

## COMMENTARY

# Standing on the shoulders of giants: the *Standardized Sleep Manual* after 30 years

Max Hirshkowitz

*Sleep Disorders and Research Center, Veterans Affairs Medical Center & Baylor College of  
Medicine, Houston, Texas, USA*

---

*More than three decades ago, the Standardized Sleep Manual defined terms and established minimal guidelines for studying sleep in humans. The advent of inexpensive, widely available, powerful computer systems places a powerful "analytical engine" (to use Babbage's term) in most of our hands. When applied to polysomnography, the potential for extending our knowledge is vast. For computerized polysomnography to realize its potential, consensus will need to be reached on terminology, methods and quantitative procedures. This was the key to the success of the Standardized Sleep Manual and it will be the key to the success of current efforts to standardized automated sleep analysis. The Standardized Sleep Manual accomplishments and limitations are reviewed, as are the problems and strengths associated with computerized polysomnography. It is argued that rather than destroying or abandoning the Standardized Sleep Manual, it would be more productive to extend it. Automated analysis should stimulate evolution of our ability to describe and understand sleep. If we stand on the shoulders of giants we will be able to see new horizons.*

© 2000 Harcourt Publishers Ltd

**Key words:** computerized polysomnography, sleep stage scoring, automatic data processing, biomedical engineering.

---

In recent years it has become fashionable to criticize the *Standardized Sleep Manual* [1]. This is especially true among sleep scientists, clinicians, and engineers who are involved with computerized analysis of sleep bioelectrical activity. Proposals range from the desire to define additional sleep stages (to improve description in clinical populations) to wholesale condemnation with a wish to eradicate sleep staging for evermore. Like many others, I initially thought automating sleep stage scoring would be a simple matter of developing appropriate rule-based algorithms. The task turned out to be much more difficult because of ambiguities, artifacts and variations in human scoring. By contrast, the sophisticated period-amplitude signal analysis part of our system proceeded with little impediment and I soon found myself wanting to blame the scoring system for my inability to produce working code. I had clearly underestimated the task's complexity. Nonetheless, there clearly are limitations to the recording and scoring system described in the *Standardized Sleep Manual*, now more than 30 years

---

Correspondence to be addressed to: Max Hirshkowitz, Ph.D., VAMC-Sleep Center 116A, 2002 Holcombe Blvd., Houston, Texas 77030, USA. Tel: +1 (713) 794-7562; Fax: +1 (713) 794-7558.

old. However, I have come to believe it is extremely important to recognize the system's value and utility.

### Creating the standardized manual

The full title of the *Standardized Sleep Manual* is *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. It has become more commonly referred to today as the *R&K System* after the two editors, Dr Allan Rechtschaffen and Dr Anthony Kales. The committee, however, included a pantheon of scientists and clinicians. These individuals were driven by their interest in human sleep and its disorders. The group included Ralph J. Berger, William C. Dement, Allan Jacobson, Laverne C. Johnson, Michel Jouvet, Anthony Kales, Lawrence J. Monroe, Ian Oswald, Allan Rechtschaffen, Howard P. Roffwarg, Bedrich Roth and Richard D. Walter. In 1968, the United States Government Printing Office published the results of this *ad hoc* committee of sleep experts' deliberations. Brain Information Service (University of California, Los Angeles) later reprinted it.

The development of the *Standardized Sleep Manual* was pivotal in our field. One can only imagine what discussions took place. There are stories, perhaps even myths, about this epic event. One myth depicts Allan Rechtschaffen barring the doors and decreeing that no one could leave until they came to agreement. Some participants disagreed vehemently with the guideline recommendations for recording and scoring. It was argued that there were too few EEG channels, that chin EMG was questionably useful, and that delta amplitude criteria were arbitrary. But indeed, the success of the system pivoted on agreement. If each participant had returned to his laboratory and ignored the guideline in favor of continuing to do things according to their own practice, the project would have not succeeded. The critical ingredient was consensus, without which the attempt at standardization would have failed. Why was the *Standardized Sleep Manual* needed? Very simply, to avoid building a Tower of Babel. Laboratories were using different terms for the same thing; however, more dangerously we were also using the same term for different things. I can recall early in my career reading journal articles and being unsure about what results meant, unless I was familiar with the specific laboratory's procedures and terminology. As years passed and I became familiar with unique and sometimes idiosyncratic practices, findings that initially seemed at odds with one another were actually confirmatory. But standardization is not always a bed of roses; it has serious potential drawbacks.

