doi: 10.1111/j.1365-2869.2008.00700.x

Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard

HEIDI DANKER-HOPFE[1], PETER ANDERER[2,3], JOSEF ZEITLHOFER[4], MARION BOECK[4], HANS DORN[1], GEORG GRUBER[3], ESTHER HELLER[1], ERNA LORETZ[3], DORIS MOSER[4], SILVIA PARAPATICS[3], BERND SALETU[2], ANDREA SCHMIDT[1] and GEORG DORFFNER[3,5]

[1]Department of Psychiatry and Psychotherapy, Charité–Universitätsmedizin Berlin, Berlin, Germany, [2]Department of Psychiatry, Medical University of Vienna, Vienna, Austria, [3]The Siesta Group Schlafanalyse GmbH, Vienna, Austria, [4]Department of Neurology, Medical University of Vienna, Vienna, Austria and [5]Institute of Medical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University of Vienna, Vienna, Austria

Accepted in revised form 28 July 2008; received 15 February 2008

SUMMARY Interrater variability of sleep stage scorings has an essential impact not only on the reading of polysomnographic sleep studies (PSGs) for clinical trials but also on the evaluation of patients' sleep. With the introduction of a new standard for sleep stage scorings (AASM standard) there is a need for studies on interrater reliability (IRR). The SIESTA database resulting from an EU-funded project provides a large number of studies (n = 72; 56 healthy controls and 16 subjects with different sleep disorders, mean age ± SD: 57.7 ± 18.7, 34 females) for which scorings according to both standards (AASM and R&K) were done. Differences in IRR were analysed at two levels: (1) based on quantitative sleep parameter by means of intraclass correlations; and (2) based on an epoch-by-epoch comparison by means of Cohen's kappa and Fleiss' kappa. The overall agreement was for the AASM standard 82.0% (Cohen's kappa = 0.76) and for the R&K standard 80.6% (Cohen's kappa = 0.68). Agreements increased from R&K to AASM for all sleep stages, except N2. The results of this study underline that the modification of the scoring rules improve IRR as a result of the integration of occipital, central and frontal leads on the one hand, but decline IRR on the other hand specifically for N2, due to the new rule that cortical arousals with or without concurrent increase in submental electromyogram are critical events for the end of N2.

KEYWORDS AASM scoring standard, interrater reliability, Rechtschaffen and Kales, SIESTA project, sleep stage scoring

INTRODUCTION

In May 2007 'The AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specification' was published by the American Academy of Sleep Medicine (Iber *et al.*, 2007). This manual is supposed to replace the 'old' scoring rules published by Rechtschaffen and Kales (1968), the gold standard for scoring since the late 1960s ('R&K standard'). For details concerning the historical background and the developmental processes of the 'AASM standard' we refer to Silber *et al.* (2007) and Iber *et al.* (2007). The authors state: 'Although the first generally accepted scoring manual by Rechtschaffen and Kales has served the field well, the need for modification and additions have been apparent to sleep scientists and clinicians for a number of years. This new scoring manual represents an attempt to combine the best available evidence with the

Correspondence: Prof. Dr Heidi Danker-Hopfe, Department of Psychiatry and Psychotherapy, Charité–Universitätsmedizin Berlin, Campus Benjamin Franklin, Eschenallee 3, D-14050 Berlin, Germany. Tel.: +49-30-8445-8600; fax: +49-30-8445-8233; e-mail: heidi.danker-hopfe@charite.de

opinion of experts in sleep science and medicine' (Iber *et al*., 2007, p. 14). Although the rules and specifications in the visual scoring of sleep retain much of the framework of the R&K standard, there are some new definitions and rule modifications. The main differences with regard to staging rules are: (1) the number of stages is reduced from 7 [Wake, S1, S2, S3, S4, rapid eye movement (REM) and MT] to 5 [W, N1, N2, N3 (collapse of the former stages S3 and S4) and R; MT is not scored as a separate stage]; (2) alpha activity is scored primarily based on an occipital derivation; (3) delta activity and K-complexes are judged primarily based on a frontal derivation; (4) there are new rules concerning the end of N2 due to a cortical arousal.

One of the limitations of the R&K standard is the reliability of the epoch-based (sleep) staging (for details see Silber *et al*., 2007). The Task Force responsible for the revision of the visual scoring reviewed seven studies examining the interrater reliability (IRR) of human expert scorers. Two of seven studies were graded as evidence level 1 (well-controlled studies – including blinded analysis and randomization, when relevant – low type I error – adequate statistical analysis and level of significance of result – and low type II error – adequate sample size – comparison to a reference standard; well-defined, homogeneous samples; for details of the evidence levels see Silber *et al*., 2007, p. 122). Two studies were graded as evidence level 2 (well-controlled studies as in 1 with higher type II error – smaller sample size), another two were graded as evidence level 3 (less-well-controlled studies; not using reference standards; less-well-defined or non-homogeneous samples) and one study was graded as level 4 (uncontrolled and observational studies with reasonably well-defined samples, adequate sample size, and standardized techniques). One of the studies graded as evidence level 1 is based on the SIESTA R&K database (Danker-Hopfe *et al*., 2004).

The SIESTA R&K database was established by the EU-funded SIESTA project and offered the opportunity to analyse interrater agreement between experienced scorers from eight European sleep labs on a large sample of healthy subjects ($n = 198$) and patients ($n = 98$) with different sleep disorders. The results for patients were documented by Danker-Hopfe *et al*. (2004).

Sharing the appraisal of the AASM Visual Scoring Task Force that 'future studies of inter- and intra-rater reliability for visual scoring of the new system are essential' (Silber *et al*., 2007, p. 129), the purpose of the present paper is to investigate the impact of the application of the revised recommendations for sleep stage scoring on IRR. Data used for this study (the SIESTA AASM database) are part of the SIESTA R&K database.

Overall there are three approaches to the evaluation of IRR: (1) visual inspection of hypnograms; (2) analysis of the degree of agreement on quantitative sleep parameters and (3) analysis of the degree of agreement on the basis of the scoring for single epochs (Danker-Hopfe and Herrmann, 2001).

METHODS

Results are based on a subsample of the SIESTA database. The SIESTA database, which was described in detail by Klösch *et al*. (2001) and Rappelsberger *et al*. (2001), comprises 572 recordings of 286 subjects [189 healthy controls and 98 patients with the ICD-10 diagnoses of non-organic insomnia related to depression or generalized anxiety disorder, Parkinson's disease, periodic limb movement disorder (PLMD) or sleep apnoea].

