Sleep and Motor Learning:

Is there Room for Consolidation?

Steven C. Pan and Timothy C. Rickard

University of California, San Diego

This manuscript was accepted for publication in *Psychological Bulletin* on December 30, 2014. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version is available at: http://dx.doi.org/10.1037/bul0000009

Steven C. Pan and Timothy C. Rickard are affiliated with the Department of Psychology, University of California, San Diego. S. Pan is supported by an NSF Graduate Research Fellowship. The authors thank Mark Appelbaum for helpful comments and consultation on meta-analytic procedures. Please address correspondence to: Timothy C. Rickard, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. Email: trickard@ucsd.edu

Abstract

It is widely believed that sleep is critical to the consolidation of learning and memory. In some skill domains, performance has been shown to improve by 20% or more following sleep, suggesting that sleep enhances learning. However, recent work suggests that those performance gains may be driven by several factors that are unrelated to sleep consolidation, inviting a reconsideration of sleep's theoretical role in the consolidation of procedural memories. Here we report the first comprehensive investigation of that possibility for the case of motor sequence learning. Quantitative meta-analyses involving 34 papers, 88 experimental groups and 1,296 subjects confirmed the empirical pattern of a large performance gain following sleep and a significantly smaller gain following wakefulness. However, the results also confirm strong moderating effects of four previously hypothesized variables: averaging in the calculation of pre-post gain scores, build-up of reactive inhibition over training, time of testing, and training duration, along with one supplemental variable, elderly status. With those variables accounted for, there was no evidence that sleep enhances learning. Thus, the literature speaks against, rather than for, the enhancement hypothesis. Overall there was relatively better performance after sleep than after wakefulness, suggesting that sleep may stabilize memory. That effect, however, was not consistent across different experimental designs. We conclude that sleep does not enhance motor learning and that the role of sleep in the stabilization of memory cannot be conclusively determined based on the literature to date. We discuss challenges and opportunities for the field, make recommendations for improved experimental design, and suggest approaches to data analysis that eliminate confounds due to averaging over online learning.

Keywords: sleep consolidation, learning, motor skills, motor sequence learning, sleep enhancement.

Sleep and Motor Learning:

Is there Room for Consolidation?

The hypothesis that sleep plays a major role in the consolidation and enhancement of learning has attracted substantial attention over the last two decades, among both cognitive scientists and the popular media. A December, 2013 literature search with the keywords *sleep* and *consolidation* identified over 700 studies, with an annual publication rate exceeding 100 and, thus far, increasing exponentially. The vast majority of those studies conclude in favor of some type of sleep consolidation, inviting wholesale reconsideration of traditional cognitive theories in which sleep is implicitly assumed to be unrelated to learning and memory. The findings also open the door to potentially important lines of translational work seeking to optimize the impact of sleep on learning and memory through sleep scheduling, pharmacological, or other manipulations.

The sleep consolidation literature is best understood through separate consideration of the declarative and procedural (primarily perceptual and motor skill) domains (Plihal & Born, 1997; Smith, 2001; Stickgold, 2005; Walker & Stickgold, 2004). In the declarative domain, the primary claim is that sleep *stabilizes new learning* (i.e., protects new learning from interference and forgetting; for recent discussion see Mednick, Cai, Anagnostaras, Shuman, & Wixted, 2011). In the procedural domain – the focus of the current work – the usual finding is that performance following sleep is not only better than performance after a matched waking period, but also better than performance at the end of the previous day's training session. That finding constitutes the basis of the *sleep-based enhancement* theory (e.g., Diekelmann & Born, 2007; Robertson, Pascual-Leone, & Miall, 2004; Walker, 2005), according to which sleep consolidation in the procedural domain enhances learning rather than merely stabilizing it.