### Danger of standardization

Charles de Gaulle is reported to have quipped "Any country that makes 365 types of cheese is ungovernable", later remarking that America has only one cheese and it is homogenized, homogenization being a procedure designed to keep the cream from floating to the top. So, we should ask, has the cream been prevented from floating to the top? Impeding progress is a major risk of standardizing too soon. Standardization can inhibit progress by stunting creative growth that ultimately leads to innovation. By contrast, however, confusion and ambiguity can also inhibit growth. Thus, when my colleagues claim that the *Standardized Sleep Manual* inhibited growth in the field of sleep research by defining things, they have a debatable point. However, my colleagues have also argued that the *Standardized Sleep Manual* inhibited growth because

it did not define certain things. This too is potentially correct. However, this is a "damned if you do and damned if you don't" situation. I suspect it is more likely that some other process is responsible for inhibiting growth of the field (if growth has really been inhibited at all). Nonetheless, the committee seemed aware of this potential pitfall and the *Standardized Sleep Manual's* foreword reflects their open-minded position:

"This proposal was prepared by the Committee with the expectation that the standardization of recording technique and scoring criteria would be widely used and would markedly increase the comparability of results reported by different investigators. An evaluation of how much such standardization contributes to reliability of scoring will have to await the development of experience with the system and empirical testing."

### Standardized sleep manual accomplishments

So what did the *Standardized Sleep Manual* attempt to accomplish? The most important outcome was creating a common language for clinicians and scientists to communicate about sleep. It defined a minimal recording technique, identified salient events and provided a system to describe and summarize sleep macroarchitecture. It allowed us to visualize a night of sleep in such a way that the patterns and alterations could be detected. The system focused on central nervous system activity; thus, breathing, leg movement and other physiological events were not defined. Table 1 summarizes sleep features the manual did and did not address.

**Table 1** Scope of the *Standardized Sleep Manual*

Defined or illustrated	Did not define or illustrate
Polysomnographic recording procedure	EEG arousals
Terminology	ANS arousals
Sleep staging in normal adults	Snoring
Quantitative sleep summary parameters	Sleep-disordered breathing events
Sleep EEG events (e.g. sleep spindle, sawtooth wave)	Sleep-related erections
	EKG events
	Electrodermal activity
	Leg movement events

Creating standardized terminology helped promote further exploration of human sleep staging. The *Standardized Sleep Manual* is sometimes regarded as just a set of scoring rules. In fact, the rules for scoring already existed, for the most part, as the Dement-Kleitman system and the Williams-Karacan system (which used frontal, central and occipital EEG channels). Some modifications, explications and simplifications were made in the interest of improving scoring reliability. A consensus agreement was reached, the work published, and the information was widely disseminated. Perhaps more importantly, the *Standardized Sleep Manual* defined terms and helped researchers deal with data overload. Calculation procedures for overall recording sleep summary parameters were established. We could now compare our data with others, compare

our data to normative values, and could begin to compare data in a standardized manner before and after experimental interventions. Additional parameters of interest certainly could be calculated and reported (as cutting edge scientists always have), but a minimal set of basic variables were recommended. Limitations were not imposed by the system but of course the committee did not foresee all of the needs of the next 3 decades. Who can? I fail to see the usefulness of criticizing trains because they don't float or automobiles because they can't fly. It is always easy to judge lack of foresight with the 20–20 vision of hindsight. For example, in 1981 I recall Bill Gates making the prediction about microcomputer memory "640K ought to be enough for anybody". We laugh now but at the time, coming from the small computer world of assembly, FORTH, GW BASIC, and C programming languages we had become used to 64–256K memory; therefore, 640K seemed reasonable. But this is beside the point. Rather than expounding or lamenting about how those old systems were limited, productive time was spent developing, building, testing, validating, and *demonstrating* superiority of the "next generation" system.

