Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 4, 834–842

0278-7393/08/$12.00 DOI: 10.1037/0278-7393.34.4.834

Sleep Does Not Enhance Motor Sequence Learning

Timothy C. Rickard, Denise J. Cai, Cory A. Rieth, Jason Jones, and M. Colin Ard
University of California, San Diego

Improvements in motor sequence performance have been observed after a delay involving sleep. This finding has been taken as evidence for an active sleep consolidation process that enhances subsequent performance. In a review of this literature, however, the authors observed 4 aspects of data analyses and experimental design that could lead to improved performance on the test in the absence of any sleep consolidation: (a) masking of learning effects in the averaged data, (b) masking of reactive inhibition effects in the averaged training data, (c) time-of-day and time-since-sleep confounds, and (d) a gradual buildup of fatigue over the course of massed (i.e., concentrated) training. In 2 experiments the authors show that when these factors are controlled for, or when their effects are substantially reduced, the sleep enhancement effect is eliminated. Whereas sleep may play a role in protection from forgetting of motor skills, it does not result in performance enhancement.

Keywords: motor skill, sleep, consolidation, enhancement

Consolidation of procedural memory has been proposed to be separable into two stages: *stabilization*, which occurs during waking periods, and *enhancement*, which occurs during sleep (Walker, 2005). The stabilization phase corresponds roughly to memory consolidation as classically defined wherein a memory representation is made more resistant to interference and forgetting (McGaugh, 2000). Some studies of motor skill tasks (Brashers-Krug, Shadmehr, & Bizzi, 1996; Muellbacher et al., 2002) have confirmed that waking periods between training and retest (from 4 to12 hrs) can stabilize performance, although that result is not ubiquitous (Caithness et al., 2004; Goedert & Willingham, 2002; Robertson, Press, & Pascual-Leone, 2005).

There is also empirical support for sleep-based enhancement for both perceptual and motor skills (for review, see Marshall & Born, 2007; Stickgold, 2005; Walker, 2005; Walker & Stickgold, 2004, 2006). For example, in one study, researchers used a sequential finger-tapping task. Participants were given 12 training blocks (each block consisting of 30 s of key-presses interleaved with 30-s rest) either at 10 a.m. or 10 p.m., and they returned 12 hrs later for 2 test blocks (e.g., Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002). In a comparison of the averaged performance on the last 2 blocks of training to that on the 2 blocks of test, the overnight-sleep group exhibited a substantial performance improvement on the test, whereas there was no improvement for the daytime-awake group. Similar performance gains following sleep have been shown to hold for delays of 24 and 72 hrs (Walker, Brakefield, Hobson, and Stickgold, 2003; Walker, Brakefield, Seidman, Morgon, Hobson, & Stickgold, 2003). These and related results have been taken to reflect learning that takes place during one or more stages of sleep (e.g., Nishida & Walker, 2007; Walker et al., 2002).

However, the proposed enhancement phase, and the claim that it occurs exclusively during sleep, is not without controversy (for reviews, see Frank & Benington, 2006; Robertson & Cohen, 2006; Vertes, 2004; Vertes & Siegel, 2005). Although overnight improvements in performance have been observed consistently for finger-tapping tasks and for visual discrimination tasks, it has not been observed consistently for rotary pursuit and arm-reaching tasks wherein test performance following sleep has in some cases declined from peak training performance (Adams, 1952; Donchin, Sawaki, Madupu, Cohen, & Shadmehr, 2002). Some evidence also suggests that enhancement is not limited to delays involving sleep (Brashers-Krug et al., 1996; Fischer, Hallschmid, Elsner, & Born, 2002; Robertson, Pascual-Leone, & Press, 2004; Spencer, Sunm, & Ivry, 2006). In the case of implicit skill learning, new evidence suggests that the enhancement effects may reflect a time-of-day confound (Keisler, Ashe, & Willingham, 2007).

In our review of the work on explicit motor sequence learning, we noted potential limitations with respect to both data analyses and experimental design that raise the possibility that sleep may play a substantially different role in consolidation than is implied by the two-stage model. In Experiment 1, we explored the possibility that the data-averaging done in prior studies results in illusory sleep enhancement (i.e., in a performance enhancement effect that is not related to sleep consolidation). In Experiment 2, we explored whether two experimental design factors—the confounding of time-of-day with the wake-sleep manipulation and the massing of practice during the training session—may also contribute to illusory enhancement. We show that when these data analyses and experimental design factors are eliminated or are substantially mitigated, there is no enhancement effect. It appears that,

Timothy C. Rickard, Denise J. Cai, Cory A. Rieth, Jason Jones, and M. Colin Ard, Department of Psychology, University of California, San Diego.

We thank Ali Sultan, Allen Kim, Angela Kuzara, Ashley Robb, Athar Haq, Blair Weaver, Christine Olandj, Christina Reno, Elizabeth Hahn, Emily Cheung, Joseph Lee, Kim Nakashima, Lisa Hecht, Samantha Yellen, and Sumeet Gupta for assistance with data collection; Daniel Bajic for assistance with programming; and John Wixted for comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Timothy C. Rickard, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: trickard@ucsd.edu

although sleep may play a role in protection from forgetting (i.e., stabilization) for the explicit sequence learning task, it does not give rise to performance enhancement; that is, it does not play an active role in learning.

Experiment 1

One characteristic of the analyses in all studies to date is substantial data averaging (1 min or more of task performance in the comparison of performance at the end of the training session with performance in the test session). This averaging could, in principle, result in two separate types of biases, either or both of which could give rise to an illusory enhancement effect. First, averaging may mask performance changes that are occurring during the training and test blocks (Robertson, Pascual-Leone, & Miall, 2004; Vertes, 2005). Consider the possibility that performance is actually worse at the beginning of test compared to at the end of training due to partial forgetting for all of or some component of the task, but the rate of improvement during test is greater than would be expected on the basis of an extrapolation of training session performance, as has been observed by Rickard (2007) for cognitive skills. If this is the case, then in the averaged data the initial slowing on the test may be masked, yielding an apparent immediate enhancement. Second, averaging of data over the last 1 min or more of the training session could result in an underestimate of actual achieved skill level due to fatigue effects that may build up over the course of the 12 training blocks. Of particular focus in Experiment 1 is the possibility that performance at the end of the training session suffers from reactive inhibition effects (Hull, 1943) that are behaviorally expressed, at least in part, as a progressive worsening of performance within each 30-s training block. This possibility is suggested by the results of Fischer et al. (2002), who gave participants continuous finger-tapping practice over a series of 5-min blocks and observed substantial worsening of performance over the course of each block followed by an immediate improvement in performance at the beginning of each new block. If this within-block reactive inhibition effect is more pronounced toward the end of the relatively long-duration training session than in the brief 2-block test session, it could result in illusory enhancement in the averaged data.