The recording protocol specified at least 16 channels of biosignals: six electroencephalogram (EEG) channels with mastoid (M1, M2) as reference (Fp1-M2, C3-M2, O1-M2, Fp2-M1, C4-M1, O2-M1) and an additional EEG channel M1-M2 for re-referencing; two EOG channels with slightly different electrode placement than the standard R&K recommendation (Häkkinen *et al*., 1993); submental electromyogram (EMG) and EMG recorded from electrodes placed at the *musculus anterior tibialis* of the left and right leg (electrodes were linked); electrocardiogram and respiratory signals (airflow, movements of the chest wall and abdomen and O_2 saturation of arterial blood). Recording of additional EEG channels was recommended whenever possible. Two of the eight sleep labs (Department of Neurology, Medical University of Vienna and Department of Psychiatry and Psychotherapy, Charité – Campus Benjamin Franklin) contributed multichannel recordings (19 EEG channels including F3-M2 and F4-M1). The sampling rates were 200 and 256 Hz for the PSGs recorded in Berlin and Vienna, respectively. The filter settings for the recordings from Berlin were 0.5–85 Hz. The filter settings for the recordings from Vienna were 0.1–70 Hz. Thus, the R&K SIESTA database comprised a subsample which could be used for analysing the impact of scoring according to the new AASM rules (Iber *et al*., 2007) on IRR (the AASM SIESTA database). The subsample consisted of studies (second night recordings after an adaptation night) of 72 subjects, 56 healthy controls and 16 patients with an ICD-10 diagnosis of non-organic insomnia related to generalized anxiety disorder ($n = 5$), Parkinson's disease ($n = 6$), or PLMD ($n = 5$). The mean age of the 38 females and 34 males was 57.7 ± 18.7 years, ranging from 21 to 81 years. There were no age differences between males (59.8 ± 17.9) and females (55.8 ± 19.5) in the sample (Wilcoxon's two sample test: $P = 0.44$). The mean age of patients and healthy controls was 58.5 ± 19.8 and 54.9 ± 14.7, respectively, with the differences again not reaching statistical significance (Wilcoxon's two-sample test: $P = 0.32$). Scorings performed according to the AASM standard will be referred to as AASM scorings, scorings according to the Rechtschaffen and Kales standard will be referred to as R&K scorings.

All recordings were done digitally. For the exchange of the digital polygraphic sleep data and the documentation of the hypnograms the EDF data format (Kemp *et al*., 1992) was used. For each of the 72 recordings two independent scorings according to the rules of R&K (R&K1 and R&K2) and a consensus scoring (R&KC) were available. The first scoring

was always performed by the recording lab, while the second scoring lab and the consensus scoring lab were randomly assigned (Danker-Hopfe *et al*., 2004, 2005). Furthermore, for all these recordings (at least) two independent scorings according to the AASM rules (AASM1 and AASM2) were available. A subset of 12 recordings (three healthy subjects, three subjects with non-organic insomnia related to generalized anxiety disorder, three with Parkinson's disease, and three with a PLMD) was scored independently by six scorers. In total, seven scorers from three labs in Austria (Department of Neurology, Medical University of Vienna: DM and MB, The Siesta Group GmbH: EL, GG and SP) and Germany (Department of Psychiatry and Psychotherapy, Charité – Campus Benjamin Franklin: EH and AS) participated in the AASM scoring task. All seven scorers have a long-standing experience and practice in sleep scoring. The number of PSGs scored previously by each scorer varied between 350 and more than 7000. As the data were not explicitly collected for the present study, it has to be considered a retrospective study.

For the quantitative sleep parameters derived from the scorings, correlation analysis (intraclass correlations) and descriptive statistics of absolute differences were used to describe the degree of agreement. For a comparison of the degree of agreement at the epoch level Cohen's kappa as well as percentages were used to evaluate the IRR. Epochs considered as not scoreable or scored as MT by at least one of the scorers ($n = 503$ for the AASM scoring and $n = 394$ for the R&K scoring) were omitted from further analysis. For the purpose of comparison, kappa coefficients for the agreement between R&K scorings, stages S3 and S4 were combined as slow-wave sleep (SWS).

While Cohen's kappa is appropriate for quantifying the level of agreement for qualitative data between two scorers, Fleiss' kappa was used to measure the level of agreement between all six scorers (for more methodological details see Penzel *et al*., 2003). The effect of age, sex and the nature of the sleep disorder was analysed by means of regression analysis. An effect was considered to be statistically significant if P of the regression coefficient was < 0.05.

RESULTS

For the purpose of demonstration, hypnograms (two independent scorings according to the R&K standard, the R&K consensus scoring, and six independent scorings according to the AASM standard) for two subjects are displayed in Fig. 1a and b.

Agreement based on quantitative sleep parameters

Interrater reliability of sleep stage scoring which is reflected in quantitative sleep parameters was analysed by means of intraclass correlations. The AASM manual (Iber *et al*., 2007) recommends that besides *lights out clock time* and *lights on clock time* the following sleep parameters should be reported: total sleep time (TST; in min), total recording time ('lights out' to 'lights on' in min), sleep latency (SL; 'lights out' to first epoch of any sleep in min), stage R latency (sleep onset to first epoch of stage R in min), wake after sleep onset (WASO; stage W during total recording time minus sleep latency in min), percent sleep efficiency (TST/total recording time), time in each stage (min), percent of TST in each stage [(time in a stage/TST) × 100]. For these parameters the degree of agreement was analysed in terms of intraclass correlations.

The data presented in Table 1 show that for 10 of the 13 variables presented interrater agreement was higher for scorings according to the AASM standard than for scorings according to the R&K standard. The difference was most pronounced for stage S1/N1 (in min as well as % of TST). There was also a considerable difference in intraclass correlation for SWS/N3 and REM/R; for the stage expressed as % of TST the difference in intraclass correlation was higher than for the stage expressed in minutes. The three variables for which agreement was lower when the AASM scoring standard was applied were stage R latency (min, 0.7306 versus 0.7735), stage N2-min and stage N2-% (min: 0.5714 versus 0.6211; % of TST: 0.3373 versus 0.4019).

Agreement based on epoch-by-epoch comparison

The overall level of agreement for the 72 subjects was $\kappa = 0.7626$ (or 82.0%) for the AASM scoring ($n = 68\,029$ epochs) and $\kappa = 0.7352$ (or 80.6%) for the R&K scoring ($n = 68\,051$ epochs). According to the arbitrary benchmarks for the evaluation of κ by Landis and Koch (1977) both values reflect a substantial ($0.61 \leq \kappa \leq 0.80$) agreement.

