Sleep Biol. Rhythms (2016) 14:257–260
DOI 10.1007/s41105-016-0048-8

ORIGINAL ARTICLE

Comparing two versions of the Karolinska Sleepiness Scale (KSS)

Anna Åkerstedt Miley[3] · **Göran Kecklund**[1,2] · **Torbjörn Åkerstedt**[2,1]

Received: 4 September 2015 / Accepted: 4 January 2016 / Published online: 1 February 2016

Abstract The Karolinska Sleepiness Scale (KSS) is frequently used to study sleepiness in various contexts. However, it exists in two versions, one with labels on every other step (version A), and one with labels on every step (version B) on the 9-point scale. To date, there are no studies examining whether these versions can be used interchangeably. The two versions were here compared in a 24 hr wakefulness study of 12 adults. KSS ratings were obtained every hour, alternating version A and B. Results indicated that the two versions are highly correlated, do not have different response distributions on labeled and unlabeled steps, and that the distributions across all steps have a high level of correspondence (Kappa = 0.73). It was concluded that the two versions are quite similar.

Keywords Sleep deprivation · Method · Drowsiness · Ratings

Introduction

Sleepiness affects a large part of the population and is usually increased in relation to shift work or disturbed sleep [1]. Subjective measures are a quick and cost-effective way of estimating sleepiness. The Karolinska Sleepiness Scale (KSS) [2] has been widely used in studies of shiftwork [3], sleep deprivation [4], and driving [5]. It has been found to correlate well with polysomnographical measurements (PSG), like alpha (8–12 Hz) and theta (4–8 Hz) activity in the EEG [2], as well as with performance-based measures [6], indicating that worsening of performance is associated with increased KSS values. A recent review summarizes a number of studies of KSS in different laboratory and field settings [7].

The scale exists in two versions. The original scale (A, see methods) had labels on every second (i.e. uneven) step (1, 3, 5, 7, and 9). Baulk et al. [8] added labels to the remaining four (even) steps. The two versions have not been compared with respect to distribution of ratings, however. With increasing use of the scale we have received requests for information on the comparability of the scales. We have, therefore, selected for analysis a previously collected material [9] from a 24 h sleep loss study, which seems to be the only available data set where both versions are used. The sleep deprivation setting has the advantage of making possible comparisons at many different levels of sleepiness.

The purpose of this brief report is to compare the distribution of ratings across the two versions and to analyze the covariation between the two scales over 24 h of wakefulness.

Method

Participants and design

Twelve participants [50 % female; mean age = 37.5 years; (SD = 7.5)] were kept awake in the laboratory from 07:00 (day 1) until 09:00 (day 2). Each participant underwent a 50 min test battery (reaction time, Stroop color

✉ Torbjörn Åkerstedt
torbjorn.akerstedt@ki.se

[1] Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden

[2] Stress Research Institute, Stockholm University, Stockholm, Sweden

[3] Karolinska Hospital, Stockholm, Sweden

word test, a tracking test and additions) every 3 h, starting at 08:00 on day 1 [9].

Procedures

Subjective sleepiness was assessed with the KSS every hour, starting at 08:00 (day 1). To compare the two versions of the KSS, the administration of the versions alternated every hour, starting with version A (for all participants). This approach was chosen since rating sleepiness on both scale at the same time would very likely have led to a strong carry-over from one scale to the other. As such, each participant rated his/her sleepiness using both versions for a total of 12 ratings for version A and 12 for version B.

The KSS spans 9 levels (1 = extremely alert, 2 = very alert, 3 = alert, 4 = rather alert, 5 = neither alert nor sleepy, 6 = some signs of sleepiness, 7 = sleepy, but no effort to keep awake, 8 = sleepy, some effort to keep awake, 9 = very sleepy, great effort keeping awake, fighting sleep). The original scale (A) included labels on every second step (1, 3, 5, 7, and 9), that is, the uneven numbers. Version B [8] had labels added to the remaining (even = 2, 4, 6, and 8) steps to smooth the scale. The instruction asks the user to circle the number that represents the perceived level of sleepiness during the immediately preceding 5 min.

Statistical analyses and ethics approval

The main question concerned whether the two versions would differ in the distribution of choices of even and odd values. For this purpose the McNemar test [10] was applied to compare the distribution of frequency of use of even and odd values for the two versions. There was also a possibility that the distribution across all the individual values 1–9 would be affected by the labeling. Therefore, we also applied Cohen's unweighted Kappa [11] across all values 1–9 between versions A and B. This tests the absolute correspondence between the ratings for each level of the two versions. However, since the KSS scale is ordered, a more appropriate method is to use Cohen's weighted Kappa [12] which takes into account how close the rating in one version is to the rating in the other version. To this was also applied Bowker's test of symmetry [13] for significance testing. The study was approved by the Regional Ethical Committee of the Stockholm region.