This experiment is a close replication of Walker et al.'s (2002). We show that when data are averaged in the usual way, there is an apparent sleep enhancement effect. Fine-grained analyses with minimal data-averaging and minimal influence of reactive inhibition, however, reveal a different pattern that could easily be overlooked. Instead of performance enhancement in the sleep condition, there is performance slowing (forgetting) in the awake condition.

Method

Participants. Fifty-three young adult students from the University of California, San Diego participated for course credit or for pay. All participants were right-handed.

Design and procedure. The sequential finger-tapping task required participants to repeatedly complete, with their left (nondominant) hand, the sequence 4–1–3–2–4 (numbering the fingers from little to index) by pressing the keyboard keys *z*, *x*, *c*, and *v*, which were labeled *1*, *2*, *3*, and *4*, respectively. Each block consisted of 30 s of key presses followed by 30-s rest. The training session consisted of 12 blocks, and the test session consisted of 2 blocks, just as in the Walker et al. (2002) study. Participants were tested on IBM-compatible personal computers programmed with E-Prime software (Psychology Software Tools, Pittsburgh, Pennsylvania). The numeric sequence (4–1–3–2–4) was displayed at the top of the screen at all times to exclude any working memory component to the task. Each key press produced a white dot below the correct digit, forming a row from left to right over the course of each key-press sequence.

Participants were randomly assigned to either the awake or the sleep group. The training session began within 1 hr of 10 a.m. or 10 p.m. for the awake and sleep groups, respectively, and participants returned for their testing session exactly 12 hrs later. Immediately after the second session, participants were administered the Stanford Sleepiness Scale (Hoddes, Zarcone, Smythe, Philips, & Dement, 1973) and completed a questionnaire in which they reported hours slept the night before the test session, whether they napped between sessions, and whether (and for how long) they practiced between sessions. In this experiment and in Experiment 2, participants were not told that sleep was relevant to the experiment and were not aware that the task would be the same for both sessions. Because our laboratory is not known on campus for studying sleep, these experiments should have high ecological validity. In informal postexperiment debriefing, no participant reported suspecting that sleep was a factor in the experiments. These facts minimize the possibility that participants in the sleep groups might perform more diligently at test to satisfy perceived demand characteristics.

Results and Discussion

Thirteen participants reported practicing between sessions and 7 additional participants in the awake group reported napping between sessions. One participant in the sleep group reported only 3 hrs of sleep between sessions. These participants were removed prior to analysis, leaving 16 "clean" participants in the awake group and 16 clean participants in the sleep group. Among these participants, there was no significant group difference on the Stanford Sleepiness Scale (which was given after the test session), and the mean values for the awake and sleep groups were 2.47 ($SD = 1.28$) and 2.50 ($SD = 1.38$), respectively. Participants in the sleep group reported having slept a mean of 6.33 hrs ($SD = 1.27$).

We used two measures of accuracy. The first was simply a determination of whether the participant's key-press response matched the correct response on each trial (we henceforth refer to each key press as a trial). Preliminary inspection of the data, however, revealed that the responses for some participants were, during one or more intervals of practice, correct with respect to the five-key sequence but offset by one or more responses from the correct response. This phenomenon created series of trials that were all incorrect by the measure described above but that nevertheless reflected accurate sequencing. To correct this, we created a second adjusted error measure in which each correct sequence of five trials was recorded as correct, even in cases in which the responses themselves were offset by one or more keys. Because this measure appears to provide a more accurate reflection of the actual participant performance level, it was used as the basis for removing incorrect trials prior to all response time (RT) analyses.

For the awake group, mean trial accuracy was .980 (.983 for the modified accuracy measure described above) and .968 (.985) for the last two blocks of training and for the two blocks of test, respectively. For the sleep group, these values were .964 (.984) and .891 (.983). For the modified accuracy measure, accuracy was nearly identical for both groups and sessions. This null effect of errors for the sleep condition contrasts with some of the prior experiments on the motor sequence task wherein a significant decrease in errors following sleep has been observed (Walker et al., 2002).

Prior to analysis, all RTs were log transformed, a procedure that minimizes the influence of outlier RTs and yields an approximately normal distribution. We present antilogs of the means of the log RTs to facilitate interpretation. Results were materially the same when analyses were performed on the raw RTs. Correct trial RTs are shown in Figure 1A as a function of group, practice block, and session. A two-sample t test on the average RT difference scores (mean of the last two blocks of the training minus the mean of the two blocks of the test; Figure 1B) was statistically significant, $t(30) = 2.48$, $p = .019$, reflecting the same Group $\times$ Session interaction observed by Walker et al. (2002). Post hoc one-sample t tests on the difference scores indicated a significant enhancement effect for the sleep group, $t(15) = 2.17$, $p = .046$, but no significant effect for the awake group, $t(15) = -1.37$, $p = .190$.[1]

Fine-grained analyses, however, suggest that the apparent enhancement effect in the sleep group is a consequence of data-averaging.

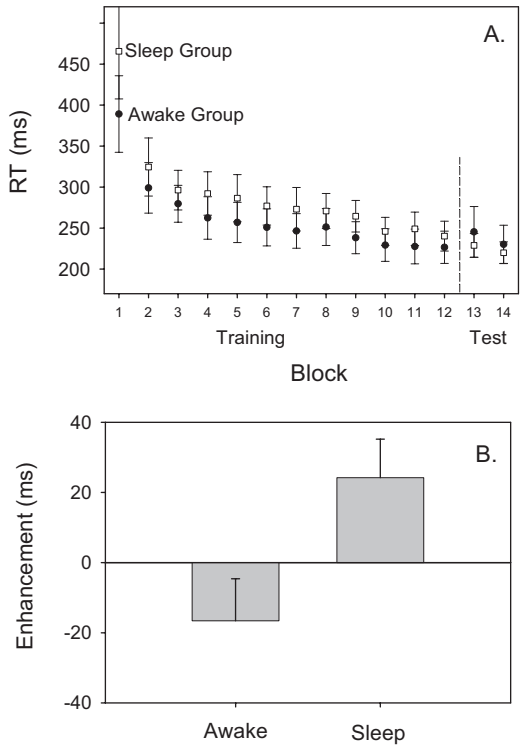

Figure 1. Panel A represents mean response time (RT) in Experiment 1 as a function of group (awake, sleep), session, and block. The vertical dotted line separates training and test sessions. Panel B represents RT enhancement effect (mean RT for the last two blocks of training minus mean RT for the two test blocks) for Experiment 1 plotted by group. Error bars represent the standard error of the mean for each data point.