### Characterizing normal sleep

The intent of the procedures described in the *Standardized Sleep Manual* was to characterize *normal* sleep. This is a point that sometimes draws criticism. The argument is that because the system was designed to characterize the normal sleep pattern, it is questionably useful in the presence of sleep pathophysiology. Anomalies such as alpha-delta, excessive spindles, K-alpha activity, and spindle intrusion into REM sleep, are legitimate issues that render scoring problematic. Moreover, pathophysiological events associated with sleep disorders related to breathing and movements are outside the realm of the *Standardized Sleep Manual*; they are not limitations. The ambiguous EEG occurring in severe sleep apnea makes applying staging rules according to the *Standardized Sleep Manual* difficult at best. However, sleep stages typically normalize when treatment is provided. "Aha, back to normal" . . . a powerful conceptual statement. Thus, even in pathological conditions we need a way to characterize normal! It is claimed that a profound weakness of the *Standardized Sleep Manual* is that only EEG, EOG and EMG are recorded. Indeed, automobiles were not designed to fly, but is that a profound weakness? In sleep-related breathing disorders, the issue of limitation is not one of computerized versus manual scoring, it is an issue of recording respiration versus not recording respiration. You can score respiration manually if you want.

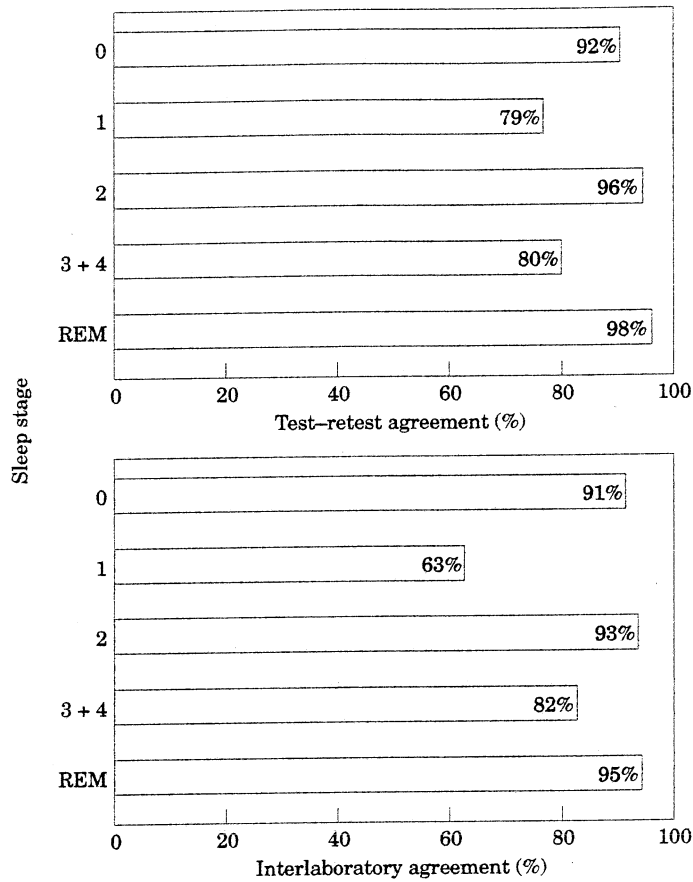
Does sleep staging characterize the EEG, EOG and EMG activity associated with normal sleep? To explore this question, let us consider an exercise I regularly conduct when training medical students, residents, and fellows in sleep medicine. I provide each student with approximately 100 individual pages from normal polysomnograms. Each page is one 30-sec epoch, four-channel (C3-A2 EEG, left eye, right eye, and submentalis EMG) epoch, recorded according to standard technique. I instruct students to sort the paper tracings into piles based on their judgement of similarities in the squiggly lines present on each page. No other direction is provided; that is, recording channels are not identified, paper speed is not indicated, page orientation is left ambiguous and the source of the recording is unstated. I have each student perform this assignment three times in order to provide familiarization and practice. When the third sorting is analysed, the result is similar more often than not. Students usually create 5–10 categories. With near universality, REM sleep is recognized. Two or three

