



The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring

Richard S. Rosenberg, Ph.D., F.A.A.S.M.; Steven Van Hout, B.S.

American Academy of Sleep Medicine, Darien, IL

Study Objectives: The program provides a unique opportunity to compare a large number of scorers with varied levels of experience to determine sleep stage scoring agreement. The objective is to examine areas of disagreement to inform future revisions of the *AASM Manual for the Scoring of Sleep and Associated Events*.

Methods: The sample included 9 record fragments, 1,800 epochs and more than 3,200,000 scoring decisions. More than 2,500 scorers, most with 3 or more years of experience, participated. The analysis determined agreement with the score chosen by the majority of scorers.

Results: Sleep stage agreement averaged 82.6%. Agreement was highest for stage R sleep with stages N2 and W approaching the same level. Scoring agreement for stage N3 sleep was 67.4% and was lowest for stage N1 at 63.0%. Scorers had particular difficulty with the last epoch of stage W before sleep

onset, the first epoch of stage N2 after stage N1 and the first epoch of stage R after stage N2. Discrimination between stages N2 and N3 was particularly difficult for scorers.

Conclusions: These findings suggest that with current rules, inter-scorer agreement in a large group is approximately 83%, a level similar to that reported for agreement between expert scorers. Agreement in the scoring of stages N1 and N3 sleep was low. Modifications to the scoring rules to improve scoring during sleep stage transitions may result in improvement.

Keywords: Scoring, reliability, stages

Commentary: A commentary on this article appears in this issue on page 89.

Citation: Rosenberg RS; Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med* 2013;9(1):81-87.

Standardized scoring of sleep stages became possible with the publication of a manual by Rechtschaffen and Kales in 1968.¹ R & K, as it came to be known, standardized terminology, recording techniques, and the rules for sleep stages for normal, young adult research volunteers. In 2007, the American Academy of Sleep Medicine (AASM) developed a *Manual for the Scoring of Sleep and Associated Events*.² The goal of the *AASM Manual* was to expand R & K to a patient population. This included standards for scoring of sleep stages, respiratory and limb movement events, as well as rules for scoring of pediatric recordings.

Grigg-Damberger, in a recent review,³ noted that there have been few complaints about the new scoring system. Moser and colleagues⁴ found significant differences in stage scores with the *AASM Manual*. Stages N1 and N3 were increased, whereas stage N2 was decreased with the new scoring as compared to the old. Scorers were drawn from a pool of 30 experienced sleep experts. Danker-Hopfe and colleagues found that use of the *AASM Manual* resulted in slightly improved inter-rater reliability to 82.0% compared to 80.6% using R & K rules.⁵ Her study was based on 72 recordings that were scored with both methods by a pool of 7 experienced scorers.

The AASM inter-scorer reliability (ISR) program was developed to aid sleep centers in fulfilling accreditation standards. The standards require that a sample of randomly chosen records be scored by the center director and each of the technologists involved in record scoring. As a means of achieving this standard, the AASM ISR program provides a record each

month, scored independently by 2 board certified sleep specialists replacing the center director scorer. Program scorers are compared and differences resolved, leading to a final “correct” answer. All participants score the same record sample using a web-based program, and scores are compared to the “correct score.” This allows instantaneous feedback to the scorer and the center manager. Feedback includes percentage agreement with the “correct score” and scores relative to all users. The intent of the program is to add standardized measurement to the quality assurance cycle. Center directors evaluate technologist performance, identify areas of weakness, provide additional training and experience in scoring, and then re-evaluate performance to close the quality assurance loop. The program began in April 2010 and has grown in use since then. At the time of this writing, approximately 2,500 technologists and physicians use the AASM ISR program. This initial review covers only sleep stage scoring, but the program also requires scoring of respiratory events, periodic limb movements, and arousals.