These analyses were based on the mean trial RTs for each five-key sequence, as shown in Figure 2, averaged over group (these patterns were materially identical for the two groups), and plotted as a function of session and sequence number for Blocks 1, 2, and 3, and for the average of Blocks 7–12 (patterns were materially identical over these blocks). For each block, only the first 12 sequences, which most participants completed, are included.

For the first two blocks of both the training and test sessions (Figure 2A, Figure 2C), mean RTs decreased as a function of sequence on a typical decelerating learning curve, indicating substantial learning within those blocks. Thus, the strategy of averaging over two or three blocks of test performance, as in the analyses above and in earlier studies, does not allow the issue of whether there is immediate enhancement following sleep to be appropriately addressed.

For the data from the last half of the training session (Blocks 7–12; Figure 2B), there is a distinctly different pattern. The RTs for the first sequence are relatively large; they decrease to their smallest values for about Sequence 2–4 and then increase in a linear fashion. In a repeated measures general linear model analysis with a continuous factor of sequence performed on the data from Sequences 2–12, the linear trend was highly significant, $F(10, 310) = 8.6$, $p < .0001$. The large RT for the first sequence in Blocks 7–12 appears to reflect a warm-up effect, and therefore that sequence is best eliminated from each block in subsequent analyses. Data from Sequences 5 onward for those blocks presumably reflect within-block fatigue or reactive inhibition, similar to that evident in the Fischer et al. (2002) data. Similar patterns begin to emerge in Blocks 4–6 (not shown in Figure 2). For participants who completed more than 12 sequences per block, this trend of a linear increase in sequence RTs continued. Inclusion of Sequences 5 and beyond in the averaged data therefore introduces a bias which, we argue, also needs to be removed to provide the most accurate possible measure of actual achieved performance toward the end of the training session.

We thus performed a supplementary RT analysis on the basis of the average of Sequences 2–4. To summarize, these sequences were selected on the bases that (a) they occur early within the practice block and are thus, by definition, least subject to buildup of within-block reactive inhibition and to within-block learning effects, (b) during the later portion of training, RTs for these sequences were the fastest (Figure 2B), supporting the assumption that they are least affected by fatigue (or warm-up), and (c) there was no hint of a visual trend toward increasing RTs for Sequences 2–4 during the later part of training, again supporting the assumption of minimal influence of reactive inhibition for those sequences. Note that qualitatively the same results are obtained when analyses are limited to Sequences 2, 3, or 4 but with more variability in the data.

The selection of Sequences 2–4 is based on averaged data and may mask individual differences in reactive inhibition. We ex-

[1] Supplementary analyses with the Walker et al. (2002) dependent measure—the mean number of correctly completed trial sequences per block for the last two blocks of the training session versus the first two blocks of the test session—yielded analogous results. For the awake group, the means were 22.5 and 22.3 for the training and test sessions, respectively. For the sleep group, these means were 22.6 and 24.1.

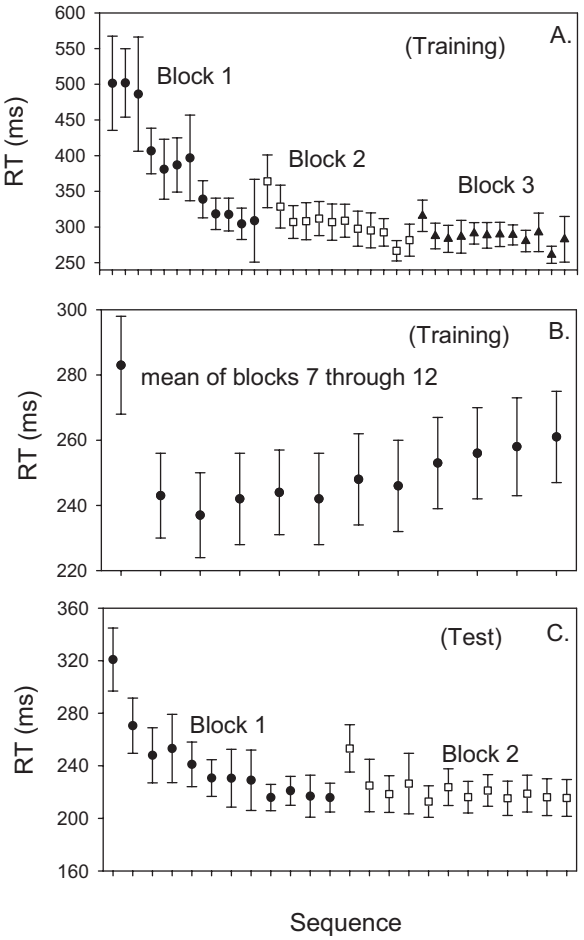

Figure 2. Mean response time (RT) in Experiment 1 as a function of sequence number and training block, averaged over participants and groups. Panel A shows data form the first three blocks of the training. Panel B shows data averaged over Blocks 7–12 of training. Panel C shows data from the two test blocks. Filled circles represent awake group. Empty squares represent sleep group. Error bars represent the standard error of the mean for each data point.

plored this possibility with individual participant analyses equivalent to the grand mean analysis in Figure 2B. Although there was more noise in these participant-level plots, most participants showed a trend toward increasing RTs over sequences similar to that of Figure 2B. Twenty-five participants had positive regression slopes (11 were statistically significant) whereas 7 participants showed negative slopes (only 1 was significant, and all had a magnitude smaller than the mean for participants who exhibited positive slopes). The negative slopes for those 7 participants do not necessarily indicate that they had no reactive inhibition but rather may indicate that the reactive inhibition was not sufficient to fully mask within-block learning effects.

