The Reliability of Arousal Threshold During Sleep

MICHAEL H. BONNET, LAVERNE C. JOHNSON,
Naval Health Research Center, San Diego

AND WILSE B. WEBB
Department of Psychology, University of Florida, Gainesville

ABSTRACT

Thirty-five subjects from two independent studies were awakened at EEG-defined periods during the night with 1000 Hz ascending tone series. Awakenings were made five to eight times per night during stage 2, stage 4, or REM sleep over a series of nights in good and poor sleepers. Reliability was assessed within stage, within night, between stages, and between nights. Good and poor sleepers did not differ in either depth of sleep or reliability of arousal threshold and were thus pooled in the analyses. From night to night, the most consistency was seen in stage 4 ($r = .74$), although REM reliability ($r = .49$) and stage 2 reliability ($r = .50$ and $r = .69$ in the two respective studies) estimates were also greater than zero. Early sleep onset and morning arousals were more variable. Reliability estimates on arousal thresholds taken within the same night for stage 2 were $r = .64$ and $r = .77$ for the two studies and $r = .96$ for REM. The depth of sleep was not correlated with awake auditory threshold. It was concluded that five or six carefully placed arousals could give a good estimate of an individual’s usual arousal threshold.

DESCRIPTIONS: Depth of sleep, Auditory arousal threshold, Sleep, Measurement reliability.

The scientific study of the depth of sleep dates to the first experimental study of sleep in 1862 (Kohlschutter). Despite a long history, measures of behavioral responsivity during sleep hold an obscure and tenuous place in the annals of sleep research. Webb and Agnew (1969) dismiss them. Williams (1967) held hope for their value but only if the mass of possibly related variables could be controlled. By 1973 even Williams (Williams, Holloway, & Griffiths) had lost hope. It is fairly common knowledge that there is wide variability both between and within subjects in their depth of sleep (Snyder & Scott, 1972), but the existence of variability does not really comment on the usefulness or even the stability of the measure. In addition to these problems, only one source (an unpublished dissertation by Zimmerman in 1967) has discussed the issue of reliability. In that work reliability was deemed “sufficient” from two nights of observation, but no statistical analysis was attempted.

Arousal threshold as a measure has potential for giving information not available from EEG recordings. For example, depth of sleep has been used to evaluate the effects of stimulus meaning and sleep deprivation. Further, the measure has considerable face validity as a measure of depth of sleep and may be viewed as the point of transition between sleeping and waking states. In view of the possible value of arousal threshold, it is of importance to know whether it is a reliable measure. Therefore, data from two independent studies—one at San Diego and one at the University of Florida—were analyzed to determine the between- and within-night stability of arousal threshold.

This research was supported in part by Department of the Navy, Bureau of Medicine and Surgery, under Work Unit M0096-PN.001-1029, and by grants from Hoffmann-La Roche, Inc. to Drs. D. Ten Eyck and L. C. Johnson at the Naval Health Research Center, San Diego, and Drs. G. Barnard, W. B. Webb, and M. H. Bonnet at the University of Florida, Gainesville. The views presented in this paper are those of the authors. No endorsement by the Department of the Navy has been given or should be inferred.

Address requests for reprints to: M. H. Bonnet, Ph.D., Naval Health Research Center, San Diego, California 92152.
SAN DIEGO STUDY

Method

Twenty-six males aged 17 to 37 (mean 21.2), divided into groups of 12 good sleepers and 14 poor sleepers, slept in the laboratory for 2 screening nights and 5 baseline nights. Good sleepers reported no sleep complaints, had sleep latencies less than 30 min, and had less than 30 min of wake time after sleep onset on both screening nights. Poor sleepers described themselves as poor sleepers. In addition, on the 2 screening nights all poor sleepers actually had a sleep latency greater than 30 min, and 2 also had more than 30 min of stage 0 after sleep onset.

Subjects were awakened on the first, third, and fifth baseline nights with 1000 Hz tones delivered through a University Sound Model MCL 8 Ω speaker clamped to the headboard of the bed 45.7 cm above the subject’s head. Two-sec tones were delivered at 30-sec intervals in approximate 5 dB ascending steps starting from waking threshold, as measured each arousal night, and continuing until the subject pressed a button three times and said, “I’m awake.” Sound intensity levels were SPL (reference .0002 dynes/cm²). All subjects had passed Navy tests for normal hearing. Further, all subjects displayed waking thresholds between 35–40 dB in the laboratory.