Most studies of sleep study stage scoring agreement evaluate differences between expert scorers, at times from the same laboratory or after an intensive training process to improve agreement. Other studies have compared expert human scorers with automated scoring systems. Some of these studies employ large numbers of epochs to be scored, but none have evaluated more than a handful of scorers. The AASM ISR program provides a unique data set. Instead of a large number of epochs scored by a relatively small number of technologists, the program provides relatively few record fragments scored by a large number

Table 1—Clinical studies used in the analysis

Date	Age	Sex	Study Type	Abnormality	Overall Stage Agreement
June 2011	65	F	Diagnostic	Minimal OSA	91.2
July 2011	58	F	Diagnostic	Mild OSA	82.6
August 2011	76	M	Diagnostic	Severe OSA/CSA	72.5
September 2011	66	M	CPAP	Severe PLMs	86.8
October 2011	53	F	Diagnostic	Mild OSA	84.5
November 2011	26	M	CPAP	Normal	77.7
December 2011	24	M	Diagnostic	Normal	83.5
January 2012	36	M	Diagnostic	Normal	87.9
February 2012	78	F	CPAP	Minimal OSA	74.9

Table 2—Results of user survey

Are you a physician?	Yes 4.3%	No 95.7%				
How long have you been in sleep?	0.0% 1.1%	< 6 Mo 0.9%	6-12 Mo 1.4%	12-36 Mo 6.7%	36-60 Mo 40.1%	60+ Mo 50.0%
What certifications do you have?	None 11.9%	CPSGT 1.1%	RPSGT 87.2%	RST 38.5%	SDS 1.1%	
How did you get your training?	None 0.2%	OJT 84.1%	ASTEP I 10.5%	ASTEP M 21.5%	CoA 9.0%	Other 47.4%

CPSGT, Certified Polysomnographic Technologist; RPSGT, Registered Polysomnographic Technologist; RST, Registered Sleep Technologist; SDS, sleep Disorders Specialty Certificate; OJT, on the job training; ASTEP I, Accredited Sleep Technologist Education Programs Introductory Program; ASTEP M, Accredited Sleep Technologist Education Programs Modules; CoA, Committee on Accreditation of Polysomnographic Technologist Education Programs approved by the Commission on Accreditation of Allied Health Education Programs.

of scorers. Most of the program users had passed a national certification examination and most had 3 or more years of experience in the field. We hypothesized that the large number of scorers would result in lower agreement than in previous studies of scoring experts. We also hypothesized that the majority of disagreements would occur in epochs in the transition from one stage to another.

METHODS

The AASM ISR program employed 200 epoch record fragments from de-identified recordings provided to the AASM by a variety of sleep centers. Polysmith reading software (provided free of charge by Nihon Kohden America) was used to create static images of 30-sec and 120-sec epochs. Users were able to access the epochs and answer 4 standard questions regarding sleep stage, the presence or absence of respiratory events, and the number of arousals and limb movements in a 30-sec epoch. Amplitude markers at $\pm 37 \mu\text{V}$ were provided for rapid measurement of slow wave activity. This methodology differs somewhat from the typical scoring system, but for sleep staging is essentially equivalent to that provided by most sleep recording systems. Beginning in January 2012, a new version of the program automatically maximized epoch display size based on available display and allowed keypad scoring. These improvements increased the speed of scoring but did not alter the nature of the task.

Table 1 provides demographic information for the 9 record fragments under consideration. Records were posted between June 2011 and February 2012. All patient information was removed with the exception of age and sex. Records were available for 3 months; data was therefore collected between June 1, 2011, and April 30, 2012. An attempt was made to include only records with minimal artifact. A variety of sleep disorders and levels of abnormality were used. Diagnostic and PAP titration studies were included. Studies were limited to adult patients.

Although individual user information is provided to the center manager for quality assurance purposes, the AASM does not retain individual identification of users participating in the program. Some indication of the training level of users was considered important for this analysis. Therefore, a survey was sent to all users in July 2012. The survey included questions about the experience level of the users and whether or not they had formal training in sleep technology. Unfortunately, there is no way to connect the scoring data with the survey responses.

Table 2 provides information about the users of the ISR program. More than 90% of users had at least 3 years of experience. Most (87%) were registered polysomnographic technologists. Fewer than 10% had completed a formal education program accredited by the Commission on Accreditation of Allied Health Education Program. Less than 5% of users were physicians.

The ISR program instructed scorers to follow the rules of the *AASM Manual* in assigning a sleep stage to each 30-sec epoch. Data were saved with each keystroke and stored on a file server in

Table 3—Percentage agreement

Majority Score	W	N1	N2	N3	R	Epochs
W	84.1%	10.0%	3.8%	0.3%	1.9%	150
N1	10.9%	63.0%	21.7%	0.1%	4.3%	150
N2	0.7%	6.3%	85.2%	6.6%	1.3%	1,045
N3	0.1%	0.1%	32.3%	67.4%	0.1%	161
R	1.1%	4.4%	3.9%	0.1%	90.5%	294
Total						1,800

Scorers in agreement with the majority score are shown in bold type.