Our goal in this selected sequence analysis with respect to the training data was to better approximate actual achieved performance level. For the participants who exhibited the negative slopes above, selection of Sequences 2–4 may not be optimal for achieving that goal. The only material consequence for the following analyses of the averaged data is that reactive inhibition may still be exerting some influence in the direction of producing illusory sleep enhancement for some participants.

Results are plotted as a function of group, block, and session in Figure 3A, along with the best fitting 3-parameter power functions fitted to Blocks 2–12 of the training session and extrapolated to the test session.[2,3] A two-sample *t* test on the difference scores (Figure 3B) again indicated a significant Group × Session interaction, $t(30) = 2.46$, $p = .02$. It is critical to note, however, that the interaction in this case reflects slowing at test for the awake group, $t(15) = 2.81$, $p = .013$, rather than enhancement for the sleep group, $t(15) = 0.31$.

Extrapolation of the power-function fit from the training session to the test session (Figure 3A) yields the most accurate measure of the expected performance had participants practiced for 14 instead of 12 blocks in the training session; that is, it allows for predictions that adjust for expected learning effects due to continued practice. In the sleep group, RTs for both test blocks were within the range of statistical error.

Although the selected sequence analysis in Figure 3 suggests that the putative sleep enhancement effect in motor sequence learning is, in part at least, a consequence of data-averaging, the relative advantage for the sleep group remained (Figure 3B). Our results are therefore consistent with the possibility that sleep affords protection from the forgetting of motor sequence skill. The data also suggest, however, that the time-of-day and time-since-sleep confounds inherent to this design may be at least partially responsible for the relative advantage for the sleep group. If participants generally perform better on motor sequence tasks at 10 a.m. (or soon after waking) than at 10 p.m. (or after having been awake for many hours), then a Group × Session interaction could be obtained even if there is no effect at all of sleep on performance. A comparison of the training session RTs for the two groups is consistent with this possibility (see Figure 1A). Training in the morning (awake group) yielded faster grand mean RTs (264 ms) than did training in the evening (sleep group; 284 ms). This effect, however, did not approach statistical significance.

It is nevertheless instructive to consider the possibility that in the population there is a small RT advantage for morning performance. It turns out to be much more likely that such an effect would lead to a significant Group × Session interaction than that it would lead to a significant effect of group in the training session.

[2] Power-function fits that included the first practice block produced nonrandom residuals as indicated by the Wald–Wolfowitz test, indicating that the data from the first block are inconsistent with power-function learning. This result may reflect (a) the fact that power-function learning has generally been found when performance is plotted as a function of practice trial, whereas these data are plotted as a function of constant time intervals (with each successive interval containing more trials) and/or (b) a rapid shift from declaratively driven to procedurally driven performance, which also results in deviation from power-function mean speedup (for related results, see Rickard, 1997, 2007). Because the goal here was solely to fit the data from practice as well as possible for purposes of making a meaningful extrapolation to the test session, data from the first block of practice were not included in the fit.

[3] The exponential function has been shown to fit best to individual data and the power function to fit best to averaged data (Heathcote, Brown, & Mewhort, 2000). We confirmed this pattern for our data. Because our analyses are for averaged data, we used the power function.

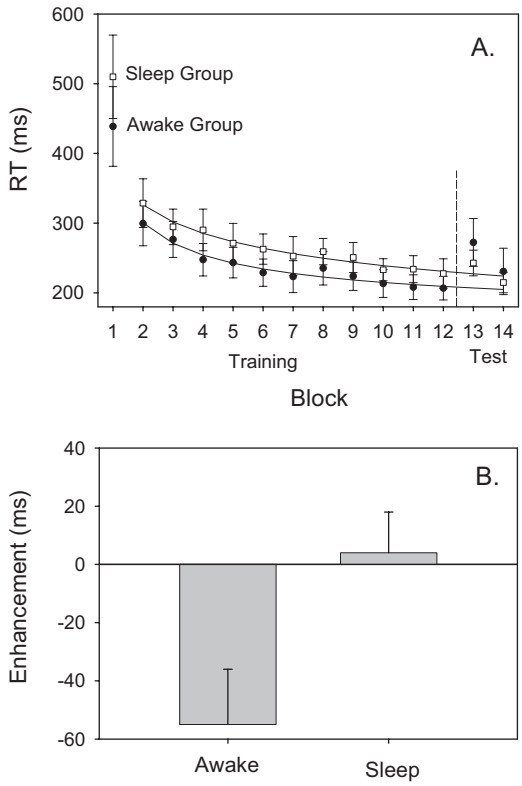

Figure 3. Panel A represents mean response time (RT) for Sequences 2–4 in Experiment 1 as a function of group (awake, sleep), session, and block, along with best three-parameter power-function ($RT = a + b^{*}N^{-c}$) fits to Blocks 2–12 of the training session and extrapolated to the test session. The vertical dotted line separates training and test sessions. Panel B represents RT enhancement effect (mean RT for the last two blocks of training minus mean RT for the two test blocks) for Sequences 2–4 of Experiment 1 plotted by group. Error bars represent the standard error of the mean for each data point.

For purposes of this discussion, we assume that in the population there is a 20-ms time-of-day (or time-since-sleep) advantage for morning versus evening performance, consistent with results for the grand mean training RTs discussed above. In light of the observed between-subject variance, the retrospective power to detect this effect in a two-sample *t* test (we used an alpha .05, one-tailed test) is only .16. Now consider the power to detect a Group × Session interaction, under the assumption that the 20-ms time-of-day effect is the only factor driving that interaction. Note first that any time-of-day (or time-since-sleep) effect is doubled in magnitude in the interaction. It would, as the sole factor in play, result in an expected 20-ms slowing for the awake group on the test and an expected 20-ms enhancement for the sleep group on the test, yielding an overall 40-ms difference. In addition, statistical power to detect the time-of-day effect in the interaction is improved by the fact that that interaction has the within-subject variability (variability in difference scores) as the error term. Again we performed a retrospective power analysis, framed as a two-sample *t* test on the difference scores using the variability estimates from the data. Power to detect the expected 40-ms effect is .84. Thus, if there is a small time-of-day effect in the population, a likely outcome is that that effect would not be significant in the group comparison of the training data but would lead to a significant Group × Session interaction.

