Sleep Accelerates the Improvement in Working Memory Performance

Kenichi Kuriyama,1,2 Kazuo Mishima,1 Hiroyuki Suzuki,1 Sayaka Aritake,1 and Makoto Uchiyama3

Introduction

Working memory (WM) performance, which is an important factor for determining problem-solving and reasoning ability, has been firmly believed to be constant. However, recent findings have demonstrated that WM performance has the potential to be improved by repetitive training. Although various skills are reported to be improved by sleep, the beneficial effect of sleep on WM performance has not been clarified. Here, we show that improvement in WM performance is facilitated by posttraining naturalistic sleep. A spatial variant of the n-back WM task was performed by 29 healthy young adults who were assigned randomly to three different experimental groups that had different time schedules of repetitive n-back WM task sessions, with or without intervening sleep. Intergroup and intersession comparisons of WM performance (accuracy and response time) profiles showed that n-back accuracy after posttraining sleep was significantly improved compared with that after the same period of wakefulness, independent of sleep timing, subject’s vigilance level, or circadian influences. On the other hand, response time was not influenced by sleep or repetitive training schedules. The present study indicates that improvement in n-back accuracy, which could reflect WM capacity, essentially benefits from posttraining sleep.

Key words: sleep; working memory capacity; memory consolidation; n-back task; skill learning; intelligence
We hypothesized that improvement in WM performance as measured by the n-back procedure could be facilitated by posttraining physiological sleep similar to that observed in other skill domains. In this study, we made a particular attempt to discriminate the possible effects of time elapsed after training, posttraining brain state (sleep or wakefulness), and circadian fluctuations in the possible role of posttraining sleep in the improvement of n-back task performance.

Materials and Methods

Participants
A total of 29 right-handed healthy subjects (mean age, 21.9 years; range, 19–26 years; 19 females) were randomly assigned to three different groups (described below). Subjects had no previous history of drug or alcohol abuse or of neurological, psychiatric, or sleep disorders, and were maintaining a constant sleep schedule. They were instructed to be drug-, alcohol-, and caffeine-free for 24 h before, and during, the study period. All procedures for the study were in accordance with the guidelines outlined in the Declaration of Helsinki. The study protocol was approved by the Intramural Research Board of the National Center of Neurology and Psychiatry, and all subjects provided written informed consent to participate in the study.

Working memory task
We used a spatial variant of the n-back WM task, which has been widely used to measure spatial WM with a sustained attention component (Gevins and Cutillo, 1993; Callicott et al., 1998, 1999; McEvoy et al., 1998), for all three experimental groups (groups A–C) (for details, see Fig. 1 and below). Subjects performed the n-back WM task with nine increasing levels of difficulty (n = 1–9), using a standard PC. Four large dots presented in a single row were displayed on the screen, indicating the four possible places where a stimulus could appear (Fig. 2). The stimulus consisted of one dot changing color. Subjects were instructed to respond by pushing one of four buttons on a response button box with the right fingers as quickly and as accurately as possible when the next stimulus appeared. The layout of the four buttons corresponded spatially to the four possible positions in which the stimulus appeared. Responses were to be made after a delay of n (load level) in n-back stimuli. The load level was shown before stimulation began throughout the entire experimental task. The different load levels were run in blocks of 20 + n stimuli; thus, 20 responses were obtained at each load level. The interstimulus interval was set at 500 ms, and each stimulus was displayed for 1500 ms; each block lasted a total of 41,500–58,500 ms. At each level, subjects performed three trials separated by 15,000 ms rest periods, with the scores being averaged at the end of the three trials. The stimuli were set in randomized order for each test session. Subjects completed all load levels (n = 1–9) three times in each test session (see below, Experimental design).

Performance measures
Performance was evaluated by using both the average percentage of correct responses (accuracy) and the average RT at each different load level. These provided measures of improvement in the throughput and the processing speed of the WM, respectively. The detection threshold for a given session was defined as the maximum n-back accuracy level (NL) at which the subject's accuracy exceeded at least 80%. RT was also calculated for each session.