Tones were presented 3 to 5 min after initial stage 2 onset, 5 min after stage 4 onset, 5 min after the beginning of the second REM period, 5 min after the following stage 2 period, and in the morning at the standard wake-up time. Because of the desire to obtain an arousal threshold during stage 4, the minimal sleep time required between first stage 2 and stage 4 arousal was 20 min. At least 30 min of sleep time intervened between all other experimental arousals. The first four arousals also were delayed for 5 additional min, or longer if the sleep stage was not firmly established, by the presence of movement artifact in the EEG.

On each night electrodes were attached as specified by Rechtschaffen and Kales (1968). Subjects retired at 2200 and were awakened at 0530 hrs. Each subject was instructed to press the button when he heard the tone, and the button press responses, as well as a signal marking stimulus onset, were written on the EEG paper as they occurred.

Results

Between-night Reliability

The reliability estimates reported are either Pearson $r$ (where two trials per subject existed) or ANOVA $r_{1}$, the Pearson $r$ equivalent (Guilford, 1954; Webb, Note 1) when more than two trials of threshold data existed. The total number of subjects for each correlation varied as a result of missing data. About half the data from the first stage 2 arousal was lost because, although the arousal series began in stage 2, subjects had often drifted into stages 3–4 before being aroused. Much of the data from the final morning arousal was not used in the reliability analyses because the nearness of body movements to time of threshold determination was not controlled (it was necessary to arouse subjects at 0530 to go to class). Arousals within 10 min of a body movement were found to occur at significantly lower dB values than those not so close to body movements ($\bar{X}_{\text{BM}}=58.9$; $\bar{X}_{\text{NoBM}}=72.76$; $t=2.616$). The actual number of subjects used in each analysis may be found in Table 1 along with the total number of observations entering into each correlation. Whenever the listed number of subjects was half of the total observations, Pearson $r$ was the statistic used. Whenever the listed number of subjects was a smaller fraction of the total observations, ANOVA $r_{1}$ was the calculated correlation. The observation/number of subjects ratio indicates the number of nights used, and that number reflects the number of nights on which that threshold information was collected. For example, for the initial stage 2 arousal, the two arousal threshold values for the 12 subjects who actually had been awakened from stage 2 on two different nights were correlated (Pearson $r$).

Initially the good and poor sleeper groups were compared for each of the five awakenings. No significant group differences in threshold were found for individual stage values or overall nightly means ($\bar{X}_{G}=74.5$; $\bar{X}_{P}=78.0$; $t=.421$). There were also no significant differences in reliability coefficients between the two. The good and poor sleepers were thus combined for the reported analyses. Further analysis of the differences between good and poor sleepers will be presented in a separate paper.

The reliability coefficients for data from the five awakenings correlated across nights can be found in Table 1. Little consistency was seen in the first stage 2 awakening, but the correlation of stage 4 thresholds was significant. Arousal threshold reliabilities for the REM and second stage 2 arousals were significantly greater than zero. The night-to-night correlation for the final arousal, which involved both REM and stage 2 awakenings with 9 subjects, was not significantly different from zero. The correlation for stage 4 was significantly larger than that for the first stage 2 arousal, but not significantly different from the reliability values for the other arousals. Because ANOVA estimators of reliability are not common in the sleep literature, the pairwise Pearson $r$ correlations were also computed for the second stage 2 awakenings and are included in Table 1.

Finally, the stage 4, REM, and second stage 2 threshold values were averaged within subjects for each of 2 nights. The correlation of the 2 nightly averages was significant, $r=.64$, which reflects the increased stability from use of three averaged observations.

1Significant means $p <.05$, two-tailed, in this paper.
# TABLE 1

**Between-night reliability**

<table>
<thead>
<tr>
<th>Sleep Stage</th>
<th>Number of Subjects</th>
<th>Total Observations</th>
<th>$r$ or $r_1$</th>
<th>Arousal Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean SPL</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td><strong>San Diego Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Stage 2</td>
<td>12</td>
<td>24</td>
<td>.22</td>
<td>68.6*</td>
</tr>
<tr>
<td>Stage 4</td>
<td>24</td>
<td>48</td>
<td>.74*</td>
<td>91.6</td>
</tr>
<tr>
<td>REM</td>
<td>25</td>
<td>75</td>
<td>.49*</td>
<td>83.4</td>
</tr>
<tr>
<td>Second Stage 2$^b$</td>
<td>26</td>
<td>78</td>
<td>.50*</td>
<td>70.4*</td>
</tr>
<tr>
<td>Morning (REM or Stage 2)</td>
<td>9</td>
<td>18</td>
<td>.44</td>
<td>67.1</td>
</tr>
<tr>
<td><strong>Florida Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Stage 2</td>
<td>6</td>
<td>24</td>
<td>.60</td>
<td>47.5</td>
</tr>
<tr>
<td>Stage 2 #1</td>
<td>6</td>
<td>24</td>
<td>.68</td>
<td>51.8</td>
</tr>
<tr>
<td>Stage 2 #2</td>
<td>6</td>
<td>24</td>
<td>.57</td>
<td>52.2</td>
</tr>
<tr>
<td>Stage 2 #3</td>
<td>6</td>
<td>24</td>
<td>.76</td>
<td>49.1</td>
</tr>
<tr>
<td>Stage 2 #4</td>
<td>6</td>
<td>24</td>
<td>.80</td>
<td>45.3</td>
</tr>
<tr>
<td>Mean</td>
<td>6</td>
<td></td>
<td>.69</td>
<td>49.2</td>
</tr>
</tbody>
</table>