In several recent studies, however, the sleep-based enhancement hypothesis has been called into question (e.g., Brawn, Fenn, Nusbaum, & Margoliash, 2010; Keisler, Ashe, & Willingham, 2007; Nemeth et al., 2009; Rickard, Cai, Rieth, Jones, & Ard, 2008; Sheth, Janvelyan, & Khan, 2008). Those researchers have identified a number of moderating variables that may account for at least a portion of the performance improvement following sleep. There are also recent data suggesting that, under some circumstances at least, skill performance after sleep may be no better than after a period of wakefulness (e.g., Cai & Rickard, 2009). Nevertheless, the general presumption that sleep enhances learning remains common in the most recent studies (e.g., Fogel et al., 2014; Tucker, McKinley, & Stickgold, 2011).

To further explore the extent to which sleep consolidation vs. other factors can account for the observed sleep effects in the motor domain, we conducted a quantitative meta-analytic review of the empirical work on sleep and explicit motor sequence learning, which constitutes the largest and most influential sub-literature in the procedural domain. That task typically involves training subjects on an explicitly defined, five (or more) element, deterministic finger tapping sequence that consists of either tapping each of the four fingers against its own keyboard key or the four fingers against the thumb. Labeling the index as finger one, the most commonly used sequence involves five taps: 4, 1, 3, 2, 4. The training session involves a series of performance-break cycles (e.g., 30 second performance blocks interleaved with 30 s breaks). The test session after a sleep delay involves the same task and pattern of performance-break cycles. In some studies, a matched wake delay group is also included as a control. The primary dependent measure is the rate of correct sequence completion, which is most often calculated as either the number of sequences correctly completed per 30 s training block or the mean latency over a fixed number of key presses or key press sequences within a block.

The core behavioral findings in that literature are that: (1) performance after a delay involving sleep exhibits a robust performance gain – often of 20% or more – compared to performance at the end of training, and (2) when a matched wake control group is included, the performance gain for that group is smaller than for the sleep group and is often not significantly different from zero. Henceforth, the term *post-delay gain* will refer to the grand average gain observed for the combined sleep and wake groups in our sample, *post-sleep gain* will refer to the gain observed for sleep groups, and *relative sleep gain* will refer to the difference between the gain following sleep and the gain following wakefulness (i.e., post-sleep minus post-wake). Although the post-sleep and relative gains are the most theoretically pertinent empirical phenomena, and the ones on which we will ultimately focus, our primary analytical approach maximizes statistical power by fitting multiple candidate predictor variables to the joint set of sleep and wake groups; the overall observed gains in those analyses will be referred to as post-delay gains. Importantly, the terms post-delay gain, post-sleep gain, and relative sleep gain are intended to refer only to the empirical phenomena, not their theoretical interpretation.

The Current Review

Our primary goals in this review are to critically evaluate both of the major theoretical claims about consolidation in the motor learning and sleep literature. The first claim is that the empirical post-sleep gain reflects sleep-based enhancement of learning. If that claim is correct, then the second claim, namely that consolidation operates more effectively during sleep than during wakefulness, follows naturally. If, however, the first claim is incorrect then the second claim may still stand; namely, sleep-specific consolidation may stabilize rather than enhance procedural learning.

Our approach to theoretical inference is two-fold. First, we consider the existing evidence that factors other than sleep consolidation may explain at least a portion of the observed gain effects. Second, we conduct quantitative meta-analyses and meta-regressions to determine the predictive power of those factors for experiments run to date. Those analyses are presented in the Results section in the following order:

1. The *primary analyses* are conducted on the full set of 88 sleep and wake groups in our sample, where a group refers to a single sample of subjects that is trained on the motor sequence task and is then tested after a delay involving either sleep or wakefulness. There are 65 sleep groups and 23 wake groups in our sample. The goal of the primary analyses is to explore the role of several previously hypothesized variables in explaining the post-delay gain. Those variables include sleep status (wake vs. sleep groups), which indexes the magnitude of the relative gain, along with a number of other variables described below that may influence the post-delay gain and hence the post-sleep gain. Assessment of the predictive power of those variables is optimized by jointly fitting both sleep and wake groups.
2. Based on the primary analysis, a *working model* of the important variables for explaining post-delay gain effects is advanced. That model allows us to evaluate the magnitude of the relative gain effect in the literature (i.e., the difference in gain for sleep and wake groups) and to determine whether the post-sleep gain effect survives after adjusting for the influence of non-consolidation related variables.
3. Following the primary analysis, a *secondary analysis* limited to the set of 23 matched sleep-wake groups in the sample is conducted, allowing for a more refined investigation of the relative gain effect and its causal basis.