The nature of the discrepancies in the scoring of all epochs, which amount to 18.0% for the scoring according to the AASM standard and 19.4% for the R&K scoring is detailed in Table 2 and Fig. 2. The most common combination of deviating scorings was between SWS/N3 and S2/N2 for both standards (AASM and R&K). The number and percentage was a little lower for the AASM standard (30.6%) than for the R&K standard (34.2%). Almost as common as the discrepancy between SWS/N3 and S2/N2 was that between S1/N1 and S2/N2 (AASM: 30.0%; R&K: 31.6%). Again the discrepancy was a little higher for the R&K scoring than for the AASM scoring. These two combinations of stages account for more than 60% of all discrepancies for both the scoring according to the R&K standard and the scoring according to the AASM standard. The combination of Wake and S1/N1 scorings is a little more common for the AASM scoring (17.1%) than for the R&K scoring (15.1%). Overall, however, this combination of stages contributes much less than that between SWS/N3 and S2/N2 or between S1/N1 and S2/N2. Less common is the combination of stages S1/N1 and REM, although it is a little more common for the AASM standard (9.6%) than for the R&K standard (7.2%). All other combinations of stages occur with a frequency of approximately 6% or less and have almost the same frequency of occurrence for both scoring standards. Overall the distribution of combinations of deviating scorings is statistically different for the two

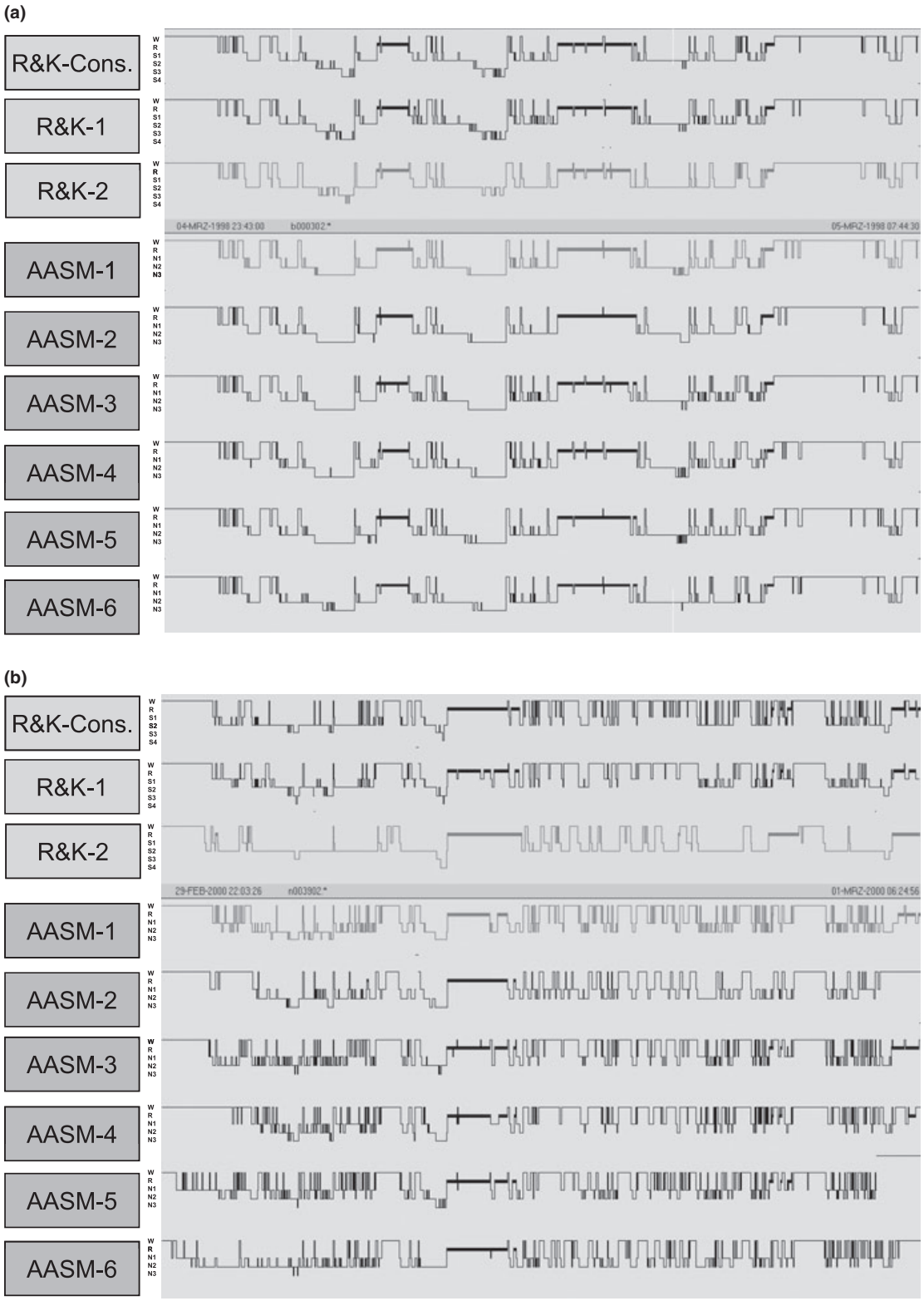

Figure 1. (a) Hypnograms of a 38-year-old male patient with a generalized anxiety disorder plus non-organic insomnia, resulting from two independent and a consensus scoring according to the R&K standard and from six independent scorings according to the AASM standard. (b) Hypnograms of a 67-year-old male patient with Parkinson's disease, resulting from two independent and a consensus scoring according to the R&K standard and from six independent scorings according to the AASM standard.

Table 1 Intraclass correlations for quantitative sleep parameters

Parameter	*R&K standard*	*AASM standard*
Total sleep time (TST, min)	0.9187	0.9251
Sleep latency (min)	0.7315	0.7909
REM latency (min)/stage R latency (min)	0.7735	0.7306
Wake after sleep onset (min)	0.8746	0.9189
Sleep efficiency index (%)/Percent sleep efficiency (%)	0.8753	0.9139
Time in stage S1 (min)/Time in stage N1 (min)	0.4389	0.7966
Time in stage S2 (min)/Time in stage N2 (min)	0.6211	0.5714
Time in stage SWS (min)/Time in stage N3 (min)	0.6277	0.6975
Time in stage REM (min)/Time in stage R (min)	0.8498	0.9320
Time in stage S1 (% of TST)/N1 (% of TST)	0.4014	0.7423
Time in stage S2 (% of TST)/N2 (% of TST)	0.4019	0.3373
Time in stage SWS (% of TST)/N3 (% of TST)	0.6168	0.7253
Time in stage REM (% of TST)/R (% of TST)	0.7875	0.9013

SWS, slow-wave sleep; REM, rapid eye movement.

