) and unguided (without assistance) application of ESS (Ugur, 2011). Ugur and coworkers concluded the administration method can increase the reliability and sensitivity of the test. We applied non-guided competition of the questionnaires.

Compared to the statistical methods used to characterize the test-retest reliability of the ESS, we assumed the differences of the analyses may have contributed to the heterogeneous result of the reproducibility. The applied statistical method varied from the correlation analysis (Johns, 1991) to the intraclass correlation coefficient (Lee, 2020) and Lin's concordance coefficient (in our study). To compare our results with the first test-retest examination of the ESS by Johns (Johns, 1991), we calculated Pearson's correlation coefficient for our datasets, which showed good reliability. On the basis of our experiment, we suppose the good reliability determined by Johns is likely to be the consequences of the applied statistical method, which was the accepted method for the test-retest reliability at that specific time. By contrast with the former investigations

using Pearson's correlation, Lee (2020) applied the intraclass correlation coefficient, which is one of the approved statistical methods in the test-retest paradigm, and proved the inadequate reliability of the Epworth Sleepiness Scale in a test-retest paradigm.

There are several limitations of the present investigation. The number of the participants is relatively low, therefore subgroup analysis cannot be done. A further limitation is that the study population was mixed (patients with different sleep disorders), but this aspect is close to the original concept of the ESS (determination of cause-independent sleepiness).

5 | CONCLUSIONS

In this study, we provided additional data in reference to the issue of reliability of the ESS test. Excluding nearly all modifying factors with the applied protocol, we demonstrated the low test-retest reliability of the ESS within a short time frame.

These results question the usefulness of the ESS, at the very least, in the follow-up procedures for patients and suggest it must be interpreted cautiously in the evaluation of excessive daytime sleepiness.

Our results and former results based on test-retest reliability of the ESS underline the importance of a reliable and relatively simple method to detect sleepiness.

ACKNOWLEDGEMENTS

Our study was supported by the NKFIH SNN125143 and the EFOP-3.6.1-16-2016-00004.

CONFLICTS OF INTEREST

All the authors have disclosed all financial and non-financial support for our work and other potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conceptualization: BF and NK. Study design: BF and NK. Data collection: RR and BF. Data analysis: RR and BF. Interpretation: RR, ID, JJ, NK and BF. Writing – review and editing: RR, JJ, NK and BF.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Béla Faludi  <https://orcid.org/0000-0003-2126-4243>

REFERENCES

- Alôe, F., Pedroso, A., & Tavares, S. M. (1997). Epworth Sleepiness Scale outcome in 616 Brazilian medical students. *Arquivos De Neuropsiquiatria*, 55(2), 220-226. <https://doi.org/10.1590/s0004-282x1997000200009>
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3-11. <https://doi.org/10.2466/pr0.1966.19.1.3>

- Bland, J. M., & Altman, D. G. (1998). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160. <https://doi.org/10.1177/096228029900800204>
- Bourke, S. C., McColl, E., Shaw, P. J., & Gibson, G. J. (2004). Validation of quality of life instruments in ALS. *Amyotrophic Lateral Sclerosis Other Motor Neuron Disorders*, 5(1), 55–60. <https://doi.org/10.1080/14660820310016066>
- Chasens, E. R., Sereika, S. M., & Burke, L. E. (2009). Daytime sleepiness and functional outcomes in older adults with diabetes. *Diabetes Education*, 35(3), 455–464. <https://doi.org/10.1177/0145721709333857>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>
- Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30, 357–370. <https://doi.org/10.1007/BF02289499>
- Guttman, L. (1944). A basis for analyzing test-retest reliability. *Pszichometrika*, 4(10), 255–282.
- Hays, J. C., Blazer, D. G., & Foley, D. J. (1996). Risk of napping: Excessive daytime sleepiness and mortality in an older community population. *Journal of American Geriatric Society*, 44(6), 693–698. <https://doi.org/10.1111/j.1532-5415.1996.tb01834.x>
- Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep*, 14(6), 540–545. <https://doi.org/10.1093/sleep/14.6.540>
- Johns, M. W. (1992). Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep*, 15(4), 376–381. <https://doi.org/10.1093/sleep/15.4.376>
- Knutson, K. L., Rathouz, P. J., Yan, L. L., Liu, K., & Lauderdale, D. S. (2006). Stability of the Pittsburgh Sleep Quality Index and the Epworth Sleepiness Questionnaires over a one year period in young and middle-aged adults: The CARDIA study. *Sleep*, 29(11), 1503–1506. <https://doi.org/10.1093/sleep/29.11.1503>
- Kovács, N., Pál, E., Janszky, J., Bosnyák, E., Ács, P., Aschermann, Z., Faludi, B. (2013). Parkinson's disease Sleep Scale-2 is more specific for PD than the Epworth Sleep Scale. *Journal of the Neurological Sciences*, 333, e139. <http://dx.doi.org/10.1016/j.jns.2013.07.462>
- Lee, J. L., Chung, Y., Waters, E., & Vedam, H. (2020). The Epworth sleepiness scale: Reliably unreliable in a sleep clinic population. *Journal of Sleep Research*, 29, 1–5. <https://doi.org/10.1111/jsr.13019>
- Lin, L.-I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268.
- MacLean, A. W., Davies, D. R., & Thiele, K. (2003). The hazards and prevention of driving while sleepy. *Sleep Medicine Review*, 7(6), 507–521. [https://doi.org/10.1016/s1087-0792\(03\)90004-9](https://doi.org/10.1016/s1087-0792(03)90004-9)
- McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient* (pp. 1–10). Hamilton, New Zealand.
- Newman, A. B., Spiekerman, C. F., Enright, P., Lefkowitz, D., Manolio, T., Reynolds, C. F., Robbins, J. (2000). Daytime sleepiness predicts mortality and cardiovascular disease in older adults. The Cardiovascular Health Study Research Group. *Journal of American Geriatric Society*, 48(2), 115–123. <https://doi.org/10.1111/j.1532-5415.2000.tb03901.x>
- Nguyen, A. T., Baltzan, M. A., Small, D., Wolkove, N., Guillon, S., & Palayew, M. (2002). Clinical reproducibility of the Epworth Sleepiness Scale. *Journal of Clinical Sleep Medicine*, 2(2), 170–174.
- Ruggles, K., & Hausman, N. (2003). Evaluation of excessive daytime sleepiness. *WMJ*, 102(1), 21–24.
- Ugur, K. S., Ark, N., Kurtaran, H., Ozol, D., Kurt, K., & Mutlu, C. (2011). Comparison of scores of application methods of the Epworth Sleepiness Scale: Self administered or nurse administered. *Journal of Otorhinolaryngological Related Specialities*, 73(5), 249–252. <https://doi.org/10.1159/000330383>
- Young, T. B. (2004). Epidemiology of daytime sleepiness: Definitions, symptomatology, and prevalence. *Journal of Clinical Psychiatry* 65(Suppl 16), 12–16.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Rozgonyi R, Dombi I, Janszky J, Kovács N, Faludi B. Low test-retest reliability of the Epworth Sleepiness Scale within a substantial short time frame. *J Sleep Res.* 2021;30:e13277. <https://doi.org/10.1111/jsr.13277>