Primary Factors Hypothesized to Moderate Gain Effects

The meta-analyses focus on five primary factors that have previously been hypothesized to influence the post-delay gain and (or) the relative gain: (1) *sleep status* (wake only vs. sleep groups), which directly indexes the magnitude of the relative gain, (2) the amount of *data averaging* in calculation of the pre-post gain scores, (3) *training duration*, (4) the build-up of *reactive inhibition* during the course of training, and (5) the effects of *time of day* on performance during the training and (or) test sessions. Based on properties to be discussed below, we hypothesize that factors 2 through 4 will primarily affect the post-delay gain but not the relative gain (i.e., their effects will be equivalent for wake and sleep groups) whereas time of training and testing (factor 5) have the potential to influence both the post delay and relative gains.

Unlike sleep status, factors 2 through 5 have not been extensively discussed in the literature. Below we elaborate on each of them, summarize prior evidence for their effects on gain scores, and specify their operationalization as predictors in the meta-analyses.

Data Averaging

A nearly ubiquitous strategy for measuring post-delay gains in this literature has been to calculate the difference between average performance over some duration or number of trials at the end of training (the *pre-test*) and average performance over a roughly equivalent duration or number of trials at the beginning of the test session (the *post-test*). The range of data averaging across studies is large, spanning from about 25 s of performance for both the pre- and post-tests to as much as 900 s per test. A large amount of averaging has the advantage of yielding more precise estimates of each subject's pre-test and post-test scores and hence more statistical power to detect a performance gain. However, calculation of gain scores using that strategy runs the

risk that learning that occurs during the pre-test and (or) post-test periods (i.e., *online learning*) is incorporated into the gain score (Rickard et al., 2008; Robertson et al., 2004).

The problem is illustrated in Figure 1. The dependent variable in this example is the mean time per finger press across a block of trials (measured in ms per key press), and hence learning yields smaller values. The delay between training and test sessions is assumed in Panel a to have no effect (including no sleep consolidation effect) on either underlying skill or observed performance. In that case, performance improvement both within and between sessions is expected to follow a smooth, monotonically decreasing curve. For data averaged over subjects, there is a preponderance of support for that expectation across a wide variety of task domains (for performance curve reviews see Heathcote, Brown, & Mewhort, 2000; Newell & Rosenbloom, 1981), with exceptions primarily in the special cases of discrete strategy shifts that can occur in some task domains in the early phase of training (e.g., Rickard, 2004) and extreme fatigue than can give rise to worsening of performance toward the end of a long training session (e.g., Adams, 1952).

Despite the fact that the between-session delay is assumed to have no effect in Figure 1a, averaged pre-post gain scores are guaranteed, at the population level, to yield a post-delay performance gain, given only the expected monotonic performance improvement. It should also be apparent that extending the range of averaging further backward through training and further forward through testing would exacerbate the effect.

Thus, the open question in this literature is not whether post-delay gains as computed using pre-post difference scores are confounded by online learning (to some extent they almost certainly are), but whether that confounding factor accounts for a theoretically meaningful portion of the performance gain. At one extreme, averaging over online learning could account

for all of the post-delay gain (Figure 1a). The alternative case, in which averaging is clearly not sufficient to explain the post-delay gain, is illustrated in Figure 1b. Remarkably, there have been no prior tests of those possibilities in the literature. In the meta-analyses below, averaging is quantified as duration in s of the pre-test (duration of the post-test among included studies was always identical to or closely approximated that of the pre-test). For studies in which each performance block was a fixed number of trials, the duration of averaging was estimated from trial latency data that was provided graphically.