*Significant ($p<.05$) from REM and stage 4.

$^b$Pairwise Second Stage 2 Correlations (all $p<.05$): Nights 1–3, .51; Nights 1–5, .43; Nights 3–5, .57;
Mean, .51.

*p$<.05.

## Within-night Reliability

The morning arousals which were not in close proximity to body movements were divided into REM arousals and stage 2 arousals. Fifteen subjects had an acceptable stage 2 arousal and 8 such a REM arousal on at least one of the nights. The arousal threshold values were correlated with the respective second stage 2 or the REM value for the same subject on the same night. Those correlations were .64 and .96 respectively. Both were significantly greater than zero, and the REM value was also greater than the REM reliability figure between nights ($z=2.36$).

Finally, the stage 4 threshold value, the REM value, and the second stage 2 value from each subject on the fifth laboratory night were entered into an ANOVA to get a reliability estimate over sleep stage (i.e., the extent to which someone with high stage 4 thresholds also had high REM and stage 2 thresholds). This correlation ($r=.55$) was significant.

## FLORIDA STUDY

### Method

Nine male subjects aged 21–23, selected for their normal sleep habits based upon responses on the Florida Sleep Inventory, spent from 1 to 4 non-consecutive nights in the laboratory. Waking threshold was tested with a Tracor RA 214 Rudmose screening audiometer, and all thresholds were within normal audiometric range (7±15dB). Subjects were awakened on each night with 1000 Hz tones delivered by the audiometer, which was built to ANSI 1969 specifications and calibrated by an audio specialist at the earphone insert before experiments were begun. Readings reported were in SPL (reference .0002 dynes/cm²) levels. Signals were presented in a 3-sec ‘on,’ 3-sec ‘off’ sequence in 2–5 dB steps. Both waking and sleeping thresholds were obtained from stepwise procedures which began with an ascending series and changed direction after tones became loud enough for the subject to begin responding or soft enough that he stopped. During sleep, an arousal was made 5 min into the first stage 2 period. Thereafter the following criteria were imposed for an arousal: at least 5 min into stage 2; at least 30 min since the last natural or experimentally produced EEG awakening; at least 10 min since the last transitory body movement of more than 6 sec. All arousals were made from clearly defined stage 2 sleep. The criteria were such that the first arousal always occurred before the appearance of stage 4 and subsequent arousals after the completion of the first stage 4 period. Arousals occurred throughout the night, but more than eight arousals per night were never attempted.

Electrodes were attached so that recording could be made from $F_1$–$F_7$, $P_1$–$T_5$, $O_2$–$O_2P_2$, and an eye channel. Upon entering the sleep room, subjects inserted the earphone into their ear and had a response button taped into
their hand. They were instructed to press the button whenever they heard the tone, and their responses, as well as a signal marking stimulus onset and offset, were fed through the EEG machine and written out as they occurred. Subjects retired at 2330 and were allowed to sleep for approximately 8 hrs.

Results

The Florida between-night results are summarized on the bottom of Table 1. These estimates were based on observations of the 6 subjects who had a 4-night protocol. The ANOVA $\bar{r}_1$ for the first awakening was .60. A minimum of four awakenings other than the initial awakening were made in stage 2 during the 4 nights, so an ANOVA was done on each of the 4 to give four strict between-nights estimates of $\bar{r}_1$. None of these estimates, including the overall average, was statistically significant, although correlations were all above .57.

Four threshold observations (the initial arousal was not included) were taken from within the same night for all 9 subjects. The $\bar{r}_1$ estimate ($\bar{r}_1 = .77$) for within-night stage 2 reliability was significant.

The reliability of auditory threshold of the subject when awake was $\bar{r}_1 = .92$. The waking threshold values were subtracted from the corresponding arousal threshold values to give a measure of "pure" depth of sleep. For both within individual subjects and between subjects, the two values, waking threshold and "pure" depth of sleep, were not significantly correlated ($|r| < .10$).