If time-of-day effects are the dominant factor underlying the interaction, then we might expect the magnitude of the RT advantage for the awake group at the end of training to be roughly the same as the magnitude of the advantage for the sleep group on the test. Indeed, inspection of Figures 1A and 3A supports that prediction.

Time-of-day cannot fully account for the wake–sleep manipulation effects in the explicit sequence learning literature, however. Fischer, Nitschke, Melchert, Erdmann, and Born (2005) conducted a sleep deprivation study using a finger–thumb opposition task very similar to the current one, in which there was no time-of-day confound. In their study, both groups were trained at 10 p.m. One group was sleep deprived on the first night after training. Both groups were tested 48 hours after training. They found relatively better test performance in the control condition than in the sleep deprived condition.[4] Thus, although the time-of-day effect appears to account for a portion of the Group × Session interaction, it is apparently not a sufficient explanation.

Experiment 2

Although the selected sequence analysis of Experiment 1 is a reasonable strategy for avoiding data-averaging effects, the experiment is still not ideal for evaluating the sleep enhancement hypothesis for several reasons. First, the selected sequence analysis ignores 88% of the training and test data. A preferred approach would circumvent the need to eliminate data prior to analyses. Second, as noted above, there are potential time-of-day and time-since-sleep confounds that could, for the sleep group in Experiment 1, contribute to an illusory sleep enhancement effect (or could mask forgetting between sessions). Third, there is the unexplored possibility of a buildup of performance fatigue over the course of training blocks (above and beyond the within-block reactive inhibition effect) that dissipates during the delay between sessions, an additional factor that could produce illusory facilitation. Finally, it remains possible that, even though performance on the first two test blocks did not exhibit enhancement in the selected sequence analysis of Experiment 1, an enhancement effect would have been observed in the sleep group had participants been given additional test blocks (i.e., RTs might eventually have fallen below the extrapolated prediction). This possibility is supported by the decreasing RTs across the two test blocks of Experiment 1 and also by similar effects that have been observed for cognitive skills wherein RTs are initially slower than predicted by extrapolation for the first few test blocks but then systematically fall below the extrapolated prediction on subsequent blocks. Rickard (2007) referred to this effect as *learning potentiation* and suggested that it might reflect a form of consolidation that does not yield immediate performance enhancement. Instead, it may establish a foundation supporting an enhanced rate of learning during the test session.

Experiment 2 was designed to simultaneously address all of the factors outlined above. All participants were tested exactly 24 hrs

[4] Fischer et al. (2005) observed enhancement for the control group. However, they used substantial data averaging similar to that used in the Walker et al. study.

after training, with 1 night of sleep between sessions; thus we controlled for both time-of-day and expected time-since-sleep in the comparison of performance at the end of training to that at the beginning of the test. In prior studies in which researchers used a 24-hr delay (Walker, Brakefield, Hobson, et al., 2003; Walker, Brakefield, Seidman, et al., 2003), an enhancement effect has been observed in the analysis of averaged data and that effect has been interpreted as reflecting sleep consolidation. Note that this experiment does not include a wake versus sleep group manipulation. Instead, our goal here is to set up conditions that eliminate or substantially reduce the effects of all factors that could result in illusory enhancement and to determine whether any enhancement remains.

The experiment involved two groups. One of the groups, which we will refer to as the massed practice group, is a close replication of the prior studies, with twelve 30-s blocks of training interleaved with 30-s rest periods. For this group, we expect to observe the same enhancement effect as in previous studies in the analyses of the averaged data. The treatment for the spaced practice group was identical to that for the massed practice group, with the exception that there were thirty-six 10-s training blocks with 30-s rest between each block. Total time performing the task (6 min) was the same for the two groups. However, total elapsed time differed (11.5 min for the massed practice group vs. 23.5 min for the spaced practice group).

By our use of 10-s instead of 30-s blocks in the spaced practice group, the potential consequences of data-averaging that were documented in Experiment 1 (i.e., averaging over within-block reactive inhibition at the end of training and over speedup during the test session) should be substantially curtailed, and the need to perform selected sequence analyses may be eliminated. Further, because the training session in the spaced practice group is spread out over about twice the time interval as for the massed practice group, the hypothesized buildup of fatigue over the course of the training blocks may also be significantly reduced.

Both groups differ from those of prior studies in that the test session had the same number of practice blocks as did the training session. These additional blocks in the test session allowed us to investigate the possibility of a learning potentiation effect.

Method

Participants. One hundred sixty-four young adults from the San Diego community participated for course credit or for pay. All participants were right-handed.

Design and procedure. The procedure and design were the same as those of Experiment 1 except as noted. Participants were randomly assigned to the massed or spaced practice groups. Each trial produced either a blue or red dot below the visually displayed digit for a correct or incorrect response, respectively, forming a row from left to right over the course of each trial sequence as in Experiment 1. The blue versus red dot feedback was not used in Experiment 1. It was introduced here in an effort to reduce the frequency with which participants' trials became desynchronized with the correct responses and thus to decrease the difference between the raw and modified error measures.

Results and Discussion

Fifty-three participants reported practicing between sessions. Two additional participants reported 3 or fewer hours of sleep between sessions. These participants were eliminated prior to data analysis, leaving 54 clean participants in the massed practice group and 55 clean participants in the spaced practice group. Among these participants, the average number of reported hours slept was 6.78 ($SD = 1.4$) for the massed practice group and 6.85 ($SD = 1.3$) for the spaced practice group.