Training Duration

Given that the rate of performance improvement decreases as a function of practice, post-delay gains as estimated by averaged pre-post difference scores should, in the population, be greater in short than in long duration training designs, holding the amount of averaging constant. This effect is illustrated by comparison of Panels a and c of Figure 1. In Panel a, a relatively large amount of training is presumed, such that the rate of block-to-block improvement towards the end of training is low. In that case, averaging over online learning is expected to have a relatively small effect on the pre-post gain score. In Panel c, there is less training, the block-to-block improvement rate is high at the end of training, and averaging over online learning is expected to have a larger effect on the gain score. Thus, the magnitude of the post-delay gain as measured by pre-post difference scores is expected to be negatively correlated with the duration of training. In the meta-regressions, training duration is operationalized in s and corresponds to total time on task (excluding breaks).

Reactive Inhibition

Empirically, reactive inhibition refers to performance worsening that can accumulate during a period of continuous training (Hull, 1943). It tends to dissipate, at least in part, when

brief breaks are inserted between blocks of training. If there are multiple performance-break cycles over a training session, as in the motor sequence literature, performance can exhibit a scalloped effect, worsening during each uninterrupted performance block but improving across blocks. Rickard et al. (2008) and Brawn et al. (2010) demonstrated highly robust scalloped reactive inhibition effects using the commonly employed 30 s-30 s performance-break cycle, as shown for Rickard et al.'s massed practice sleep group in Figure 2. The scalloped effect is evident for that group after the first few 30 s blocks of each session. The absence of the scalloped effect during the first few blocks of training in the massed group suggests that rapid learning during that period masks any reactive inhibition effect. A briefer effect of the same type at the beginning of the test session suggests that the magnitude of reactive inhibition increases across the first few performance-break cycles within a session.

In agreement with the bulk of the literature, in which 30 s-30 s cycles have been used, the massed practiced group in Figure 2 exhibited a highly significant post-sleep gain after the 24-hour delay between sessions, as indicated by both analysis of averaged pre- and post-test data and by application of a novel (in this literature) continuity test to be described later. However, that gain may result from the differences in magnitude of reactive inhibition at the end of training vs. the beginning of the test, and may not require a sleep consolidation interpretation. As a test of that possibility, Rickard et al. reduced the reactive inhibition effect in their spaced practice group by using a 10 s-30 s performance-break design. As shown in Figure 2, there was no evidence for a post-sleep gain for that group. For additional evidence that is consistent with the hypothesis that reactive inhibition resolves after a several minute delay see Brawn et al. (2010) and Hotermans, Peigneux, de Noordhout, Moonen, and Maquet (2006).

The two design factors that could influence the magnitude of reactive inhibition are duration of performance and duration of break within each performance-break cycle. There is no empirical evidence to date as to which is more important across the range of values that have been used in the literature (only performance duration was manipulated by Rickard et al. and Brawn et al.). To explore that question, we treat *performance duration* and *break duration* per cycle as two separate predictors in the meta-analyses, each measured in s.

Time of Training and Testing

Two physiological variables that determine sleep propensity are known to influence performance across a variety of tasks: circadian rhythms and homeostatic sleep drive.

Circadian rhythms. Among the non-consolidation factors that may affect the observed post-delay and relative gains, circadian rhythms, which vary on a 24-hr cycle under naturalistic conditions, have received the most attention. In several studies, training performance has been compared for matched morning and evening groups, with no significant differences observed (Albouy et al., 2013; Brawn et al., 2010; Doyon et al., 2009; Korman, Raz, Flash, & Karni, 2003), suggesting that circadian influences on the pre-post gain scores may be negligible. Tempering that inference, however, is the conclusion of Keisler et al. (2007) that circadian rhythms may account for sleep gain effects for implicit motor learning tasks (e.g., the implicit serial reaction time task). Further, the possibility of a selective circadian effect at time of testing has not previously been explored.