DISCUSSION

A measure of the resistance of a sleeping subject to arousal deserves consideration as a behavioral index of the sleep process. It is a crucial linkage point between the "unconsciousness" of sleep and sensitivity to the environment. The Florida data, which indicated that depth of sleep was not correlated to waking auditory threshold, imply that arousal threshold is a function of more than simply the processes which mediate waking audition. This conclusion is further supported by the fact that although a very strict EEG criterion was imposed, thresholds during sleep had a variability three times as great as during waking and such variability was present in both studies. Regardless of such variability, however, the data from the two studies support the general belief that people can be reliably divided into "light" and "deep" sleepers and that such a division is meaningful within sleep stage, within night, between sleep stages, and between nights.

Thresholds measured during stage 2 sleep showed a good deal of consistency when measured within the same night in both studies ($r = .64$, San Diego; $\bar{r}_1 = .77$, Florida) and slightly less consistency when measured across nights. The similarity in reliability is even more impressive when it is noted that there was a significant difference in arousal threshold from stage 2 sleep in the two studies ($t = 2.68$). This 20 dB difference, however, can be entirely accounted for by the difference in auditory thresholds of subjects measured when they were awake (waking thresholds were around 15 dB at Florida and around 35 dB at San Diego). This difference is most likely attributed to methodologies. For example, the physiology of the human auditory system is such that thresholds obtained from earphone apparatus are about 10-12 dB lower than those obtained from freestanding speakers (Yost, Note 2). Also, substantial learning effects were present in both studies and the Florida study increased the probability of a learning effect by 1) having pre-experimental practice on the threshold task, 2) scheduling arousals on each laboratory night, and 3) giving subjects more practice each night as a result of one more awakening and waking threshold determination at each awakening.

The data from the San Diego study indicated that the arousal threshold from stage 4 tended to have a higher reliability than the comparable measures from stage 2 and REM. This relative order is similar to that found in studies which have reported the reliability of sleep stage amounts from night to night (Webb & Agnew, 1969; Moses, Lubin, Naitoh, & Johnson, 1972; Clausen, Sersen, & Lidsky, 1974).

Several reports have found thresholds during both REM (Shapiro, Goodenough, & Gryler, 1963; Goodenough, Lewis, Shapiro, & Sleser, 1965) and stage 2 (Rechtschaffen, Hauri, & Zeitlin, 1966; Watson & Rechtschaffen, 1969) to be higher early in the sleep period than later in the night. In the present studies, some threshold curvilinearity was found (i.e. arousal thresholds at sleep onset and shortly before the normal end of the sleep period tended to be lower than those from other sleep times). These tendencies, however, can be explained to be a function of the nearness of the arousal to awakenings or body movements. The first arousal was about 5 min after initial sleep onset. While early morning stage 2 threshold in the San Diego study was significantly lower than the middle night stage 2 value, this difference disappeared when all arousals which occurred within 10 min of a body movement were dropped from the analysis. With body movements controlled, no significant time-of-night effects were found in either study for stage 2 or REM. Because the first REM period during the night was not tested, conclusions concerning REM must be qualified. However, it is possible that previous findings of significant within-stage shifts in threshold as a function of time of night may have resulted from the combination of the increase in
body movements across the night (see Kleitman, 1963) and minimal concern for postponing threshold determinations in close proximity to body movements (a 1-min criterion was used in the Rechtschaffen studies).

The lack of difference in arousal between good and poor sleepers in the San Diego study substantiates the findings of Zimmerman (1967, 1968), who found that few EEG or subjective characteristics, such as general good and poor sleep, differentiated light from deep sleepers. This lack of difference on a very basic dimension (sleep quality) indicates that the tremendous between-subject variability in depth of sleep poses a challenge for future studies.

The presence of multiple measurements during the night in both studies in stage 2 or in multiple stages made reliability estimates based upon several observations possible through the use of the Spearman-Brown Formula (Webb, Note 1). By use of this formula, an estimate was made of the increase in reliability for the Florida stage 2 data that resulted from using five observations recorded during each of 2 nights as compared to the usual single observation from 2 nights. The correlation, based on five observations, was estimated to be .92 versus the single observation value of .69. The implication is that by use of a sufficiently large number of observations, the measure reliability can be increased to almost any desired point. From the present stage 4, REM and stage 2 data, it appears that the results from five or six awakenings should be a fairly stable predictor of individual arousal threshold.

REFERENCES


REFERENCE NOTES

1. Webb, W. B. The reliability of sleep stage scores. 1974 APSS discussion available from Dr. Ismet Karacan, Baylor Medical School, Houston, TX.


(Manuscript received June 7, 1977; revision received December 8, 1977; accepted for publication December 19, 1977)