categories of slow wave sleep are derived, with stage 4 sleep being the most consistent. Low voltage, mixed frequency epochs with sleep spindles (stage 2 sleep) are grouped together with great regularity. Wakefulness (stage W) is reliably clustered when recordings manifest clear alpha activity. The remaining epochs, those with stage 1 sleep, poorly formed or absent alpha activity, stage transitions, movement artifact, slow eye movements, snoring artifact (on the EMG channel) at sleep onset, and other anomalous activity are variably classified into several categories. However, the majority of polysomnographic pages are sorted into categories that roughly correspond to stages defined in the *Standardized Sleep Manual*. Therefore, I would speculate that if the *Standardized Sleep Manual* were developed tomorrow, with no knowledge of the past, it would bear remarkable similarity to the one we currently use, precisely because it characterizes the normal human sleep.

### Scoring reliability

Another criticism leveled at the *Standardized Sleep Manual* revolves around the issue of scoring reliability. Ironically, poor inter-scorer reliability was a major motivation for forming the *ad hoc* committee to develop the manual. A triad of objections to the scoring rules is raised in one of the other articles in this journal. One is that the "rules are difficult or impossible to follow". The rules are certainly not impossible to follow. Hundreds of American Board of Sleep Medicine certified sleep specialists, and thousands of registered polysomnographic technologists have been tested on these rules and apply them everyday. Certainly the rules are not simple; however, most individuals master them with sufficient practice and minor difficulty. To be sure, there are epochs that present ambiguity and both the experienced and inexperienced sleep stage scorer will find these difficult to classify.

It is also decried that the rules have never been validated. It is not made clear what would constitute validation. There is no standard against which the rules can be compared; they represent the defined standard. Stage scoring can be tested for reliability, and it has. In the paper by Karacan and colleagues [2], intra-scorer, inter-scorer and inter-laboratory reliability was assessed (see Fig. 1). After standardized training sessions at three different laboratories, respectable scoring agreement rates were found, especially for REM, stage 2 sleep and wakefulness. Much of the variance in stages 3 and 4 sleep categorization were misclassifications between those two stages. As expected, stage 1 sleep had the poorest overall consistency. In addition to published reliability studies, part of the American Academy of Sleep Medicine Standard for Accreditation of sleep disorders centers requires formal, periodic reliability testing of polysomnographic scoring (including sleep staging) as part of quality control. This requirement extends to all individuals (professional and technical) who interpret or score sleep studies. What possible tests could be performed to assess convergent or divergent validity is not clear. It is known that self-reported sleep latency, sleep efficiency, number of awakenings, and wake after sleep onset can differ from objective parameters. In some instances the self-reported versus objective parameter difference is so extreme that it is pathognomonic for the International Classification of Sleep Disorders [3] diagnostic entity "Sleep State Misperception". Additionally, even patients suffering from objectively documented insomnia commonly overestimate their latency to sleep and underestimate their



**Figure 1** Sleep stage scoring reliability. Intra-scorer and interlaboratory scoring reliability from data presented by Karacan *et al.* [2]. Three different laboratories participated after standardized training sessions were conducted.