Mean correct trial RTs are shown in Figure 4A as a function of group, practice block, and session. No sequences were removed (other than error trials). For the massed practice group in that figure, each 30-s block has been divided into three blocks

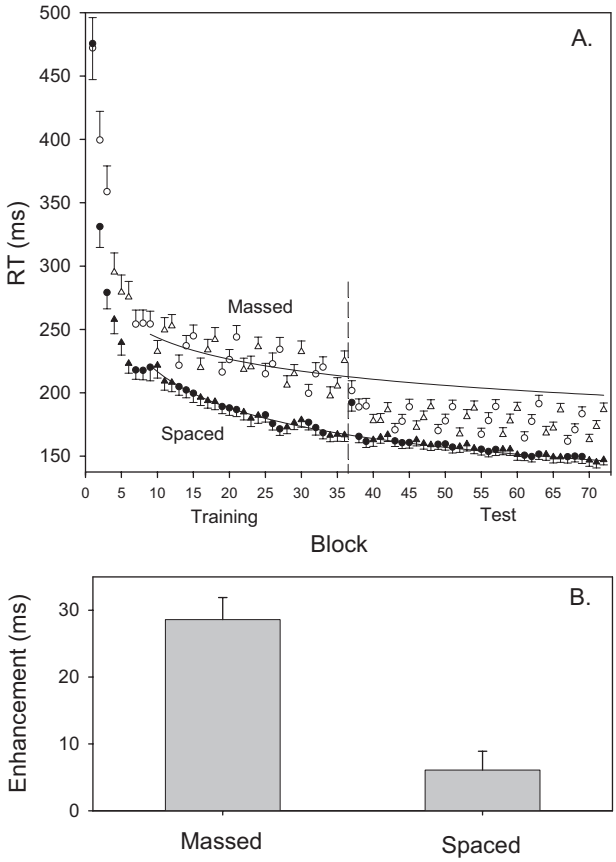

Figure 4. Panel A represents mean response time (RT) in Experiment 2 as a function of group (massed, spaced), session, and block, along with best 3-parameter power-function ($RT = a + b^{*}N^{-c}$) fits to the last 4 min of training and extrapolated to the test session. The vertical dotted line separates training and test sessions. Panel B represents RT enhancement effect (mean RT for the last 90 s of training minus mean RT for the first 90 s of test) for Experiment 2 plotted by group. Empty symbols represent the massed practice group. Filled symbols represent the spaced practice group. Circles represent odd-numbered training and test blocks. Triangles represent even-numbered training and test blocks. Error bars represent the standard error of the mean for each data point.

of approximately 10 s each, yielding 36 blocks per session.[5] Note that the data symbols in the graph alternate between triplets of three circles and triplets of three triangles. This alternation of symbols visually delineates each 30-s practice block of the massed practiced group (for consistency the same pattern of symbol alternation is used for the spaced practice group). Also shown is the best fitting 3-parameter power functions fitted for the last 4.5 min of the training session and extrapolated to the test session.[6] The interaction of Group × Block was significant in a 2 (group; a between-subjects factor) × 36 (block; a within-subjects factor) analysis of variance (ANOVA) for both the training session, $F(35, 3744) = 4.9$, $p < .0001$, and the test session, $F(35, 3744) = 7.5$, $p < .0001$. As hypothesized, spaced practice facilitated the rate of performance improvements during each session.

Consider next the effect of the 24-hr delay between sessions on accuracy and RT. Following Walker and colleagues' (Walker, Brakefield, Hobson, et al., 2003; Walker, Brakefield, Siedman, et al., 2003) 24-hr delay experiments, the last 90 s of performance during training (i.e., the last 3 training blocks for the massed condition and the last 12 training blocks for the spaced condition) was compared to the first 90 s of performance during test. For the massed practice group, mean trial accuracy was .954 (.959 for the modified accuracy measure described earlier) and .977 (.980) for the last three blocks of training and for the first three blocks of test, respectively. For the spaced practice group, these values were .967 (.972) and .962 (.970). Note that the raw and modified accuracy measures are more similar in this experiment than they were in Experiment 1, a result that is likely due to the slight change in error feedback in Experiment 2. A signed rank test on the accuracy difference scores (training minus test) confirmed a significant accuracy improvement on the test for the massed practice group ($p < .001$). There was no significant improvement for the spaced practice group ($p = .78$; t tests yielded the same patterns of results). The results for the massed practice group replicate the accuracy improvement on the test found by Walker and colleagues (Walker, Brakefield, Hobson, et al., 2003; Walker, Brakefield, Siedman, et al., 2003). The spaced practice manipulation eliminated that effect.

The enhancement effect (mean RT for the last 90 s of training minus the mean RT for the first 90 s of the test) is illustrated in Figure 4B. A two-sample t test on the enhancement data was statistically significant, $t(107) = 5.25$, $p < .0001$, indicating a strong group difference. For the massed practice group, there was an enhancement effect of 29 ms (13.3%), $t(53) = 10.00$, $p < .0001$. This effect is of similar magnitude to that reported previously for experiments with a 24-hr delay and in which 30-s practice was interleaved with 30-s rest (Walker, Brakefield, Seidman, et al., 2003). For the spaced practice group, there was an enhancement effect of 6 ms (3.5%), $t(54) = 2.18$, $p = .03$.

There is no indication, however, that the 6-ms improvement for the spaced practice group reflects active sleep learning processes. Instead, improvement falls on the extrapolated learning curve (power function), indicating that improvement reflects only normal trial-to-trial learning that is occurring within the last 90 s of training and the first 90 s of the test. It is crucial that there is no evidence of a downward-going discontinuity in the RT data between the end of the training session and the beginning of the test, as would be expected if there were a sleep-based enhancement mechanism.

Inspection of the successive triplets of data points for the massed practiced condition in Figure 4A reveals a pattern of within-block reactive inhibition similar to that observed in Experiment 1 (although at a coarser grain-size). For the first two triplets in the training session and for the first triplet in the test session there is speedup. However, from the fourth triplet onward in both sessions there is a systematic slowing effect within each triplet followed by a return to a better performance level for the first data point of the next triplet. This reactive inhibition, which is not evident for the spaced practice group in the 10-s block averages (although it is likely still present within each 10-s block to a lesser extent), explains part of the speedup advantage for the spaced practice group (Figure 4A).

Whether the within-block reactive inhibition can explain all of the speedup advantage for the spaced group can be evaluated by limiting analysis to only the first 10-s subblock of each 30-s block, which eliminates most of the reactive inhibition for the massed practice group and equates any remaining effect for the two groups. The results are shown in Figure 5. We used the same ANOVA design that was described above, and the Group × Block interaction was again significant for the training session, $F(11, 1177) = 3.0$, $p < .001$, although it was not significant for the test session. This result indicates that part of the spacing advantage during training reflects some type of fatigue that builds up over successive training blocks in the massed condition.