Table 2 Combinations of deviating expert scorings: number of epochs according to the AASM scoring (lower triangle matrix, total n = 12 244); number of epochs according to the R&K scoring (upper triangle matrix, total n = 13 186)

Stage	*Wake*	*S1*	*S2*	*SWS*	*REM*
Wake		2252	585	9	174
N1	1918		3954	21	1272
N2	472	3869		4036	878
N3	29	41	4187		5
R	154	883	689	2	

SWS, slow-wave sleep; REM, rapid eye movement.

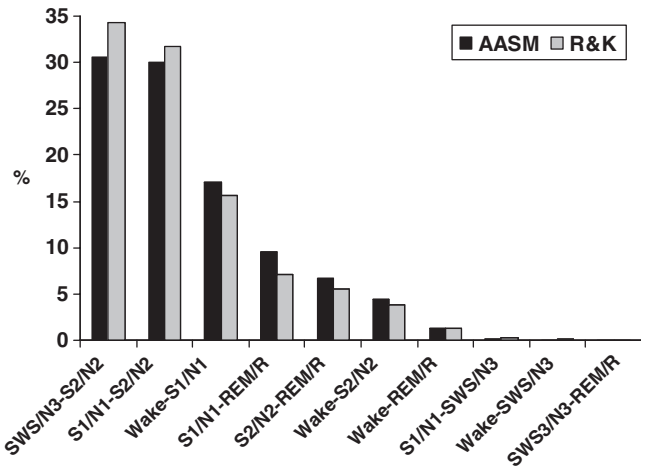

Figure 2. Combinations of deviating expert scorings expressed as percentages of overall mismatches.

scorings: ($\chi^2_{(9)}$ = 120.297, P < 0.0001). The differences observed in the occurrence of combination of S1/N1 and REM (more frequent in the AASM scoring) contribute most to the value of the test-statistic, followed by the differences in the occurrence of deviating S2/N2 and SWS/N3 scorings (less frequent in the R&K scoring).

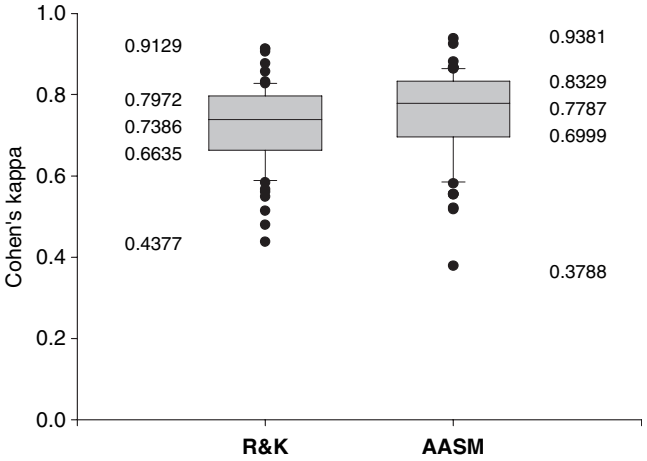

Figure 3. Boxplots of Cohen's kappa for n = 72 studies scored according to the R&K and the AASM standard.

The distribution of individual kappa values, i.e. values calculated for single recordings, is displayed in Fig. 3.

Although the range of kappa is a little wider for AASM scorings (0.3788–0.9381) than for R&K scorings (0.4377–0.9129), Fig. 3 underlines that the overall agreement is higher for AASM than for R&K scorings (mean ± SD: AASM: 0.7505 ± 0.1096; R&K: 0.7236 ± 0.0979). This is also reflected in higher median values as well as higher quartiles. Furthermore, the differences between scorings according to the different standards were tested on the basis of the distribution of recording-specific differences of kappa by means of a t-test for paired observations. The normality assumption as tested by means of a Kolmogorov D-test (appropriate if sample size is >50) with a type I error P < 0.01. The distribution of recording-specific differences is displayed in Fig. 4.

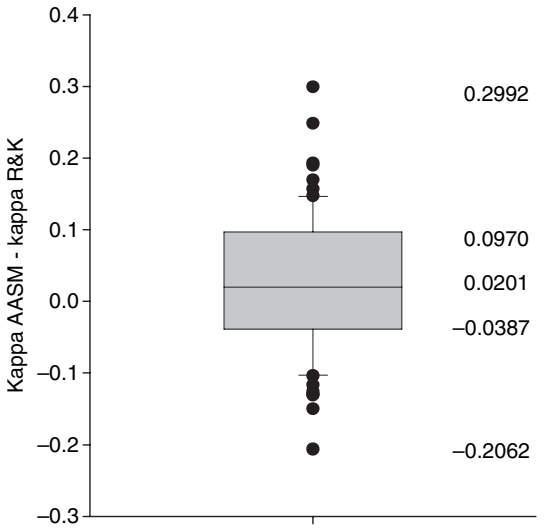

Figure 4. Boxplot of study-specific differences between Cohen's kappa for R&K and AASM scorings (n = 72 studies).

The figure demonstrates that there is a shift towards higher levels of agreement, as reflected by Cohen's kappa coefficients, for AASM. This shift was found to be statistically significant by means of a *t*-test for paired observations ($P = 0.02$).

According to the arbitrary benchmarks for the evaluation of κ by Landis and Koch (1977) agreement was perfect ($\kappa > 0.80$) for 42.3% of the AASM scorings and 22.5% of the R&K scorings, substantial ($0.61 \leq \kappa \leq 0.80$) for 42.3% of the AASM scorings and 66.2% of the R&K scorings, and moderate ($0.41 \leq \kappa \leq 0.60$) for 14.1% of the AASM scorings and 11.3% of the R&K scorings. Only one kappa was in the range of a fair ($0.21 \leq \kappa \leq 0.40$) agreement and this was for a scoring according to AASM standards. The differences in the distributions of agreement between the two scoring standards according to categories of Cohen's kappa were also statistically significant ($\chi^2 = 15.54$; df = 3; $P = 0.02$, note: to meet test assumptions categories moderate and fair were combined).

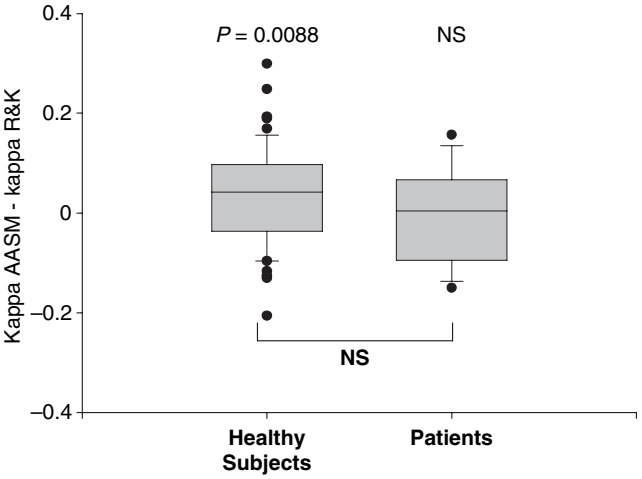

Figure 6. Boxplots of study-specific differences between Cohen's kappa for R&K and AASM scorings by patient status.