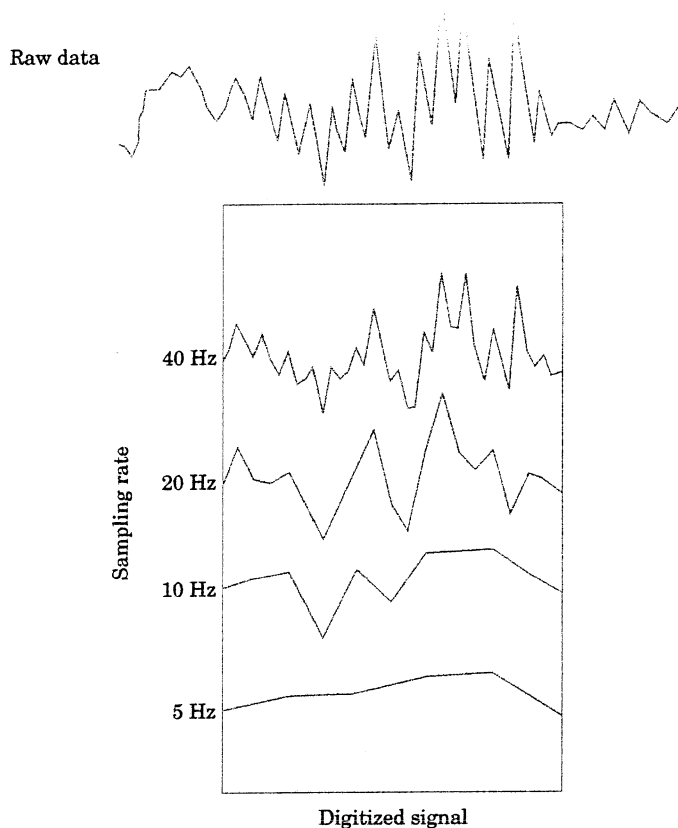
sleep efficiency. By contrast, young healthy subjects' estimations of sleep quantity and quality correlate well with objective parameters.

The final objection to *Standardized Sleep Manual* scoring reliability is that deviations are common, though not reported. Interestingly, I have not seen any statistics to support this contention. I suspect that unreported deviations may occur but I have doubts that such scientific fraud is *common* in the sleep research community. Furthermore, if an individual fraudulently represents their methodology, is it the fault of the scoring system? That would be akin to arguing that criminality is the fault of the law. Perhaps the notion is that the *Standardized Sleep Manual* is analogous to an unjust or unenforceable law and thus should be repealed. And in its place would be: nothing with a standardized recording method; nothing with an agreed analytical procedure; nothing with an established set of summary parameters; nothing that has a track record of utility; and nothing with normative values. I suspect most sleep specialists would say "No, thank you".

I find there are confusions in the logic that argues the problems associated with poor reliability of non-computerized scoring using the *Standardized Sleep Manual* will be solved if computerized scoring were used. To be sure, doing away with the *Standardized Sleep Manual* will solve any problems that may be associated with it. If a platform-independent, fully implemented, widely available, well validated, inexpensive, non-proprietary computer program capable of scoring sleep stages existed, then improved reliability and greater standardization would result. I have high hopes that such a program will eventually become available and have no doubt that a reliable program to score sleep stages will be written.

The notion that computerized analysis of sleep is somehow intrinsically reliable is laughable. That might be the case if identical programs were running on identical hardware with identical calibration. Otherwise, parameters such as delta amplitude would vary more than manually scored sleep stages. In fact, parameters may not be at all comparable between systems due to differences in cosine-ramp sampling windows, time versus frequency domain measures, artifact rejection techniques, assorted anti-aliasing strategies, sampling rates, digital filtering techniques, signal calibration, bandwidth definitions and overall analytical approach. Thus, unless a standardized computerized analysis is developed, a consensus is reached, the guidelines are properly followed and normative values are established, computerized sleep analysis will have equivalent or more variability than human scoring. It will also probably be more difficult to evaluate for purposes of quality control. I have used a variety of different computerized polysomnographic systems to analyse a benchmark set of FM instrument tape-recorded polysomnograms. This procedure was needed to validate the computer before using automated analysis in several studies our laboratory was conducting [4, 5]. Polysomnograms had to be hand-scored for microarchitectural parameters such as delta wave count, individual delta wave amplitudes, sleep spindle count, sleep spindle duration, sleep spindle peak amplitude, rapid eye movement count, slow eye movement count, alpha wave duration, K-complex count, and K-complex duration. This was an extremely time consuming, almost Herculean, endeavor. Differences between computer systems were larger than differences between scorers. However, test-retest reliability within a computer system was very high. The important implication of this is that while computerized microarchitectural parameters may be exquisitely sensitive and reliable within a system, comparison across systems and to universal clinical normative values will be problematic.