Table 4—Number of scores used to calculate percentages in Table 3

Majority Score	W	N1	N2	N3	R	Epochs	All Scores
W	250,434	29,658	11,222	779	5,703	150	297,796
N1	32,545	187,634	64,700	205	12,910	150	297,994
N2	13,080	116,276	1,570,861	121,141	23,180	1,045	1,844,538
N3	224	298	87,033	181,337	350	161	269,242
R	6,624	25,875	22,883	342	531,611	294	587,335
Totals						1,800	3,296,905

Scorers in agreement with the majority score are shown in bold type.

the AASM office. The current aggregate data do not include individual scorer information. The score chosen by the most scorers was used as the “correct” score. This means that for most epochs, the minimal level of agreement with the correct score was 50%. A few epochs had scores evenly divided over 3 stages; these epochs were assigned the plurality score as the correct score.

Aggregate data were collected for each epoch used in the analysis. For example, 2,246 users scored epoch 41 from the February 2012 record. Of these, 12 scored stage W, 801 scored stage N1, 1,414 scored stage N2, 1 scored stage N3, and 18 scored stage R. N2 was therefore the majority score and was used as the correct answer. The percentage of scores for each stage was calculated for each epoch, then summed by sleep stage and subsequently summed over all of the epochs scored. This provided a percentage correct for each sleep stage as well as percent scored for each of the incorrect possibilities.

A second analysis focused on common transitions. This arbitrarily used transitions where 3 consecutive epochs of 1 stage were followed by 3 consecutive epochs of another stage. In the 9 records used, there were 7 instances meeting this criterion for a transition from stage W to stage N1, 9 instances of stage N1 to stage N2, 6 transitions from stage N2 to stage R, 7 transitions from stage N2 to stage N3, and 9 transitions from stage N3 to stage N2. Epochs were analyzed relative to the stage change, ranging from the third epoch before the change to the third epoch after the change. As with the previous analysis, the percentage scored for each stage in each of the epochs was calculated.

RESULTS

Overall agreement is shown in **Table 3**. Majority scores are shown on the left. Scorers in agreement with the majority score are shown in bold type. The first row shows the epochs scored as stage W by a majority of scorers. The sample of 9 records included 150 such epochs. For stage W a total of 84.1% agreed

with the majority score. Stage N1 was scored by 10.8%, stage N2 by 3.8%, stage N3 by 0.3%, and stage R by 1.9%. The overall agreement with the majority score for all scores for all epochs was 82.6%. Agreement for stages W, N2, and R were above this average, whereas agreement for stages N1 and N3 were below average. The sample is weighted heavily toward stage N2 sleep (58.1% of all epochs), which is consistent with normal stage percentages encountered in a sleep center patient population.

The number of decisions used in the calculation of each of the boxes in **Table 3** is shown in **Table 4**. This shows that 3,296,905 individual scoring decisions contributed to the findings. The number of users of the ISR program increased over time. This resulted in slightly increased weighting of the more recent records as compared to the older records. However, weighting each month’s record equally by averaging the monthly agreement percentages resulted in an overall agreement of 82.4%, very similar to the overall agreement using the aggregate method.

As expected, most of the disagreements occurred with the scoring of “adjacent” sleep stages. For example, the overall agreement with the majority score for stage N1 sleep was 63%. Nearly all of the disagreements were with stage W (10.9%) and stage N2 (21.7%). Scoring of stage N3 sleep also had low agreement at 67.4%, with virtually all of the scorers who disagreed scoring these epochs as stage N2 sleep (32.3%).

The fact that disagreements occurred with adjacent stages did not systematically translate into higher disagreements at sleep transitions. **Figure 1** shows the transition from stage W to stage N1 sleep. Percentage agreement is shown in **Figure 1A**. The majority scores for these 6 epochs were W, W, W, N1, N1, N1. The first 2 bars indicate scoring agreement at about 90% for stage W. In the final stage W epoch, 21.9% of scorers shifted to stage N1 sleep. In the first epoch of stage N1, 25.5% of scorers continued to score stage W. The final epoch of stage N1 shows a substantial number of scorers (18.3%) had already switched to stage N2 sleep. The next question asked was whether the

Figure 1—Transition from stage W to stage N1 sleep (n = 7)