Results

Figure 1 shows the distribution of ratings for the two versions. Version A had 88 ratings for labeled (odd) KSS values (i.e. 1, 3, 5, 7, and 9) and 56 ratings for non-labeled (even) values (i.e. 2, 4, 6, and 8). Version B had 77 ratings for values 1, 3, 5, 7, and 9 and 66 ratings for the ratings 2, 4, 6, and 8. A McNemar test was used to compare the change in odd and even values between the two scales (Table 1). This yielded a $\text{Chi}^2 = 0.06$ ($df = 1$, $p > 0.05$), that is, the distributions for odd–even values in the two versions did not differ significantly.

Also the distribution of all the ratings of 2, 3, 4, 5, 6, 7, 8, and 9 was compared between the two versions (the value "1" was removed since there was only one such rating). This yielded an unweighted Kappa value of 0.23 with a standard error of 0.03, $p < 0.001$. That is, there was a significant difference between the two versions.

However, taking the ordered scale into account and using the weighted Kappa analysis the symmetry statistic (S) was = 20.7 ($p = 0.84$ for $df = 28$), that is, the two scales did not differ significantly. The weighted Kappa values was $K = 0.73$, with a 95 % confidence interval of 0.65–0.83. Usually a $K = 0.61$–80 is considered "good" agreement (and $K > 0.80$ = very good) [11, 12]. The number of ratings for each level of the two scales is presented in Fig. 1.

The values from version A and version B were also correlated for each individual across time points. Thus, the

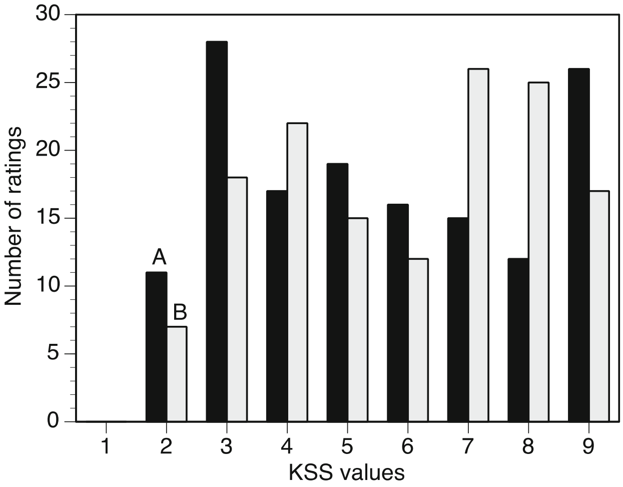

Fig. 1 Distribution of ratings of each scale value [1–9] for versions A and B

Table 1 Cross tabulation of the number of ratings for the uneven and even values of versions A and B

	Version B		
Version A	Uneven	Even	Row total
Uneven	39	31	70
Even	34	21	55
Column total	73	52	125

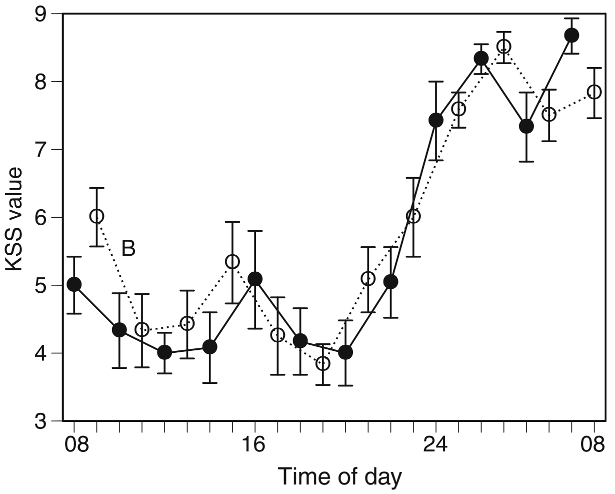

Fig. 2 Mean ± se KSS values from two-way ANOVA for each point in time for versions A and B during 24 h of continuous wakefulness. Adapted from [9]

value of version A at 08:00 was paired with the value at 09:00 of version B, etc. The average correlation between version A and version B was $r = 0.66$ ($s.e = 0.04$, $p < 0.0001$). Furthermore, the average variance explained was $r^2 = 0.45$ ($s.e = 0.05$, $p < 0.0001$). The regression coefficient was $ß = 0.69 \pm 0.03$ ($p < 0.0001$) with a mean intercept of 1.61 ($s.e = 0.31$).

For better understanding of the results the variation across time is illustrated in Fig. 2. The ANOVA results were presented in a previous paper [9] comparing mean levels of the two versions A and B. The effect of time was highly significant ($p < 0.001$), but, the difference between the versions was not.