When it comes to computerized polysomnographic scoring, perhaps the worst *fly in the ointment* has to do with artifact recognition and rejection. Artifacts wreak havoc with computerized systems. Humans can note and adjust to a wide variety of artifacts occurring intermittently, periodically, or consistently. The clues to artifact origin often derive from coincidence with activity on other channels; for example, respiration artifact on an EEG channel or ECG artifacts on an EMG channel. The assortment of possible artifacts is large, some are quite rare, and we note new ones from time to time. Most recently we found a digital pager transmission artifact in one recording and I understand a recording exists of a California earthquake artifact on a Multiple Sleep Latency Test. A member of the technical support group for a particular system we were getting ready to evaluate once advised me that artifact-free recordings *were required* for proper function. I informed him that they needed to develop a computer system that would work in the real world.



**Figure 2** Digital signal produced by different sampling rates. A raw data sample of a sleep spindle was digitized at four different sampling rates. This figure illustrates digital filtering and signal amplitude reduction at lower sampling rates.

### Sampling rate

We surely are living in the digital age. Analog recordings are becoming a thing of the past. Digital representation actually means that the original information no longer exists. Thus, the sampling rate becomes one of the two most critical elements for recreating the data for inspection or analysis (amplitude resolution is the other). The higher the sampling rate, the more information is accurately preserved. However, the cost of a higher sampling rate is paid for in file size.

The Nyquist law concerning analysis indicates that sampling must occur at a minimum of twice the frequency of highest frequency waveform desired for analysis. This, however, is not adequate for reconstructing graphic representations. To provide a smooth image, sampling must probably exceed the highest frequency of interest by a factor of four or five. Thus, the proposed sampling rate of 100 Hz for sleep EEG would allow reasonable visualization of waveforms up to 20–25 Hz. By contrast, 40 Hz EMG would undergo obvious distortion. Why? Sampling intervals act as digital filters. For example, if you were only taking one sample every fifth of a second, a sleep spindle would be a completely missed (see Fig. 2). The sampling rate, because it acts

as a digital filter, even when it exceeds the highest frequency of interest, may still introduce significant amplitude attenuation. For patients with parasomnias where the differential includes nocturnal seizure, the proposed 100 Hz sampling rate is inadequate. Spikes occurring in the 100 Hz range would be missed. Similarly, EEG correlates of hypoxic seizure would likely be missed with 100 Hz sampling in patients evaluated for sleep disordered breathing.

### Time domain rigidity

Certainly one of the limitations imposed by the *Standardized Sleep Manual* is that it uses a rigid time domain. That is, stages broadly characterized events within a specific time frame of set duration. Thus, the event "sleep-onset" may not occur in the epoch of sleep onset. If sleep onset occurs in the last 5 sec of an epoch, the epoch is scored as wakefulness. Furthermore, assuming there is not arousal in the next epoch, the next epoch is scored as the first epoch of sleep. On paper polysomnograms, epoch beginning and ending is determined accidentally by where the paper tracing folds. This sometimes led to "framing errors" in attempts to compare human and computer scoring if the computer epochs were not precisely synchronized to paper records. More importantly, this time domain rigidity can gloss over important transitions in sleep state. Shorter epochs represent one alternative. Switching to an event-driven domain poses another alternative. Extending the scoring system and defining specific polysomnographic events of interest would also provide a solution. The real challenge will be to attain general consensus for such definitions.

### Creating a new standardized technique

Creating a new standardized technique for computerized polysomnography will be difficult. The well-reasoned four jobs for the computer discussed in another article in this journal serve as a much-needed starting point. As described, the processes include (1) chart writing, (2) technical note integration into the recording, (3) automatic data analysis and (4) report generation. If a consensus is reached on sampling speed and minimum amplitude resolution, chart writing standardization should pose little controversy. Specifically defined *back-paging* capability during continued data collection and storing pre- and post-calibrations with the recording are essential. The need to store permanently notes made by the night technologist with the polysomnographic recording is self-apparent. On traditional polygraphs, the technologist wrote their comments directly on the paper record. Any system that can satisfy the first two processes (chart writing and note integration) allows substitution of digital polysomnography for paper sleep study recording. For the most part, extant commercial computerized polysomnography systems satisfy these requirements. These tasks posed mainly engineering problems. The remaining two processes (automatic analysis and report generation) require charting new conceptual territory.