Experimental design
The 29 subjects were assigned to the three experimental groups listed below, and each group underwent a specific schedule consisting of an initial training session and two retest sessions. Subjects performed a spatial n-back WM task (n = 1–9) in each test session. All morning retests were performed at least 1 h after awakening. Just before the initial training session, each subject performed the spatial n-back task (n = 0–4) to become familiar with the PC procedure. Retest schedules (Fig. 1) were as follows.

Group A: continued WM task training across wakefulness. To determine whether the simple passage of time (across wakefulness) led to improvement in WM performance, eight subjects (mean age, 22.3 years; range, 19–26 years; 5 females) were retested at 7 h intervals across the day after initial training at 8:00 A.M. (i.e., retests at 3:00 P.M. and 10:00 P.M.)

Group B: continued WM task training followed by 10 h wakefulness and then sleep. To determine whether subsequent sleep showed any marked improvement in WM performance over wakefulness, 11 subjects (mean age, 21.4 years; range, 19–26 years; 7 females) were trained at 12:00 P.M. (midday) and retested once at 10:00 P.M. after 10 h of wakefulness, and then again at 8:00 A.M. the next morning after a night of sleep.

Group C: continued WM task training followed by the immediate 8 h sleep and then wakefulness. To determine whether the improvement of WM performance required sustained wakefulness just after the initial training, 10 subjects (mean age, 20.2 years; range, 19–22 years; 7 females) were trained at 10:00 P.M. followed by an immediate 8 h sleep and then retested at 8:00 A.M. the next morning, and again later at 6:00 P.M. on the same day.

At each training and retesting point, all subjects performed a simple reaction task, which provided their simple reaction time (SRT), a standard measure of subjective alertness (Lorenzo et al., 1995; Corsi-Cabrera et al., 1996). The amount of overnight sleep for each subject in each experimental group was estimated using a self-recorded sleep log.

Statistics
Two-way factorial ANOVA was applied to detect the group and test-session differences in SRT performance. One-way factorial ANOVA was applied to compare the amount of sleep in the previous night of the experiment among three groups. The χ² test was used to compare gender distribution of the study subjects among the three groups.

Two-way factorial ANOVA was applied to detect the load level and gender differences in baseline n-back task performance in the three experimental groups (A–C), as well as to compare the improvement of n-back task performance among the three groups by 3 (experimental groups) × 3 (test sessions) and by 3 (experimental groups) × 2 (inter-sessions; retest 1 minus initial training vs retest 2 minus retest 1) comparison. After the analyses, we used one-way factorial ANOVA to detect the possible role of posttraining sleep in the improvement of n-back task performance. All ANOVA were followed by Bonferroni’s post hoc test. Results are shown as mean and SEM values. A p value of <0.05 (<0.0167 in Bonferroni’s post hoc analysis) was considered to indicate significance.

Results
Sleep quality and alertness
Two-way ANOVA revealed no significant differences in SRT within the three experimental groups (F(2, 28) = 1.920; p = 0.1535; 507.5 ± 18.0 vs 468.7 ± 12.3 vs 486.1 ± 10.4 ms for groups A–C, respectively) or within the three test sessions (F(2, 28) = 0.076; p =
either among load levels (2.39% vs 67.18% for load level 5 and load level 6). Two-way ANOVA revealed significant differences in accuracy for the load level revealed a significant decrement in accuracy (0.9267; 488.6 ± 11.5 vs 485.1 ± 11.4 ms for initial training, retest 1, and retest 2, respectively). In addition, no interaction in SRT between the experimental groups and test session (F(4,78) = 0.250; p = 0.9090) was found. One-way ANOVA detected no significant difference in the amount of sleep among the experimental groups (F(2,26) = 2.872; p = 0.0747; 7.31 ± 0.19 vs 7.55 ± 0.22 vs 6.85 ± 0.22 h for groups A–C, respectively). These findings indicate that there were no clear differences in vigilance level among the subjects in each test session.

**Initial training analyses**

We analyzed the difficulty profiles of accuracy and RT for each load level in the n-back task for each experimental group (A–C). Two-way ANOVA revealed significant differences in accuracy among load levels (F(8,294) = 114.1; p < 0.001) (Fig. 3), but not among experimental groups (F(2,294) = 2.905; p = 0.0567). No interaction was seen between load levels and experimental groups (F(16,294) = 0.954; p = 0.5086) in terms of accuracy. A post hoc test for load level revealed a significant decrement in accuracy between load level 5 and load level 6 (p < 0.0001), in that task difficulty gradually increased with an increase in trial number (n) up to 5; it then sharply increased at trial 6 and thereafter remained high.