The probable form of any circadian influence can be inferred based on the results of desynchronization experiments conducted in a variety of task domains (for discussion see Blatter & Cajochen, 2007). Desynchronization experiments allow circadian influences to be assessed while holding time since sleep (i.e., homeostatic effects) constant. Although the pattern can be

task dependent, for simple skills and automatized memory retrieval among healthy subjects, circadian factors tend to yield relatively poor performance in the early morning, improvement through the early afternoon, and worsening into the late evening. Among some individuals there is a second order performance dip between approximately 2 and 4 pm, although that effect is not aways observed in group-level data (Monk, 2005).

Homeostatic sleep drive (time since sleep). Physiological sleep drive is determined jointly by circadian rhythms and time since the last sleep period. As time since sleep increases, the *homeostatic* component of sleep drive is said to increase. Although less frequently acknowledged in the literature, some experimental designs may have homeostatic confounds that stand independently of any circadian effects. If, for example, subjects are trained in the evening (in the context of relatively high homeostatic sleep drive) and tested in the morning after sleep (lower sleep drive), and if the level of homeostatic drive is negatively correlated with task performance, then a post-sleep gain could be observed based on homeostatic influences alone.

Joint circadian and homeostatic effects. In the current meta-analyses it is not possible to separate the effects of circadian and homeostatic factors. Drawing on the literature outlined above, however, joint circadian and homeostatic factors are expected to exert a concave downward effect on performance across either time of training, time of testing, or both. Specifically, the observed post-delay gain should be relatively small in the morning, reach a peak around mid-day, and become smaller again in the evening. Examples of tasks exhibiting that effect include simple addition (e.g., Hull, Wright, & Czeisler, 2003), mirror drawing, multiplication, and code transcription speed (e.g., Kleitman, 1933), digit symbol processing and verbal fluency (e.g., Allen, Grabble, McCarthy, Bush, & Wallace, 2008), and psychomotor vigilance (e.g., Jewett et al., 1999). To test for a concave downward effect of time of day – or

indeed any linear or quadratic effect – we fitted both linear and quadratic variables for both time of training and time of testing, measured as number of hrs past midnight.

Secondary Candidate Predictors

Analyses of a set of secondary predictors of potential interest were also conducted. Those predictors include the *delay* between training and testing in hrs, whether the task involved keyboard tapping or finger-to-thumb tapping (*task type*), whether subjects were children (<18 years; *child status*), and whether subjects were elderly (>59 years; *elderly status*). Among the sleep groups, the effect of a nap vs. a full night of sleep (*nap status*) was tested. Among the full night sleep groups for which sleep time was reported, the effect of number of *hours slept* was tested.

Methods

Literature Search

An extensive literature search was conducted to obtain a comprehensive set of empirical research studies on sleep and explicit motor sequence learning. Included were online searches of four databases to obtain peer-reviewed research articles, correspondence with authors to obtain additional data and unpublished manuscripts, ancestral searches of article reference lists, and further searches of dissertation and other databases. In all searches, we applied a 34-year date range from January 1, 1980 to June 17, 2014, the end date being the day on which the online search was completed. This range well exceeded the entire span of published papers in the sleep and motor sequence learning literature (which at the time of this review has primarily occurred in the past decade and a half).

Database searches. Four online databases for peer-reviewed research articles were searched: EBSCOhost Academic Search Premier, MEDLINE, PsychINFO, and Thomson

Reuters Web of Science. All searches involved the keyword *sleep* in combination with each of these individual or binary terms: *finger, finger-sequence, finger-tapping, finger-thumb, finger-thumb opposition, motor learning, motor sequence, sequence,* and *tapping*. The four database searches yielded 2,299 hits; 1,367 of them were duplicated across databases, leaving 932 references for further review. Those references were then entered into a three-stage review process (see Figure 3) to determine suitability for inclusion in the meta-analyses.