Effect of patient status on the IRR for R&K and AASM scorings

A stratified analysis by patient status (healthy subject or patient, Fig. 5) showed no statistically significant differences in the distribution of kappa values calculated for R&K scorings. With regard to AASM scorings agreement seems to be better for healthy subjects. This, however, was not found to be statistically significant, which might be because of the comparatively small sample size for patients ($n = 16$). The difference in agreement between the scoring according to the two standards was statistically significant ($P = 0.009$) for healthy subjects, but not for patients (Fig. 6).

Effect of age on the IRR for R&K and AASM scorings

The level of agreement in sleep stage scoring was found to strongly depend on subjects' age. This was true for scorings according to both the standards. The level of agreement decreased with increasing age of the subject. This correlation was more pronounced for scorings according to the R&K standard (R&K: $r_s = -0.49$, $P = 0.0002$; AASM: $r_s = -0.31$, $P = 0.008$). However, the individual differences in Cohen's kappa for the two scorings did not correlate with age: $r_s = 0.09$, $P = 0.44$.

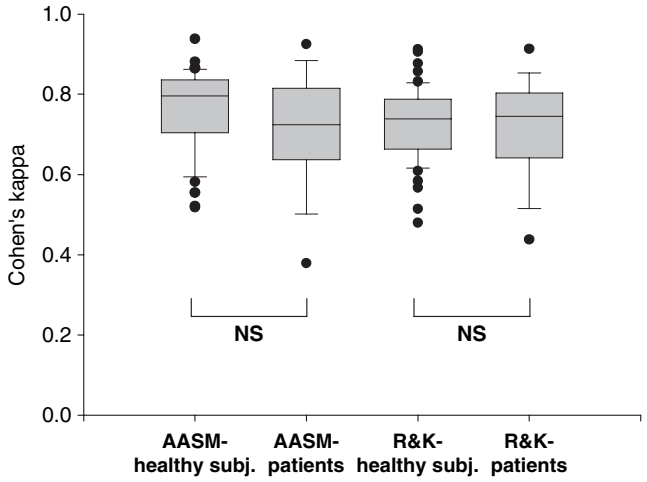

Figure 5. Boxplots of Cohen's kappa for R&K and AASM scorings by patient status.

Effect of gender on the IRR for R&K and AASM scorings

At a univariate level there was no gender difference in the distribution of Cohen's kappa, neither for the ones calculated for R&K scorings nor for those calculated for AASM scorings. However, looking at the individual differences between Cohen's kappa for the R&K and for the AASM scoring separately for males and females, the differences were found to marginally reach statistical significance for females ($P = 0.04$), but not for males.

Sleep stage-specific analyses

A sleep stage-specific analysis of the agreement between two experts revealed the highest average agreement (over all epochs combined) for stage REM/R, followed by Wake, SWS/N3, NREM2/N2, and NREM1/N1 (Table 3). The ranking is exactly the same for agreement expressed in percent. With regard to kappa SWS/N3 and S2/N2 change the position.

The stage-specific distribution of kappa for single recordings is shown in Fig. 7. Except for the SWS scoring, according to R&K the median of kappa for single recordings is higher than the corresponding overall average. Individual differences of Cohen's kappa between R&K scorings and AASM scorings are statistically significant for NREM1/N1 sleep, REM/R sleep and Wake. This indicates that the IRR is higher when the AASM standard is applied.

The rate of interrater agreement for the 12 studies scored by all six human experts, as reflected by Fleiss' kappa is displayed

Table 3 Sleep stage-specific degree of agreement

R&K standard				*AASM standard*			
Stage	*Kappa*	*%*	*Median (studies)*	*Stage*	*Kappa*	*%*	*Median (studies)*
REM	0.8739	96.6	0.8892	R	0.9054	97.5	0.9144
Wake	0.8104	94.7	0.8285	Wake	0.8608	95.6	0.8695
SWS	0.7109	94.1	0.6968	N3	0.7285	93.8	0.7414
S2	0.7223	86.3	0.7247	N2	0.7188	86.5	0.7382
S1	0.4087	88.9	0.4151	N1	0.4608	90.1	0.5694

REM, rapid eye movement.

in Fig. 8. The effect of age (continuous variable), sex (dichotomous with female as default) and patient status (dichotomous with non-patient status as default) was analysed by means of regression analysis based on Cohen's kappa estimated for single recordings while distinguishing five stages (Table 4).

The results show that overall age has a statistically significant effect on IRR of sleep stage scorings. With increasing age of the subjects reliability decreases, while gender and patient status (yes/no) have no statistically significant effect. This is true for scorings according to both standards. Stage-specific analyses reveal that the age effect is mainly because of the age-dependent reliability of the scoring of S2/N2 and SWS/N3. Both stages are less reliably identified with increasing age of the subject. This is again true for scorings according to both the AASM and the R&K standard. Only for scorings according to the R&K standard reliability of identifying S1 also decreases with increasing age of the subjects. While the reliability of scoring Wake and REM/R is not at all affected by age, gender and patient status (neither for AASM nor for R&K scorings), gender has a significant effect on the reliability of SWS/N3 scorings. For both standards reliability is lower for males (Table 4).

Overall the independent variables explained 16.7% of the variance of the kappa coefficients in R&K scorings and 13.2% in AASM scorings. This indicates that the variables introduced in the model only partly explain the scoring divergence/convergence.