Two-way ANOVA revealed no significant differences in RT either among load levels (F(8,294) = 0.304; p = 0.9641) (Fig. 3) or among experimental groups (F(2,294) = 2.520; p = 0.0826). No interaction was seen between load levels and experimental groups (F(16,294) = 0.122; p > 0.9999) in terms of RT.

**Gender effects on WM performance**

We therefore examined gender distribution in each experimental group. The χ² test revealed no significant gender distribution among the experimental groups (χ² = 0.138; p = 0.9331), suggesting that each group included almost equal gender distribution. Two-way ANOVA revealed no significant gender difference in accuracy (male vs female; 68.34 ± 2.39% vs 67.18 ± 1.84%; F(1,243) = 0.696; p = 0.4048), but there were significant load level differences in accuracy (F(8,243) = 108.3; p < 0.001) on the WM task, although no significant interaction was observed between gender and load level in terms of accuracy (F(8,243) = 0.908; p = 0.5106). Likewise, two-way ANOVA revealed no significant gender difference in RT (male vs female; 299.5 ± 13.4 ms vs 325.7 ± 13.0 ms; F(1,243) = 1.580; p = 0.2100) and no significant load level difference (F(8,243) = 0.245; p = 0.9817) in RT. Moreover, no significant interaction was seen between gender and load level in terms of RT (F(8,243) = 0.294; p = 0.9842). These findings indicate that there was no gender difference in initial training performance.

**Analyses for experimental group × test-session interaction**

Two-way ANOVA (3 experimental groups × 3 test sessions) showed significant group and test-session effects on NL (F(2,78) = 4.147, p = 0.0194; F(2,78) = 7.019, p = 0.0016; respectively) (Fig. 4), but no significant group × test-session interaction (F(4,78) = 1.453; p = 0.2246). Two-way ANOVA (3 groups × 2 intersessions) showed significant group effects on NL improvement (F(2,52) = 3.686; p = 0.0318) and a significant group × intersession interaction (F(2,52) = 15.857; p < 0.0001) (Fig. 5), but no significant intersession effect on NL improvement (F(1,52) = 0.248; p = 0.6205). A post hoc test for NL improvement revealed a significant difference between groups A and C (p = 0.0109), and a trend toward intergroup differences between groups A and B (p = 0.0491). These findings suggest that the three experimental groups showed different time profiles of NL improvement; NL improvement during posttraining sleep was significantly greater than that during wakefulness, and the acquired NL improvement seemed to be maintained for at least 10 h after posttraining sleep. As a result, average NL improvement in subjects who experienced posttraining sleep (groups B and C) was greater than that in subjects who went through the same period of wakefulness (group A) (Fig. 5).

Two-way ANOVA (3 experimental groups × 3 test sessions) showed neither significant group nor test-session effects on RT (F(2,78) = 0.719, p = 0.4904; F(2,78) = 0.027, p = 0.9738; respectively) (Fig. 4); moreover, no significant group × test-session interaction was observed (F(4,78) = 0.307; p = 0.8723).

**Group A: continued WM task training across wakefulness**

One-way ANOVA revealed no significant test-session difference in NL in group A subjects (F(2,21) = 0.218; p = 0.8058). Compared with the NL for the initial training (3.13 ± 0.72), we observed a subtle but not statistically significant improvement in NL at 3:00 P.M. (3.63 ± 0.71, by 16.0% vs initial training) and at 10:00 P.M. (3.75 ± 0.70, by 19.8% vs initial training) (Fig. 4A), suggesting that the simple passage of time across wakefulness produced no significant improvement in WM performance beyond that expected on the basis of continued rehearsal.
One-way ANOVA revealed significant test-session differences in NL in group C subjects (F(2,30) = 14.678; p < 0.0001). A post hoc test for NL revealed significant differences between the following test sessions: initial training versus retest 1 (3.30 ± 0.35 vs 4.90 ± 0.20; p = 0.0002) and initial training versus retest 2 (3.30 ± 0.35 vs 5.10 ± 0.28; p < 0.0001).