The first stage, title-level review, determined whether articles had any possible relevance to sleep or memory consolidation. This stage involved two raters (the authors of this review) independently reading only the titles of each article. If the title referred to (1) sleep in learning and memory, (2) time-based consolidation, or (3) sleep-based consolidation, it was flagged for inclusion. If the title unambiguously focused on other topics, it was not. Titles for which no clear determination of the article content could be made based on those criteria were also flagged for inclusion. If at least one rater flagged an article, that article remained in consideration for the next phase. Overall rater agreement was high (Cohen's kappa = 0.86). Of the 932 articles entered into this stage of review, 603 were excluded and 329 survived.

The second stage, abstract-level review, identified empirical research articles pertaining to sleep and motor memory consolidation. This stage involved the same two raters independently reading the abstracts of all articles that had survived the first stage of review. If the abstract addressed sleep and motor learning, the article was flagged for inclusion. If the abstract indicated that the article was not an empirical research study (e.g., a review paper or commentary), it was excluded; if animal populations were used, or if clinical populations were used (e.g., if participants in the study had been diagnosed with developmental, neurological, physical, psychiatric, or sleep disorders), it was also excluded. As in the first stage, articles were always

included if at least one rater indicated that it should remain under consideration. Overall rater agreement was high (Cohen's kappa = 0.89). Of the 329 articles entered into this stage of review, 222 were excluded and 107 survived.

The third stage, article-level review, served as the final assessment for inclusion. This stage involved the same two raters reading the full text of articles that survived the second stage of review. There were six instances of disagreement between raters; those discrepancies were resolved by subsequent discussion and mutual agreement between raters. Of the 107 articles entered into this stage of review, 78 were excluded and 29 were selected for inclusion.

Inclusion criteria for third stage. In addition to eliminating groups for which the necessary statistics were not reported, the inclusion criteria below served to minimize study heterogeneity beyond the predictor variables to be tested, as is generally recommended in quantitative meta-analyses. Exclusion of papers was based solely on the criteria summarized below and not on an independent assessment of the research quality or on the appropriateness of the experimental design for the intended purpose.

1. The paper must have involved an explicit motor sequence learning task in which participants articulate the fingers on one hand in accordance with a repeating pattern. Two closely related tasks qualified: finger-keyboard (or button-box) tapping and finger-thumb tapping.
2. The paper must have included at least one group in which either a full night of sleep or a daytime nap intervened between training and test sessions. For each paper with at least one qualifying sleep or nap group, any matched waking control groups were included. A small number of partial night sleep studies, in which subjects were awakened mid-sleep for training or testing, were excluded.

3. Identical motor sequence tasks must have been used during both training and testing. Groups for which another motor sequence task intervened between the training and test sessions, or was interleaved during testing, were excluded. Groups with longer breaks inserted between the end of the main training session and the pre-test were excluded if the break resulted in significant performance improvement on the pre-test. Inclusion of those groups could have contaminated the assessment of reactive inhibition effects (which may be sensitive to break duration as discussed earlier) on gain scores. By that criterion, only the two massed practice groups of Brawn et al. (2010) were excluded.
4. Both the sample size and the post-delay effect size measuring performance rate must have been reported or derivable. Design descriptions must have allowed for determination of the values of all primary predictor variables (sleep status, averaging, training duration, performance and break duration during each performance-break cycle, time of training, and time of testing). Where only ranges were reported (e.g., for time of training or testing), the midpoint of the range was used to estimate the average value for the group. Results for accuracy in this literature generally converge with those for correct performance rate. Because accuracy has not always been reported and because statistical tests on accuracy are likely to have lower power, analyses were performed only on correct performance rate.
5. Experiments involving pharmacological manipulations were excluded, unless there was a healthy control group or groups. In such instances, data from that group or groups were included.
6. Data averaging in calculation of the pre-post gain scores must have occurred over

approximately the same duration or number of blocks for the pre-test and the post-test, and those blocks must have been contiguous. Studies in which pre- and post-test averaging encompassed the entire training and test session were excluded. Averaging over entire sessions is guaranteed to yield a large post-delay gain given the large expected performance improvement, particularly during approximately the first half of the training session. Among the remaining groups, the longest duration of averaging was 120 s and not more than one-third of the training session. Among those groups, which constitute the great majority of the literature, the extent of influence of pre-post averaging on the observed post-delay gain is unknown. They could in the current analyses prove to be either negligible or substantial.