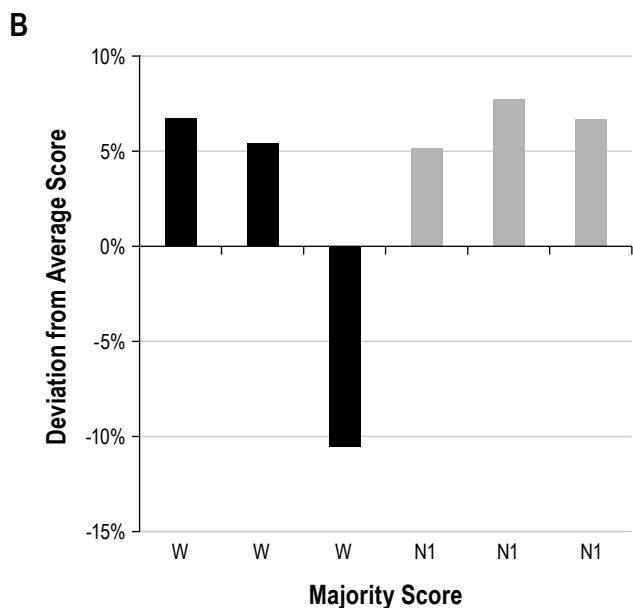
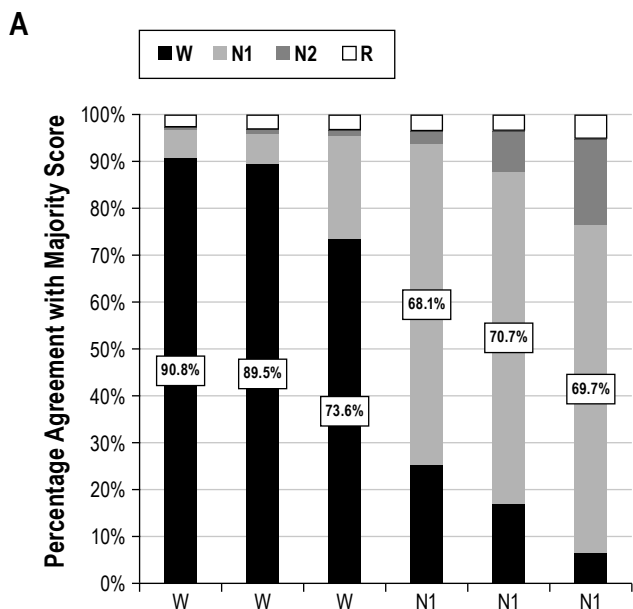


Figure 2—Transition from stage N1 to stage N2 sleep

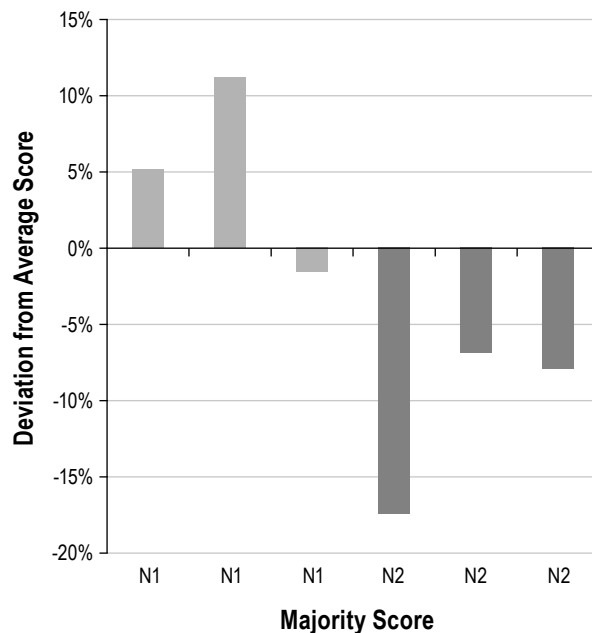
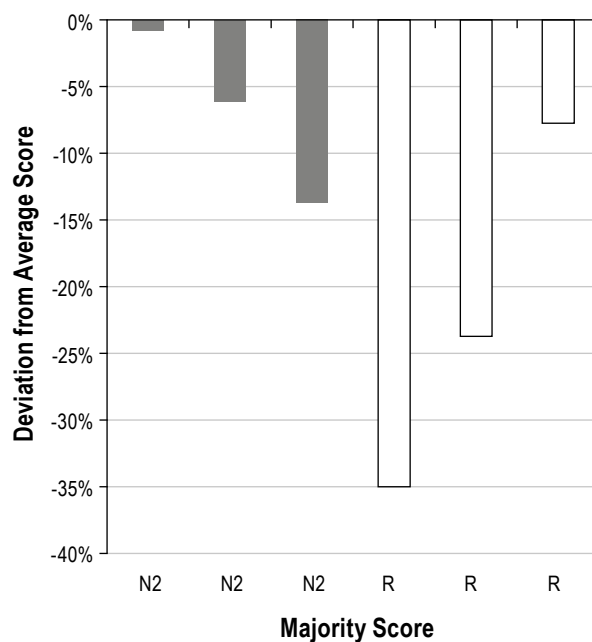