After a night of posttraining sleep, a significant increase in NL was apparent at 8:00 A.M. the next morning compared with the initial training scores (by 48.5%) (Fig. 4C). However, an additional 10 h of wakefulness produced no significant change in NL compared with the retest 1 scores (by 4.08%; p = 0.4577).

Discussion
Sleep-dependent facilitation of WM performance improvement
Although the NLs at the initial training session were comparable among all experimental groups or by gender, subjects demonstrated remarkably different time courses of subsequent NL improvements that were specifically dependent on the timing of posttraining sleep. Subjects trained at 12:00 P.M. (midday) demonstrated no significant improvement when retested after 10 h of wakefulness, but showed a significant improvement at 8:00 A.M. after a night of posttraining sleep (by 40.1% in group B) (Fig. 4). Similarly, subjects trained at 10:00 P.M. showed a significant overnight improvement (by 48.5% in group C) (Fig. 4), but no significant additional improvement during a further 10 h of wakefulness. Thus, significant improvements were acquired only across a night of posttraining sleep and not over a similar period of wakefulness, regardless of whether the time awake or time asleep came first.

The possibility that circadian factors confounded the learning profiles after 10 h of wakefulness is unlikely. The initial training session was similar for subjects trained at 8:00 A.M., 12:00 P.M., or 10:00 P.M., as was the case for objective ratings of alertness across all testing points. Thus, we consider sleep itself to be the most likely source of the improvement in NL on the n-back task.

RT has been considered to be a good indicator of skill performance improvement (Fischer et al., 2002; Walker et al., 2002; Kuriyama et al., 2004), but in the present study, it rarely seemed to reflect improvement in WM task performance. In our subjects, RT values varied independently of the comparative difficulty of the n-back task, which is in contrast to the values observed in the accuracy profiles for initial training. A previous study involving repetitive WM task training has also shown marked improvement in accuracy values over a 1–2 d period, although RT values improved slowly over a 4–5 d period (Olesen et al., 2004). Taking together, these findings suggest that RT may reflect different levels of improvement in WM task performance on the basis of accuracy values. WM performance is considered to be a result of plural cognitive processing (Gevins and Cutillo, 1993; Owen et al., 1996). The RT value possibly reflects the total skill performance of WM, whereas accuracy reflects the capacity limitation of WM.

Possible contribution of improvement in n-back accuracy to generalized improvement of WM performance
The results of recent investigations using a spatial variant of the WM task to examine the feasible number of items for both storing temporary and manipulating data converged on around three or four items (Luck and Vogel, 1997; Cowan, 2001; Saults and Cowan, 2007). The accuracy index for the n-back task has been established as a reflection of WM capacity in previous studies, and has been used for investigating individual or age variation in WM capacity (Oberauer, 2005; Mattay et al., 2006). Consistent with these previous findings, our subjects in all experimental groups showed NL around 3 or 4 at the initial training session. In the postsleep session, the NL for groups B and C increased up to around 5 and 6. Thus, the NL improvement was acquired across sleep, suggesting that dynamic changes in the WM process were executed during sleep, as has been observed in other cases of skill...
Some studies focusing on sleep-dependent skill learning have emphasized that the sleep-dependent gains seen in procedural learning (Stickgold et al., 2000; Walker et al., 2002; Kuriyama et al., 2004). Thus, the sleep-dependent benefit could not affect skill performance using other stimuli (Korman et al., 2003; Walker et al., 2003). Taatjes et al. (2003) showed that the sleep-dependent benefit seen in procedural learning is specific to various stimuli that drive cognitive demands. Trends Neurosci 23:475–483.

Conclusions

Together, these findings suggest that improvement in WM performance might not depend on the type of stimulus used, and that the sleep-dependent improvement in WM performance seen in the present study may lead to various improvements in WM performance, and furthermore, in the general capacity of WM. WM capacity is an important factor in a wide range of cognitive abilities, including general fluid intelligence (Conway et al., 2003; Colom et al., 2007), and our finding together with that of Jaeggi et al. (2008) suggests that posttraining sleep with appropriate timing could be a potent facilitating factor in WM training, leading to the advancement of individual general fluid intelligence.

References


Kyllonen PC, Christal RE (1990) Reasoning ability is (little more than) working memory capacity?! Intelligence 14:389–433.