Figure 5 distinguishes between two candidate mechanisms that could underlie the slower rate of speedup over blocks for the massed practice group. One possibility is that massed practice results in the buildup of performance fatigue. By this account, trial-to-trial learning is the same for the two groups, but the expression of that learning in performance becomes progressively less efficient over the course of each block in the massed practice group. The second possibility, not mutually exclusive with the first, is that massed practice reduces trial-to-trial learning rate but has no effect on the expression of that learning in performance (i.e., massed practice results in learning fatigue). The buildup of reactive inhibition within each block, for example, might result in a progressive decrease in the rate of trial-to-trial learning over the course of each block. This learning fatigue hypothesis is conceptually similar to the mechanism that Pavlik and Anderson (2003) proposed in their Adaptive Control of Thought—Rational model of spacing effects in foreign vocabulary learning.

If massed practice results exclusively in learning fatigue, then the performance difference between the massed and spaced practice groups at the end of training should have the same magnitude at the beginning of the test. Alternatively, if massed practiced results exclusively in performance fatigue and if that fatigue completely dissipates during the 24-hr delay between sessions, then the performance differences at the end of training should vanish on the first block of the test. A comparison of the last data point of

[5] Because participant trials did not line up exactly with the 10-s boundaries in the massed practice condition, actual time of the 10-s subblocks that were identified for each participant varied from about 9,800 ms to 10,200 ms.

[6] Elimination for the first 90 s of training was required in this case to achieve random residuals in the fits.

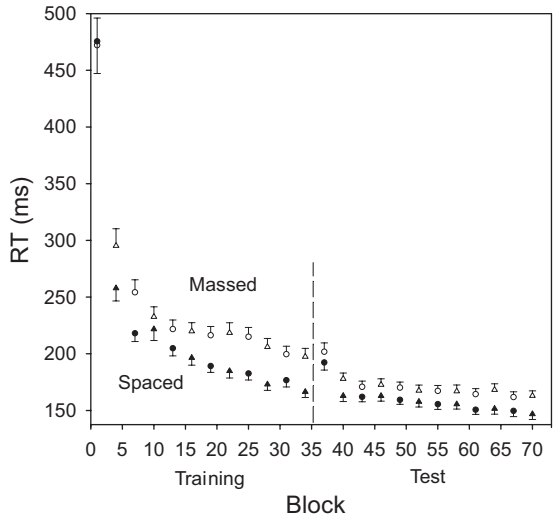

Figure 5. Mean response time (RT) for the first 10 s of each 30-s practice interval in Experiment 2 as a function of group (massed, spaced), session, and block. The vertical dotted line separates training and test sessions. Empty symbols represent the massed practice group. Filled symbols represent the spaced practice group. Circles represent odd-numbered training and test blocks. Triangles represent even-numbered training and test blocks. Error bars represent the standard error of the mean for each data point.

practice to the first data point of the test in Figure 5 supports the performance fatigue account. A 2 (group) $\times$ 2 (session) ANOVA limited to those data points confirms the Group $\times$ Session interaction, $F(1, 107) = 8.3$, $p < .005$. The group difference was significant at the end of training, $t(107) = 3.33$, $p < .002$, but not at the beginning of the test, $t(107) = 0.83$, $p = .40$.

General Discussion

In both experiments, the sleep enhancement effect that has been reported in prior studies was replicated when each performance block lasted for 30 s and when data were averaged in the usual manner. However, we identified four aspects of that approach to design and analysis that can lead to an enhancement effect that is unrelated to sleep consolidation: (a) within-block reactive inhibition, (b) averaging over learning, (c) time-of-day and time-since-sleep confounds, and (d) the progressive buildup of fatigue over the course of training. When these factors were addressed in the data analyses or were controlled for in the design, no sleep enhancement was observed, as measured by either accuracy or RT. We conclude that sleep does not enhance learning for the explicit motor sequence task.

It has recently been proposed that the sleep enhancement that has sometimes been observed in procedural memory tasks is greatest when the training session saturates the potential for within-session learning (Hauptmann & Karni, 2002; Hauptmann, Reinhart, Brandt, & Karni, 2005). According to that hypothesis—which has not yet been tested using the motor sequence task—a training session with a large number of practice blocks is more likely to saturate learning and to produce sleep enhancement than is a training session with fewer practice blocks. Our results provide an alternative account of such effects in terms of performance fatigue. It is exactly in the context of a long training session that substantial fatigue is likely to build up and that an apparent asymptote in learning (saturation) will be observed (i.e., given enough practice within a session, the rate of buildup of fatigue might completely mask any remaining learning effects). Fatigue dissipates between sessions, yielding a strong illusory sleep enhancement effect on the test.

In light of our results, we submit that the design of Experiment 2, or its conceptual equivalent, should be employed in future investigations of the hypothesis that sleep enhances performance in any skill domain. There is already evidence that the issues we raise for the explicit motor sequence tasks are relevant to other skill tasks. The fast rate of performance improvement at the beginning of a test that follows sleep is demonstrated, for example, by Brashers-Krug et al. (1996; Figure 2A), and by Rickard (2007). An important role of time-of-day effects in generating illusory sleep enhancement has recently been demonstrated for implicit motor skill learning (Keisler et al., 2007). Dramatic effects of massed versus spaced practice on both the rate of performance improvement within each session and on the effects of delays between sessions are demonstrated by the Adams (1952) rotary pursuit study. In one condition of that study, participants practiced for 6 (uninterrupted) min in each session, and in another condition they practiced over 36 blocks of 10 s each, with 1-min breaks between blocks. Achieved performance at the end of training was much better in the distributed practice than in the massed practice condition. There was performance enhancement between sessions (each separated by 1 day) for the massed practice group, but performance worsened between sessions for the spaced practice group.

Our results are consistent with the possibility that sleep, although apparently not producing performance enhancement for motor skills, may yield protection from forgetting (i.e., stabilization). Consolidation as protection from forgetting could take either active or passive forms. In the active case, there is a special mechanism unique to one or more sleep stages that complements waking consolidation processes in achieving stabilization. In this case, the sleep consolidation mechanism would presumably have a unique, albeit perhaps subtle, role that is distinct from waking consolidation. Alternatively, the effect of sleep on protection from forgetting may be passive. That is, rather than there being a special mechanism that operates only during sleep, sleep may allow a purely time-based consolidation mechanism to operate more efficiently. During sleep there is no new motor learning to interfere with ongoing consolidation, and this fact alone might lead to relatively better performance on a test that follows a period of sleep. An analogous mechanism for explaining sleep effects for declarative memory tasks has been suggested by Wixted (2004).