Figure 3—Transition from stage N2 to stage R sleep

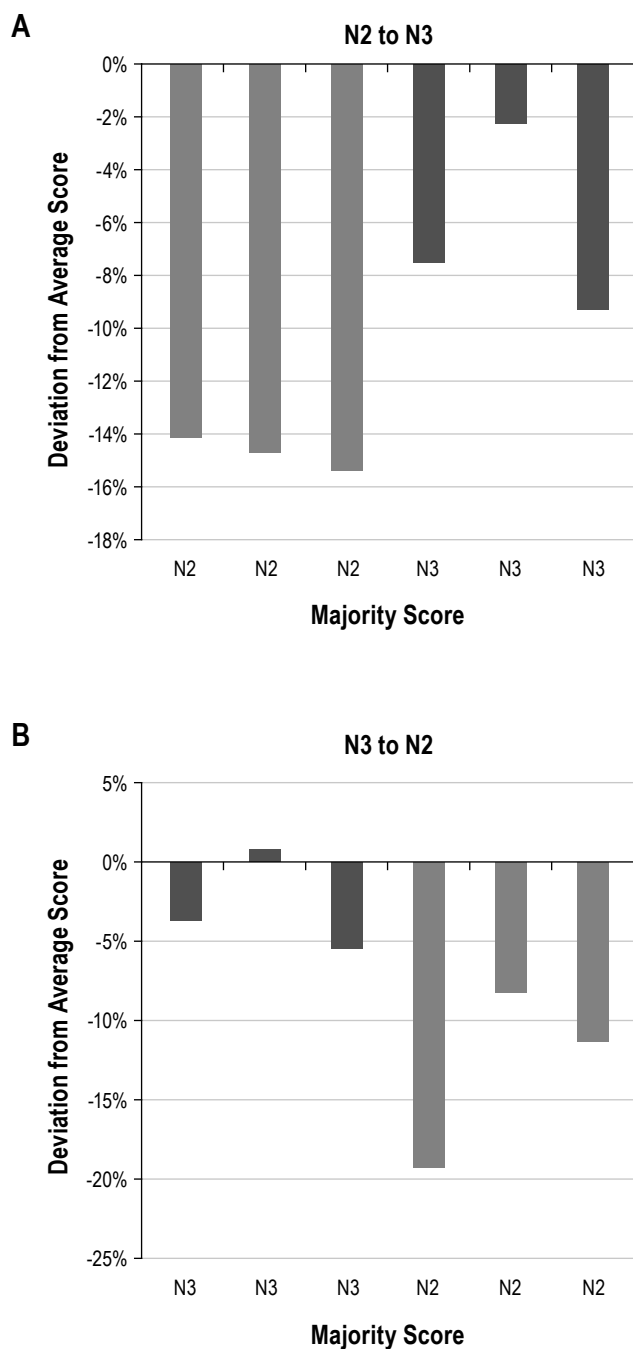


scorers did better or worse than the average agreement in these transition epochs. **Figure 1B** shows variation from the average agreement for each of the epochs in the W to N1 transition. The average score for stage W was 84.1% and for stage N1 was 63%. Agreement with the majority score was above average for the first 2 epochs of stage W in **Figure 1B** (90.8% and 89.5% vs. the average of 84.1%), but dropped to 73.6% in the epoch just before the shift to stage N1. Agreement with the majority score was better than average for the 3 epochs of stage N1 sleep.