DISCUSSION

Based on Cohen's kappa the overall interrater agreement between human expert scorers is substantial ($0.61 \leq \kappa \leq 0.80$). This is true (1) for estimates from the pooled data as well as for means calculated from individual studies; and (2) for AASM scorings as well as for R&K scorings. However, agreement is always higher for AASM scorings than for R&K scorings. The distribution of individual differences of Cohen's kappa between AASM scorings and R&K scorings differs statistically significantly from a mean of zero. The higher level of agreement is also reflected by statistically significant differences in the distribution of Cohen's kappa classified according to Landis and Koch (1977). As stated in methods, the AASM scoring was performed after a 2-day training symposium by experienced sleep scorers. A training symposium was necessary, as the scorings were done in July and August 2007, shortly after the publication of the AASM manual and consequently no scorers with experience in AASM scoring were available. For the R&K scorings, however, which were done during the SIESTA project in 2001, a large pool of experienced scorers was available in the participating

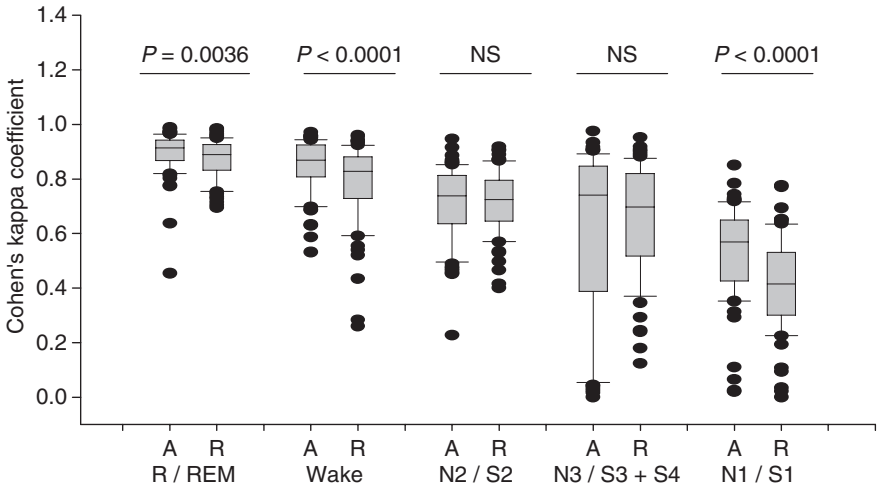

Figure 7. (Sleep) stage-specific boxplots of Cohen's kappa for $n = 72$ studies scored according to the R&K and the AASM standard. A, AASM; R, R&K.

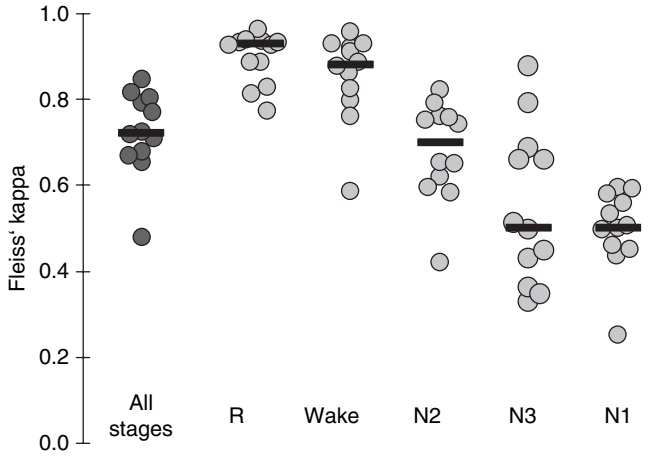

Figure 8. Distribution of Fleiss' kappa for $n = 12$ studies scored according to the AASM standard by six independent human experts.

European sleep labs and thus no such training symposium was necessary. Nevertheless, the SIESTA R&K scorers were encouraged to score as close as possible in accordance with the R&K manual at the SIESTA kick-off meeting. Moreover, both the R&K scorers as well as the AASM scorers had been informed before the start of the scoring task that their scorings will be compared with at least one second scoring. Thus, even though we tried to achieve a similar scoring situation for R&K and AASM scoring, the smaller number of experts as well as the AASM training symposium risks to inflate the AASM scoring agreement when compared with the R&K one.

Regression analyses of the effects of gender, age and patient status revealed very similar results. Neither scoring is affected by patient status, i.e. agreement for patient studies is not worse than the studies conducted in healthy subjects. This result may be because of the small size of the patient sample ($n = 16$). In this study, which comprises 16 patients and 56 healthy subjects, Cohen's kappa for R&K scorings is 0.7352. This is substantially higher than Cohen's kappa of 0.6816 for the whole sample of patients only from the SIESTA R&K database (Danker-Hopfe *et al*., 2004) and for patients' data from the literature: 0.59 in a sample of OSAS patients (Norman *et al*., 2000). In a study on differences in sleep stage scoring between accredited sleep laboratories in Germany, Penzel *et al*. (2003) underline that the degree of agreement for OSAS patients is lower than for healthy subjects and patients with PLMS and a possible reason for this observation is discussed.

Regression analyses revealed further that with both the AASM and the R&K scoring gender effects were only observed for stage SWS. Only age (which has a statistically significant effect on Cohen's kappa for both standards when looking at the five stages simultaneously as well as for stages S2/N2 and SWS/N3 separately) has a statistically significant effect on the rate of agreement for S1 (R&K scoring), but not for N1 (AASM scoring).

Looking at sleep stages separately it emerges that agreement is best for stage REM/R, followed by Wake, S2/N2, SWS/N3, and finally S1/N1. This is true for the distribution of Cohen's kappa for pairwise comparisons ($n = 72$ studies) as well as for Fleiss' kappa for agreement between all six human experts ($n = 12$ studies). A recent review of agreement for R&K scorings by the AASM Visual Scoring Task Force summarized: 'inter- and intra-rater reliability were substantial for staging of records as a whole with the use of the R&K montages. ...Greatest interrater accuracy was achieved for REM sleep followed by stage 2 sleep. Lowest reliability was found for stage 1 sleep, while reliability for wake and SWS was moderate. … It was concluded that scoring rules for stage 1 and SWS sleep needed reassessment' (Silber *et al*., 2007, p. 124).

Looking at intraclass correlations, agreement is best for the following four variables: TST, WASO, sleep efficiency index

Table 4 Effect of gender, age and patient status on the interrater reliability: results of multiple regression analysis (*b*: regression coefficient and its level of significance: *P*)

	Five stages		*Wake*		*S1*/N1		*S2*/N2		*SWS*/N3		*REM*/R	
Variable	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*
AASM standard												
Intercept	0.8865	< 0.0001	0.8887	< 0.0001	0.6291	< 0.0001	0.8798	< 0.0001	0.9802	< 0.0001	0.9386	< 0.0001
Gender (male)	−0.0327	ns	−0.0138	ns	0.0198	ns	−0.0292	ns	**−0.1397**	**0.0300**	−0.0136	ns
Age	**−0.0019**	**0.0060**	−0.0006	ns	−0.0016	ns	**−0.0024**	**0.0040**	**−0.0047**	**0.0080**	−0.0006	ns
Patient (healthy)	−0.0493	ns	−0.0119	ns	−0.0687	ns	−0.0598	ns	−0.1011	ns	−0.0033	ns
Adj. R^2	0.1324		0.0000		0.0134		0.1221		0.1626		0.0000	
R&K standard												
Intercept	0.8644	< 0.0001	0.8738	< 0.0001	0.6231	< 0.0001	0.8887	< 0.0001	0.8954	< 0.0001	0.8942	< 0.0001
Gender (male)	−0.0205	ns	0.0120	ns	0.0307	ns	−0.0431	ns	**−0.1062**	**0.0200**	0.0133	ns
Age	**−0.0022**	**0.0003**	−0.0015	ns	**−0.0036**	**0.0006**	**−0.0026**	**0.0002**	**−0.0037**	**0.0030**	−0.0005	ns
Patient (healthy)	−0.0120	ns	−0.0458	ns	−0.0841	ns	−0.0070	ns	0.0547	ns	0.0049	ns
Adj. R^2	0.1669		0.0062		0.1511		0.1996		0.1869		0.0000	