Although an effect very similar to the learning potentiation effect reported by Rickard (2007) for cognitive skills was observed for the massed practice group in Experiment 2, that effect was no longer observed in the spaced practice group. That result raises the possibility that spaced practice would also eliminate learning potentiation for cognitive skills. If so, then a general rule for the effects of delays between sessions for a broad class of skills may emerge: With respect to the underlying achieved performance (i.e., controlling for the various factors discussed here that may lead to illusory enhancement), delays between sessions will always produce equivalent or somewhat worse performance than would be expected by extrapolation of performance from the previous ses-

sion. This hypothesis is qualitatively consistent with the model of multisession cognitive skill learning proposed by Anderson, Fincham, and Douglass (1999). The extent of the worsening after delay may in turn be determined by several factors, including the length of the delay and whether participants got a normal night of sleep following the training session.

Finally, note that the effect of spaced versus massed practice observed here is markedly different from that generally observed in the memory literature. In the case of memory recall tasks, for example, spacing of practice results in a slower rate of learning (usually indexed exclusively by accuracy) during training but better performance on a delayed test. In the current experiments, spaced practice resulted in better performance during training but no significant differences on the test. Similar results were obtained in the Adams (1952) rotary pursuit study. It remains an important goal in future work to isolate the principles that determine which type of spacing effects will hold for any given task.

References

Adams, J. A. (1952). Warm-up decrement in performance on the pursuit-rotor. *American Journal of Psychology, 65,* 404–414.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1120–1136.

Brashers-Krug, T., Shadmehr, R., & Bizzi, E. (1996). Consolidation in human motor memory. *Nature, 382,* 252–255.

Caithness, G., Osu, R., Bays, P., Chase, H., Klassen, J., Kawato, M., et al. (2004). Failure to consolidate the consolidation theory of learning for sensorimotor adaptation tasks. *Journal of Neuroscience, 24,* 8662–8671.

Donchin, O., Sawaki, L., Madupu, G., Cohen, L. G., & Shadmehr, R. (2002). Mechanisms influencing acquisition and recall of motor memories. *Journal of Neurophysiology, 88,* 2114–2123.

Fischer, S., Hallschmid, M., Elsner, A. L., & Born, J. (2002). Sleep forms memory for finger skills. *Proceedings of the National Academy of Sciences, USA, 99,* 11987–11991.

Fischer, S., Nitschke, M. F., Melchert, U. H., Erdmann, C., & Born, J. (2005). Motor memory consolidation in sleep shapes more effective neuronal representations. *The Journal of Neurosciences, 25,* 11248–11255.

Frank, M. G., & Benington, J. H. (2006). The role of sleep in memory consolidation and brain plasticity: Dream or reality? *The Neuroscientist, 12,* 477–488.

Goedert, K., & Willingham, D. (2002). Patterns of interference in sequence learning and prism adaptation inconsistent with the consolidation hypothesis. *Learning and Memory, 9,* 279–292.

Hauptmann, B., & Karni, A. (2002). From primed to learn: The saturation of repetition priming and the induction of long-term memory. *Cognitive Brain Research, 13,* 313–322.

Hauptmann, B., Reinhart, E., Brandt, A., & Karni, A. (2005). The predictive value of the leveling off of within-session performance for procedural memory consolidation. *Cognitive Brain Research, 24,* 181–189.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review, 7,* 185–207.

Hoddes, E., Zarcone, V., Smythe, H., Philips, R., & Dement, W. C. (1973). Quantification of sleepiness; a new approach. *Psychophysiology, 10,* 431–436.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Keisler, A., Ashe, J., & Willingham, D. T. (2007). Time of day accounts for overnight improvement in sequence learning. *Learning and Memory, 14,* 669–672.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences, 11,* 442–450.

McGaugh, J. L. (2000). Memory—A century of consolidation. *Science, 287,* 248–251.

Muellbacher, W., Ziemann, U., Wissel, J., Dang, N., Kofler, M., Facchini, S., Boroojerdi, B., et al. (2002). Early consolidation in human primary motor cortex. *Nature, 415,* 640–644.

Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PLoS ONE, 4,* Article e341. doi:10.1371/journal.pone.0000341

Pavlik, P. I., Jr., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Doerner, & H. Schaub (Eds.), In *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 177–182). Bamberg, Germany: Universitats-Verlag Bamberg.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311.

Rickard, T. C. (2007). Forgetting and learning potentiation: Dual consequences of between-session delays in cognitive skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 297–304.

Robertson, E. M., & Cohen, D. A. (2006). Understanding consolidation through the architecture of memories. *The Neuroscientist, 12,* 261–271.

Robertson, E. M., Pascual-Leone, A., & Miall, R. C. (2004). Current concepts in procedural consolidation. *Nature Reviews Neuroscience, 5,* 576–582.

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. *Current Biology, 14,* 208–212.

Robertson, E. M., Press, D. Z., & Pascual-Leone, A. (2005). Off-line learning and the primary motor cortex. *Journal of Neuroscience, 25,* 6372–6378.

Spencer, R. M. C., Sunm, M., & Ivry, R. B. (2006). Sleep-dependent consolidation of contextual learning. *Current Biology, 16,* 1001–1005.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature, 437,* 1272–1278.

Vertes, R. P. (2004). Memory consolidation in sleep: Dream or reality. *Neuron, 44,* 135–148.

Vertes, R. P. (2005). Sleep is for rest, waking consciousness is for learning and memory—of any kind. *Behavioral and Brain Sciences, 28,* 86–87.

Vertes, R. P., & Siegel, J. M. (2005). Time for the sleep community to take a critical look at the purported role of sleep in memory processing. *Sleep, 28,* 1228–1229.

Walker, M. P. (2005). A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences, 28,* 51–104.

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature, 425,* 616–620.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron, 35,* 205–211.

Walker, M. P., Brakefield, T., Seidman, J., Morgon, A., Hobson, J. A., & Stickgold, R. (2003). Sleep and the time course of motor skill learning. *Learning and Memory, 10,* 275–284.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron, 44,* 121–133.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psychology, 57,* 139–166.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55,* 235–269.

Received October 15, 2007
Revision received February 4, 2008
Accepted March 13, 2008 ■