Figure 2 shows the transition from stage N1 to stage N2 sleep. Again, results were above average for the epochs scored as stage N1 by the majority of scorers. Stage N2 epochs at the transition were scored with below-average agreement. One possible explanation for the transition data for stage N1 is bout length. The average bout length for stage N1 in this sample was

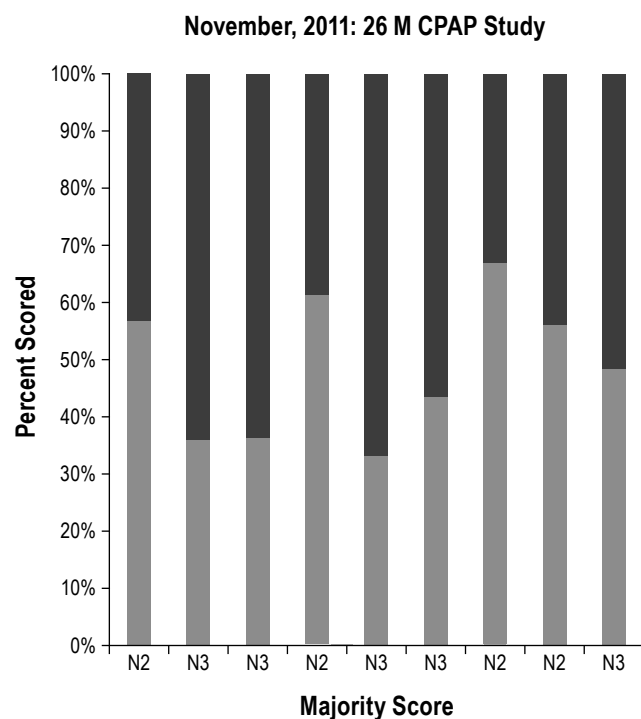
2.03 epochs. Only 12% of N1 bouts were 3 epochs or longer. Thus, the requirement that there be 3 epochs of a stage for inclusion in the transition analysis meant most of the epochs of stage N1 were excluded. There was lower scoring agreement for the common bouts of stage N1 than for the bouts that lasted 3 or more epochs required to be included in the transition analysis. In contrast, the average bout length for stage N2 was 10.1 epochs, and 47% of N2 bouts were longer than 3 epochs. Low agreement on the first epoch of stage N2 suggests some disagreement on the recognition of the first K complex or sleep spindle as required by the scoring rules.

Figure 4—Transitions involving stage N3 sleep



The transition from stage N2 to stage R is shown in **Figure 3**. This was the only transition that showed the expected pattern in that scoring was below average for stage and the greatest disagreement occurred at the border between stages. The epochs at the beginning and the end of the graph were nearly at the average level for stage in agreement. Some scorers begin stage R a bit earlier, whereas other scorers continued to score stage N2. Disagreements occurred on both sides of the transition, with the stage R disagreements appearing larger because the average agreement was higher. Agreement improved as stage R sleep was established, presumably due to the appearance of rapid eye movements and “definite” stage R.

Figure 5—Stage scores during a CPAP titration study



Systematic variations were less evident in the scoring of stage N3. The transition from stage N2 to stage N3 is shown in **Figure 4A**, and the transition from stage N3 to stage N2 is shown in **Figure 4B**. None of the scores improved as the distance from the transition increased; stage N3 was never “established.” This suggests that there is no real transition to stage N3, and that scoring criteria need to be reevaluated on an epoch-by-epoch basis rather than entering a stage for an extended period of time. This is further illustrated by the graph in **Figure 5**, which shows stage scoring for 15 consecutive epochs in a 26-year-old man undergoing a PAP titration study. None of the epochs in this series have more than 70% agreement for either sleep stage. Virtually all of the scores were either stage N2 or stage N3, but there was no transition into or out of stage N3. Further, there was no confusion with other sleep stages.

DISCUSSION

Nine sleep recording samples of 200 epochs each were scored by a large pool of sleep scorers using the scoring methodology of the *AASM Manual*. The overall agreement for sleep stages between scorers in the AASM ISR program was 82.6%. This is remarkably similar to the 83% agreement reported by Ruehland and colleagues⁶ using the *AASM Manual* scoring methodology, a study that compared 2 experienced scorers and 1 scorer with a single year of experience. Danker-Hopfe and colleagues⁵ reported 82% agreement using a pool of 7 experienced scorers scoring 72 records. This provides encouraging news in that agreement among a large group of scorers is similar to agreement between small samples of experienced scorers. Head to head comparisons of the R&K vs. *AASM Manual*