Discussion

The testing of the correspondence between the two scales did not show any significant difference with respect to the labeled and unlabeled parts of the scales. When all scale values were compared the exact comparison showed a significant difference whereas the weighted one did not. The Kappa value was in the range of "good" correspondence. In addition, the two versions were highly correlated across time.

The results do not seem to indicate that a step being labeled or not will determine rating behavior. This is somewhat unexpected, although the participants using version A are always encouraged to use also the unlabeled steps when carrying out their ratings. Inspection of Fig. 1 verifies this finding; the steps without labels in version A were not consistently less used than the corresponding labeled steps in version B. The similarity of the distribution on the two scales is supported by the non-significant weighted analysis across all 9 steps on the scale and the high Kappa values. Still, this analysis largely disregards discrepancies that are small through the weighting procedure across the scale. Close values are considered "almost" similar while distant values are considered very different.

Not viewing the scales as ordinal and computing absolute differences regardless of closeness showed a significant difference across the 1–9 scale. Inspection of Fig. 1 shows that there was more frequent use of steps 7 and 8 in version B and a less frequent use of step 9 and 3. These findings suggest that the labeling of step 8 may have made it easier not to rate 9 when very sleepy. The two scales were used with 1 h intervals and it seems unlikely that the context of continuous wakefulness should have had differing effects on the two versions

The rather high mean correlation between the two versions across time suggests that they measure the same state. It should be emphasized that this not a measure of reliability as that should be based on simultaneous administration of the two versions. As suggested above, simultaneous administration of the two scales would probably have presented spuriously high correlation because most participants would have been influenced by their previous rating. The present mean correlation of $r = 0.66$ should probably be seen as quite high since sleepiness is a very volatile state and may change within a few minutes depending on the level of stimulation. Walking, standing, being alone, or social interaction, all influence sleepiness ratings [14].

The present study made use of already collected data, which means the study was not explicitly designed for its purpose. Still, the study gives a reasonable impression of the similarity of the two versions, also when considering the pattern across time [9]. One weakness is the modest number of participants, which may have affected the analysis of the distribution of ratings. The fact that the ratings started with version A and then were followed by alternations of B and A did not create an order effect since all version B ratings were followed by a version A rating except for the first one.

In conclusion, the present study has shown that the distribution of the two versions did not differ with respect to labeled and unlabeled steps, nor when closeness of ratings were allowed for in the analysis across the whole scale. The study also showed a reasonable correlation between the two versions across time. We suggest that the present results make it possible to compare results from studies where the two scales have been used.

Acknowledgments This study was supported by the Swedish Research Council for Health, Working Life and Welfare.

Compliance with ethical standards

Conflict of interest None of the authors have declared any conflict of interest.

References

1. Ohayon MM. Determining the level of sleepiness in the American population and its correlates. J Psychiatry Res. 2012;46(4):422–7.
2. Åkerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Int J Neurosc. 1990;52:29–37.
3. Harma M, Tarja H, Irja K, Mikael S, Jussi V, Anne B, et al. A controlled intervention study on the effects of a very rapidly forward rotating shift system on sleep-wakefulness and well-being among young and elderly shift workers. Int J Psychophysiol. 2006;59(1):70–9.
4. Lo JC, Groeger JA, Santhi N, Arbon EL, Lazar AS, Hasan S, et al. Effects of partial and acute total sleep deprivation on performance across cognitive domains, individuals and circadian phase. PLoS One. 2012;7(9):e45987.
5. Sagaspe P, Taillard J, Akerstedt T, Bayon V, Espie S, Chaumet G, et al. Extended driving impairs nocturnal driving performances. PLoS One. 2008;3(10):e3493.
6. Kaida K, Akerstedt T, Kecklund G, Nilsson JP, Axelsson J. Use of subjective and physiological indicators of sleepiness to predict performance during a vigilance task. Ind Health. 2007;45(4):520–6.
7. Akerstedt T, Anund A, Axelsson J, Kecklund G. Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. J Sleep Res. 2014;23(3):240–52.
8. Baulk SD, Reyner LA, Horne JA. Driver sleepiness–evaluation of reaction time measurement as a secondary task. Sleep. 2001;24(6):695–8.
9. Kaida K, Akerstedt T, Takahashi M, Vestergren P, Gillberg M, Lowden A, et al. Performance prediction by sleepiness-related subjective symptoms during 26 h sleep deprivation. Sleep Biol Rhythms. 2008;6:234–41.
10. Mc NQ. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–7.
11. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20.
13. Bowker AH. A test for symmetry in contingency tables. J Am Stat Assoc. 1948;43(244):572–4.
14. Eriksen CA, Åkerstedt T, Kecklund G, Åkerstedt A. Comment on short-term variation in subjective sleepiness. Percept Mot Skills. 2005;101:943–8.