Default: gender = female, patient status = no sleep disorder. Values in bold indicate statistically significant regression coefficients.

and stage REM in minutes as well as % of TST. The present results for a comparison by epochs show that – as for R&K scorings – for AASM scorings agreement is best for stage R. This is not only reflected in the highest mean Cohen's kappa (which according to the classification reflects *perfect* agreement) and Fleiss' kappa, but also in a comparatively small range of kappa values. The application of the new standard even resulted in a statistically significant improvement ($P = 0.004$) of agreement, as reflected in Cohen's kappa.

Based on the assumption that agreement is best for stage REM, in the revision of the rules the AASM Visual Scoring Task Force 'concentrated predominantly on refining and simplifying sleep scoring rules, especially with regard to the start and end of periods of REM sleep, an area of considerably complexity in the R&K manual' (Silber *et al*., 2007, p. 128). The statistically significantly higher level of agreement of AASM scorings when compared with R&K scorings underlines that the goal to simplify rules might have been achieved, even though the increase might be biased by the limitations of the study design as discussed in the first paragraph of the discussion.

In accordance with data from the literature for R&K scorings, agreement of AASM scorings is worst for stage N1. The mean of Cohen's kappa for R&K scorings is at the borderline between fair ($0.21 \leq \kappa \leq 0.40$) and moderate ($0.41 \leq \kappa \leq 0.60$) agreement, while Cohen's kappa for AASM scorings clearly falls in the range of moderate agreement (Landis and Koch, 1977). Despite the still not convincing results there is a substantial (statistically significant: $P < 0.0001$) improvement in agreement for this stage. This is also reflected in a considerably higher intraclass correlation for variables derived from AASM scorings when compared with R&K scorings. The low level of agreement for N1 is also reflected by Fleiss' kappa and by comparatively low intraclass correlations.

Part of the problem with scoring S1/N1 is that it has always been difficult to identify the transition from Wake to S1/N1. S1/N1 is scored when EEG theta activity predominates over alpha. The introduction of an occipital derivation, where in relaxed wakefulness with closed eyes alpha is maximal, obviously seems to facilitate identification of the transition. Although agreement for S1/N1 has improved considerably, it is still far below that observed for other stages, which might be explained by the fact that 10–20% of the population generate little or no alpha hampering the determination of sleep onset (Silber *et al*., 2007). Basner *et al*. (2008) discuss that kappa values correlate positively with the amount of time spent in the respective sleep stage. At a first glance Cohen's kappa is indeed the lowest for S1/N1, e.g. the sleep stage in which the patients spent least of the time. However, for SWS/N3 – a stage in which the patients spent only 5 min more – kappa is much higher (see Table 2). And finally S2/N2, the stage in which patients spent by far most of their time, Cohen's kappa is much lower than for stages REM/R, Wake/Wake and N3.

As for the patient data from the SIESTA R&K database (Danker-Hopfe *et al*., 2004), agreement for stage Wake was the second best of all stages (which is also reflected by the stage-specific Fleiss' kappa and the intraclass correlations). For the present SIESTA AASM dataset this agreement can be considered as perfect ($\kappa > 0.80$) according to the classification by Landis and Koch (1977). Although this is true for both R&K and AASM scorings, agreement is statistically significantly better ($P < 0.0001$) for scorings according to the AASM standard than for scorings according to the R&K standard. Interestingly, there has not been a substantial change in the scoring rules for this stage, apart from the fact that an occipital derivation is used to evaluate alpha activity, which alone leads to a measurable increase in scoring reliability.

For stages S2/N2 and SWS/N3 agreement can be considered as substantial ($0.61 \leq \kappa \leq 0.80$). Agreement for N3 is slightly better than for the combination of stages S3 and S4 (as reflected by Cohen's kappas and intraclass correlations). At the same time the range of Cohen's kappa is much broader for the AASM scoring than for the R&K scoring. Individual differences of Cohen's kappa between the two scoring standards were not statistically significant. The distribution of Fleiss' kappa indicates that there is a high variability of agreement; for some studies Fleiss' kappa is comparatively high (for single studies even reaching values in the range of R and Wake), while for others it is even worse than for N1. The median of Fleiss' kappa for N3 is equal to the median of agreement for N1.

The major change in the rules for scoring N3 is that a frontal derivation, where the peak-to-peak amplitude of delta waves is usually the highest, is now available for judging delta activity (0.5–2 Hz). The amplitude criterion (>75 μV) was not changed. The present data show that this change alone did not lead to a substantial improvement of IRR for scoring this stage of sleep.

For scorings according to the R&K standard, agreement for stage S2 was higher than for SWS. This was not true for scorings according to the AASM standard, where agreement for N2 was lower (although not statistically significant) than for N3. Stage S2/N2 is the only stage for which agreement is lower when applying the new scoring standard. The higher agreement for S2 in the R&K scoring is also reflected in a higher intraclass correlation for this variable when based on the R&K scoring.

Stage S2/N2 is characterized by specific grapho-elements, i.e. K-complexes and spindles. Although studies on this topic are scarce, a study on spindle detection reveals an agreement of 80–90% (Campbell *et al*., 1980), whereas for K-complexes it is only approximately 50% (Bremer *et al*., 1970). Spindle activity is optimally recorded with central electrodes, while K-complexes and delta-waves are optimally recorded with frontal electrodes (Rodenbeck *et al*., 2006; Silber *et al*., 2007). Although the decision of the AASM Visual Scoring Task Force to discard the '3-min rule' of R&K for scoring S2 (a maximum of 3 min can be scored as S2 if there is no K-complex and/or spindle and no movement arousal) might be expected to result in higher

scoring agreement, this was not the case. On the contrary, scoring reliability for N2 decreased, predominantly because of the prominent role of cortical arousals in the scoring of N2. While according to R&K only a 'movement arousal or a pronounced increase in muscle tone' indicates the change from S2 to S1, any arousal that fulfils the arousal criteria (see arousal rule V.1 in the AASM Manual) leads to a change from N2 to N1. Indeed, no increase in muscle tone is required for arousals in NREM sleep. Therefore, the relatively low IRR for the scoring of arousals that are not associated with an increase in muscle tone (for a review see Bonnet *et al*., 2007) led to the observed significant decrease in the scoring agreement for N2 as compared with S2. Moreover, the decreased reliability may also be a result of introducing a frontal derivation where K-complexes are most marked. Together with the low agreement in the detection of K-complexes (Bremer *et al*., 1970) this can lead to a higher uncertainty with regard to the identification of this grapho-element.