Table 5—Epochs of highest disagreement and potential solutions

Type of Epoch	Percentage Agreement	Possible Explanation	Potential Solution
Last epoch of stage W before sleep onset	73.6%	Confusion between attenuation of alpha and appearance of low voltage mixed frequency EEG	Improve definition of sleep onset as attenuation of alpha rhythm OR replacement with low voltage mixed frequency EEG
First epoch of stage N2 sleep	67.7%	Failure to recognize K complex or sleep spindle	Add amplitude requirement for K complexes and sleep spindles
First epoch of stage R after stage N2	55.4%	Confusion caused by complicated “look back” rule and failure to recognize K complex or sleep spindle	Begin scoring stage R with first epoch of definite stage R
First epoch of stage N3 after stage N2	59.9%	Failure to accurately measure slow wave frequency, amplitude or duration	Use automated analysis of slow wave activity
First epoch of stage N2 after stage N3	65.9%	Failure to accurately measure slow wave frequency, amplitude or duration	Use automated analysis of slow wave activity

scoring methods by Danker-Hopfe⁵ found an increase in overall agreement from 80.6% to 82%. Ruehland found no improvement with the new methodology.⁶ The improvement in scoring agreement is more impressive when compared to historical values for R&K scoring, such as the 73% agreement among 5 experts from different centers reported by Norman and colleagues.⁷ Our data, like those of other researchers, indicates that the best agreement is achieved with stages W, N2, and R. Disagreement with the scoring of stage N1 includes scoring of stage W and scoring of stage N2 sleep. Disagreement with the scoring of stage N3 sleep is almost entirely based on confusion with scoring stage N2 sleep.

We expected that scoring agreement would be low in the epochs at the transition from one stage to another due to differences in recognition of key waveforms. This was not the case. An exception to this was the transition from stage N2 to stage R, which relies on complex scoring rules from the *AASM Manual*. The agreement on epochs in the transition from stage W to stage N1 was above average, with the exception of the final epoch of stage W before the change. The *AASM Manual* instructs the scorer to score stage N1 sleep when “alpha rhythm is attenuated and replaced by low amplitude mixed frequency activity for more than 50% of the epoch.”² Attenuation of alpha rhythm often occurs prior to its replacement with low amplitude mixed frequency EEG. Using one or the other definition might improve agreement in scoring of the transition to sleep. The transition from stage N1 to stage N2 again showed higher than average agreement for the epochs of stage N1. Scoring of stage N2 in this transition was less than average. This may reflect disagreement in the waveforms that define stage N2. Low amplitude or poorly formed K complexes or sleep spindles may lead some begin to score stage N2, whereas others wait for more definitive waveforms before making the change. In the absence of amplitude criteria for K complexes or spindles, this type of disagreement is inevitable.

Rules for the transition from stage N2 to stage R were developed by consensus.⁸ The rules hinge on the scoring of epochs with low amplitude mixed frequency EEG without K complexes, spindles, or rapid eye movements. These epochs may be stage N1, stage N2, or stage R, depending on the scoring of epochs that precede or follow them. The rules tip the scales in

favor of stage R; epochs between definite stage N2 and definite stage R are scored in most cases as stage R. Agreement in this transition might be increased by making the onset of stage R similar to the onset of stage N2. That is, stage R sleep might be defined as starting with the first epoch of rapid eye movements, low amplitude mixed frequency EEG, and low chin EMG tone. This would also simplify the scoring of stage N2. Stage N2 would end when the first epoch of any other sleep stage is scored. However, it could be argued that this would provide a less accurate reflection of underlying physiology.

The average agreement for epochs where the majority scored stage N2 sleep was 85.2%. Only 6.6% of scores for these epochs were for stage N3. However, the average agreement for epochs scored stage N3 by the majority was only 67.4%, and virtually all of the disagreement (32.3%) was with stage N2. The pattern at the transitions (**Figure 4**) does little to explain this disagreement. Consider the sample of 10 consecutive epochs shown in **Figure 5**. This record is from a 26-year-old man undergoing CPAP titration. With the age of the patient and conditions of the study, high amplitude slow waves and prolonged bouts of stage N3 sleep were expected. The pool of scorers, however, failed to agree on scoring of this sample. None of the epochs had more than 70% agreement for either stage N2 or stage N3. The high level of disagreement between scorers occurred despite the fact that the rule for scoring stage N3 sleep includes amplitude and frequency requirements, and is therefore more precise than most stage definitions. Slow waves must account for more than 20% of the epoch. But the measurement of wave duration and amplitude may be too complex and onerous for visual scoring. An alternative would be to use frequency analysis or some other automated method to perform this task. This would greatly increase reliability of scoring and might provide a better measure of the hypothetical process “S” as promoted by Borbély and colleagues.⁹

Table 5 summarizes the 5 types of epochs with significant disagreement and provides possible causes and potential changes to scoring methodology that might result in improved agreement. The table points out the inherent contradiction of encouraging additional criteria for amplitude of spindles and K complexes, while at the same time acknowledging the inaccuracy of scorers in measuring slow wave amplitude.