7. If a study had multiple training and test sessions for a given group of subjects, then the following rules applied. For sleep groups, data from the training and test sessions immediately adjacent to the first chronological sleep delay interval were extracted for meta-analyses. In most cases, this meant selecting the first chronological night sleep delay; in two cases, this meant selecting a nap delay over a subsequent night sleep delay (Korman et al., 2007; Korman, Dagan, & Karni, in preparation). For the wake groups, data from the training and test sessions immediately adjacent to the first waking delay interval were extracted.

Ancestral searches and unpublished data. We conducted ancestral searches on the reference lists of articles that survived the three-stage review process, seeking to identify any additional peer-reviewed research studies. That search resulted in the addition of four articles to the meta-analyses, yielding a total of 33.

To combat publication bias and the “file drawer” issue (Strube & Hartmann, 1983), we

also contacted eleven sleep consolidation researchers who have published recently on this topic to request unpublished data, receiving nine responses. One unpublished data set was obtained (Korman et al., in preparation). The other eight researchers informed us that they had no unpublished data. In addition, we performed online searches of ProQuest Dissertations and Theses and Google Scholar, using the exact set of keywords as used in the preceding online database searches (and specifying master's and doctoral dissertations in the former and keyword hits in abstracts in the latter). We determined that all of the dissertations with relevance to this review had been subsequently published in peer-reviewed journals (and were already flagged for inclusion in prior searches). Similarly, relevant hits on Google Scholar were also duplicated in prior online or ancestral searches.

Missing or incomplete information. We contacted seven authors to request clarifications and additional data on papers that were included in the meta-analyses; all but one responded. In all of these instances, author contact was necessary either to (a) obtain necessary information to calculate effect sizes for specific groups, or (b) quantify primary predictor variables (e.g., time of day for training and testing).

Summary of literature search results. Overall, 34 papers met the criteria for inclusion. Of these, print publication dates ranged from July 2002 to August 2014. In total, 88 groups (65 sleep groups and 23 wake groups) were extracted from those studies, encompassing 1,296 unique subjects. Study, group, the values of the primary set of predictor variables, and the statistical results are shown for all 88 groups in Table 1. Appendix A lists the values of the secondary candidate predictors for each group, when reported. As indicated in Table 1, 76 of the groups involved independent sets of subjects and twelve groups involved the same subjects that were used in twelve of the other 76 groups (as indicated in Table 1).

Random Effects Meta-Analyses with Robust Variance Estimation

Random effects meta-analyses (Borenstein, Hedges, Higgins, & Rothstein, 2010; Raudenbush, 2009) were performed on the gain score effect sizes,

$$d = (gain\ score)/s, \quad (1)$$

where *gain score* is the pre-test mean minus the post-test mean for each group (or vice versa depending on whether the variable was time to complete a fixed number of trials or the number of sequences completed in a fixed amount of time) and *s* is the standard deviation of the subject-level pre-post difference scores for each group. Where the *gain score* was not reported but paired *t* tests or *F* tests for the gain score were, we derived the effect size as follows:

$$d = t/n^{.5} \quad \text{or} \quad d = F^{.5}/n^{.5},$$

where *n* is the sample size.

When neither *t* nor *F* tests were reported but bar graphs of gain scores with standard error bars were reported, *d* was estimated by the following method. First, the height of the bar on the y-axis scale (corresponding to the mean gain score) was estimated to the precision of one row of computer screen pixels. An analogous pixel analysis was then used to estimate the standard error. The effect size for each group was computed as:

$$d = (gain\ score)(n^{.5})/standard\ error.$$

The sampling variability (s*v*) for each effect size was estimated following Morris and DeShon (2002) for the case of repeated measures gain scores:

$$sv = (1/n)[(n\text{-}1)/(n\text{-}3)](1 + nd^2) - d^2/c^2,$$

where *c* is computed using the bias function (Hedges, 1982).