LIMITATIONS

One limitation of this study is that not all scorers who did the scoring according to the AASM standard were the same as the ones who did the R&K scoring, although the AASM scorers were from the same labs. Furthermore the number of R&K scorers (from eight labs) was higher than the number of AASM scorers (from three labs). And finally the R&K scoring was performed without prior training (Danker-Hopfe *et al*., 2004), while the AASM scoring was preceded by a 2-day training symposium with detailed discussions and four test scorings (which of course are not included in the present study) to ensure that the new AASM standard was correctly interpreted by all scorers.

CONCLUSION

Silber *et al*. (2007, p. 129) state: 'No visual based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-second epoch. Nevertheless, we believe it is possible to develop a rigorous, biologically valid scoring system that can be applied meaningfully in clinical and research settings. The new scoring system is presented as a step forward along this path'. Despite the above-mentioned limitations of this study and possible inconsistencies, the results underline that the integration of frontal, central and occipital leads improve the IRR on the one hand. This advantage, however, is counteracted on the other hand by the impairment of the IRR due to the new rule, that any cortical arousal, whether or not it is associated with an EMG increase, determines the end of stage N2. Indeed, the reliability for arousal scoring is specifically low for arousals without concomitant increase in submental EMG activity.

ACKNOWLEDGEMENTS

This study was supported by The Siesta Group Schlafanalyse GmbH, Vienna. The authors would like to express their thanks to Mag. Elisabeth Grätzhofer for her valuable editorial assistance.

REFERENCES

Basner, M., Griefhahn, B. and Penzel, T. Inter-rater agreement in sleep stage classification between centers with different background. *Somnologie*, 2008, 12: 75–84.

Bonnet, M. H., Doghramji, K., Roehrs, T., Stepanski, E. J., Sheldon, S. H., Walters, A. S., Wise, M. and Chesson, A. L., Jr. The scoring of arousal in sleep: reliability, validity, and alternatives. *J. Clin. Sleep Med*., 2007, 3: 133–145.

Bremer, G., Smith, J. R. and Karacan, I. Automatic detection of the K-complex in sleep electroencephalograms. *IEEE Trans. Bio.-Med. Emg. BME*, 1970, 17: 314–323.

Campbell, K., Kumar, A. and Hofmann, W. Human and automatic validation of a phase-locked loop spindle detection system. *Electroencephalogr. Clin. Neurophysiol*., 1980, 48: 602–605.

Danker-Hopfe, H. and Herrmann, W. M. Interrater reliability of sleep stage scoring according to Rechtschaffen and Kales rules (RKR): A review and methodological considerations. *Klin. Neurophysiol*., 2001, 32: 89–99.

Danker-Hopfe, H., Kunz, D., Gruber, G., Klösch, G., Lorenzo, J. L., Himanen, S. K., Kemp, B., Penzel, T., Röschke, J., Dorn, H., Schlögl, A., Trenker, E. and Dorffner, G. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J. Sleep Res*., 2004, 13: 63–69.

Danker-Hopfe, H., Schäfer, M., Dorn, H., Anderer, P., Saletu, B., Gruber, G., Zeitlhofer, J., Kunz, D., Barbanoj, M. J., Himanen, S. L., Kemp, B., Penzel, T., Röschke, J. and Dorffner, G. Percentile reference charts for selected sleep parameters for 20- to 80-year-old healthy subjects from the SIESTA database. *Somnologie*, 2005, 9: 3–14.

Häkkinen, V., Hirvonen, K., Hasan, J., Kataja, M., Värri, A., Loula, P. and Eskola, H. The effect of small differences in electrode position on EOG signals: Application to vigilance studies. *Electroenceph. Clin. Neurophysiol*., 1993, 86: 294–300.

Iber, C., Ancoli-Israel, S. and Chesson, A. and Quan, S. F. for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, American Academy of Sleep Medicine, Westchester, IL, 2007.

Kemp, B., Värri, A., Rosa, A. C., Nielson, K.-D. and Gade, J. A simple format for exchange of digitized polygraphic recordings. *Electroencephalogr. Clin. Neurophysiol*., 1992, 82: 391–393.

Klösch, G., Kemp, B., Penzel, T., Schlögl, A., Rappelsberger, P., Trenker, E., Gruber, G., Zeitlhofer, J., Saletu, B., Herrmann, W. M., Himanen, S. L., Kunz, D., Barbanoj, M. J., Röschke, J., Värri, A. and Dorffner, G. The SIESTA project polygraphic and clinical database. *IEEE Eng. Med. Biol. Mag*., 2001, 20: 51–57.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, 33: 159–174.

Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A. and Rapoport, D. M. Interobserver agreement among sleep scorers from different centers in a large data set. *Sleep*, 2000, 23: 901–908.

Penzel, T., Behler, P.-G., von Buttlar, M., Conradt, R., Meier, M., Möller, A. and Danker-Hopfe, H. Reliability of visual evaluation of sleep stages according to Rechtschaffen and Kales from eight polysomnographs by nine sleep centres (in German). *Somnologie*, 2003, 7: 49–58.

Rappelsberger, P., Trenker, E., Rothmann, C. H., Gruber, G., Sykacek, P., Roberts, S. T., Klösch, G., Zeitlhofer, J., Anderer, P., Saletu, B., Schlögl, A., Värri, A., Kemp, B., Penzel, T., Herrmann, W. M., Hasan, J., Barbanoj, M. J., Röschke, J., Kunz, D. and Dorffner, G. The SIESTA project (in German). *Klin. Neurophysiol.*, 2001, 32: 76–88.

Rechtschaffen, A. and Kales, A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. University of California, Brain Information Service/Brain Research Institute, Los Angeles, CA, 1968.

Rodenbeck, A., Binder, R., Geisler, P., Danker-Hopfe, H., Lund, R., Raschke, F., Weeß, H. G. and Schulz, H. A review of sleep EEG patterns. Part I: a compilation of amended rules for their visual recognition according to Rechtschaffen and Kales. *Somnologie*, 2006, 10: 159–175.

Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M., Kapen, S., Keenan, S. A., Kryger, M. H., Penzel, T., Pressman, M. R. and Iber, C. The visual scoring of sleep. *J. Clin. Sleep Med.*, 2007, 3: 121–131.