Perhaps the most interesting finding of this analysis is that in terms of agreement, a large group of scorers with varied backgrounds fared just as well as small groups of highly trained scorers. One interpretation of this finding is that a basic understanding of the rules for scoring is sufficient to produce competence, and additional training in a search for excellence may not be fruitful. Ambiguities in the scoring rules and variations in waveforms both across and within subjects may make further improvements in agreement difficult. Returning to the issue that prompted the development of the AASM ISR program, center managers may wish to set a relatively low cut point (such as 75%) that reflects basic stage scoring competence. A consistent score above this threshold could be used as a criterion for entrance to the pool of scorers. It is not recommended that a high threshold, such as 95%, be used as a criterion for promotion or bonuses. Such a criterion would be in excess of the agreement between expert scorers and is likely to occur only as a result of chance. It is important to note that this percentage agreement applies only to stage scoring. The other aspects of the ISR program (arousals, limb movements, and respiratory events) may have different levels of agreement. Further changes to the scoring rules, such as those proposed here, may also result in modest additional improvements.

Summary

A unique data set from the American Academy of Sleep Medicine Inter-scorer Reliability program was explored. More than 3 million individual stage decisions were analyzed.

Overall agreement was 82.6%, a finding similar to published literature comparing expert scorers with each other. Sleep stage transitions where there were at least 3 epochs of a sleep stage followed by at least 3 epochs of another stage were examined. Scoring of stage N1 sleep was better than average at these transitions. Only the last epoch of stage W prior to the onset of stage N1 had reduced agreement. The transition from stage N2 to stage R showed progressive disagreement prior to the shift of stage and gradual improvement after the shift. Disagreements in the transition from stage N2 to stage N3 and vice versa were high, with most of the disagreement in scoring of stage N3 occurring due to confusion with stage N2.

The authors propose changes to the scoring rules that might improve agreement: (1) changing the definition of stage N1 to require either the attenuation of alpha rhythm or the onset of low amplitude mixed frequency EEG but not both; (2) providing more objective criteria for key waveforms including alpha rhythm, K complexes, and sleep spindles; (3) changing the on-

set of stage R sleep to the first epoch of “definite” stage R; and (4) using alternative methods for analysis of slow wave activity, such as frequency analysis.

REFERENCES

1. Rechtschaffen A, Kales A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington, DC: US Department of Health, Education and Welfare Public Health Service -- NIH/NIND, 1968.
2. Iber C, Ancoli-Israel S, Chesson A, Quan S, for the American Academy of Sleep Medicine. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. Westchester, IL: American Academy of Sleep Medicine, 2007.
3. Grigg-Damberger MM. The AASM scoring manual four years later. *J Clin Sleep Med* 2012;8:323-32.
4. Moser D, Anderer P, Gruber G, et al. Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. *Sleep* 2009;32:139-49.
5. Danker-Hopfe H, Anderer P, Zeithofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res* 2009;18:74-84.
6. Ruehland WR, O'Donoghue FJ, Pierce RJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. *Sleep* 2011;34:73-81.
7. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep* 2000;23:901-8.
8. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. *J Clin Sleep Med* 2007;3:121-31.
9. Borbely AA. From slow waves to sleep homeostasis: new perspectives. *Arch Ital Biol* 2001;139:53-61.

ACKNOWLEDGMENTS

The authors appreciate the careful reading of the manuscript by Drs. Michael Silber, Joan Fisher, and Douglas Kirsch. The authors acknowledge the support of Jerry Barrett, the Executive Director of the American Academy of Sleep Medicine.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication September, 2012

Submitted in final revised form October, 2012

Accepted for publication October, 2012

Address correspondence to: Richard S. Rosenberg, PhD, American Academy of Sleep Medicine, 2510 N Frontage Road, Darien, IL 60561; Tel: (630) 737-9700; Fax: (630) 737-9790; E-mail: rrosenberg@aasmnet.org

DISCLOSURE STATEMENT

This is not an industry supported study. Richard S. Rosenberg and Steven Van Hout are full time employees of the American Academy of Sleep Medicine. The American Academy of Sleep Medicine developed the inter-scorer reliability program and provides salary support for the authors.