Standardizing automatic data analysis will require consensus agreement on the type of analysis to perform. There are diverging opinions concerning strengths and weaknesses of differing approaches. Time domain, frequency domain, demodulation and periodicity analysis all offer potential, exquisitely sensitive, techniques for extending our ability to describe human sleep. There is little doubt that micro-architectural analysis of the sleep process represents a step closer to underlying processes than macro-architectural analysis [6-8]. But what does it mean? Can I compare delta power from a

patient recorded using my computerized analysis with standardized normative values? The *Standardized Sleep Manual* was not designed as model of the sleep process, although it has certainly contributed to our understanding of sleep. If agreement about micro-architecture analysis can be reached, it could turn out to be a giant step forward for the field. Parameter definition will largely depend upon the analytical method adopted; thus, design of the report writer for micro-architectural variables will have to wait.

### **Extending the *Standardized Sleep Manual***

It is argued that sleep staging provides only a superficial description of sleep. Some investigators have suggested and debated possibly adding additional sleep stages. Whether such would increase our knowledge or would serve only to increase our specimen categories (like a butterfly collection) remains unresolved. Another approach has been to collapse the rigidity of the time domain. This provides identification and detection of transient stage changes that would otherwise be missed. This would be particularly useful for providing greater sensitivity to brief awakening. However, scoring EEG arousals could accomplish the same result. An alternate strategy would be to *extend* rather than *replace* the *Standardized Sleep Manual*. If the newer micro-architectural parameters prove more useful then one might expect them to eventually replace staging, leading practice to abandon the old system. In this manner our systems will evolve. The change will occur not because it was too difficult or we were too impatient to accurately automate sleep staging, not because we thought it was the direction to turn, and not because we became enamored by our own technology. Change will occur because it proves useful by living up to its potential. As with the work of the *ad hoc* committee, bold but well-reasoned consensus is needed. We cannot wait until everything is known before starting, otherwise we will never start. The evolution of technique will require taking a chance using our best wisdom, and some small measure of luck will not hurt. The *Standardized Sleep Manual* has survived 30 years and it is hard to think of comparable guidelines in medicine and research that have lasted that long. It is my view that we will do well to stand on the shoulders of the giants who created the *Standardized Sleep Manual* and to use its vantage point to look to new horizons. This seems more productive than spending our effort trying to slay a creature that unwittingly grew to the proportions of Goliath. If the new standardization effort is even half as successful as the past efforts of our field's pioneers, it will be a great accomplishment.

### **References**

- 1 Rechtschaffen A, Kales A, eds. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC: U.S. Government Printing Office, NIH Publication No. 204, 1968.
- 2 Karacan I, Orr WC, Roth T, Kramer M, Shurley JT, Thornby JL, Bingham SF, Salis PJ. Establishment and implementation of standardized sleep laboratory data collection and scoring procedures. *Psychophysiology* 1978; 15: 173-179.
- 3 Diagnostic Classification Steering Committee, Thorpy MJ, Chairman. *ICSD—International Classification of Sleep Disorders: Diagnostic and Coding Manual*. Rochester, Minnesota: American Sleep Disorders Association, 1990.
- 4 Hirshkowitz M, Thornby JL, Karacan I. Sleep pharmacology and automated EEG analysis. *Psychiatr Ann* 1979; 9: 510-520.

- 5 Hirshkowitz M, Thornby JI, Karacan I. Sleep spindles: pharmacological effects in humans. *Sleep* 1982; 5: 85-94.
- 6 Armitage R. Microarchitectural findings in sleep EEG in depression: diagnostic implications. *Biol Psychiatr* 1995; 37: 72-84.
- 7 Borbely AA, Tobler I, Loepfe M, Kupfer DJ, Ulrich RM, Grochocinski V, Dorman J, Mathews G. All-night spectral analysis of the sleep EEG in untreated depression and normal controls. *Psychiatr Res* 1984; 12: 27-33.
- 8 Kupfer DJ, Ulrich F, Coble PA, Jarrett DB, Grochocinski V, Doman J, Matthews G, Borbely AA. Application of automated REM and slow wave sleep analysis: 1. normal and depressed subjects. *Psychiatr Res* 1984; 13: 325-334.