Two random effects, study and group (within study), were estimated hierarchically, using the model:

$$T_{ij} = \mathbf{X_{ij}\beta} + \theta_i + \eta_{ij} + \varepsilon_{ij},$$

where T_{ij} is the estimated effect size for group i in study j, $\mathbf{X_{ij}}$ is the design matrix in study j, $\boldsymbol{\beta}$ is the vector of regression coefficients, θ_i is the study-level random effect, η_{ij} is the group-level random effect, and ε_{ij} is the sampling error.

Random effects meta-analysis assumes that the observed effect size at each level of the hierarchy (i.e. for each study and each group) is a random deviate from its own population effect size distribution. The random effects approach thus accommodates (and provides a quantitative estimate of) possible heterogeneity of population effect sizes over papers and groups due to differences in experimental design, sampled population, or other factors. In the current model the residual variation of the effect size estimate $\boldsymbol{T}_{ij}$ can be decomposed as:

$$V(T_{ij}) = \tau^2 + \omega^2 + \upsilon_{ij},$$

where τ^2 is the variance of the between-study residuals, θ_i, and ω^2 is the variance of the within-study residuals, η_{ij}, and υ_{ij} is the known sampling variability of each group. Estimates of τ^2 and (or) ω^2 that are greater than zero raise the possibility that heterogeneity is present and that fixed effects predictor variables may be needed to fully explain differences in effect sizes over papers and (or) groups within papers.

Given that the covariance structure of the effect size estimates is unknown in this literature, we employed robust variance estimation (Hedges, Tipton, & Johnson, 2010; Tanner-Smith & Tipton, 2012) in the model fitting. All meta-analyses were performed using Stata (StataCorp LP, College Station, TX, USA) and the macro robumeta.ado, which can be downloaded from the Stata Statistical Software Components archive (SSC).

The majority of effect size dependencies in our sample are in the form of multiple independent subject groups within study (paper). There were only 11 pairs of correlated groups in which the same subjects served in each group (each of those pairs is identified in table 1). We thus used the macro's hierarchical weight type option, which accommodates dependencies due to independent groups within study, as the best estimate of the effect size weights (Tanner-Smith & Tipton, 2012). An advantage of the robust variance estimation approach in this case is that it is resistant to potential biasing effect for the correlated group pairs. For reference, the Stata code for the final working model that will be described later is provided in Appendix B.

Results

The primary meta-analyses were performed on the full set of sleep and wake groups. After estimating the aggregate post-delay effect, fits of individual predictors are described. Following that, simultaneous fits of previously hypothesized or individually significant predictors are discussed, culminating in a final working model of important predictors of the post-delay gain. That model is then used to estimate the magnitude of the post-sleep gain for representative and theoretically important values of the predictor variables. Finally, to achieve further insight into relative gain, a secondary meta-analysis limited to the 23 matched pairs of sleep-wake groups is reported, and the effect of experimental design on the observed relative gain is summarized.

Primary Meta-Analyses

In this analysis, a primary goal was to maximize statistical power to detect effects of candidate predictor variables on the post-delay gain. Toward that end, all 88 groups, including 23 pairs of wake-sleep groups (46 total groups) and 42 additional sleep groups, were included. Secondary analyses limited to the 23 pairs or wake-sleep groups are reported later.

The weighted mean effect size for the post-delay gain was highly significant, $d = 0.83$, $p < .0001$, 95% confidence interval (CI): 0.61, 1.05, confirming a large post-delay gain aggregated across wake and sleep groups. There was also a large between-study residual variance component, $\tau^2 = 0.27$, the magnitude of which can be appreciated by comparison to the median within-group sampling variability, *sv*, of 0.14. The within-study residual variance, ω^2, was 0.13. If both of those values were near zero, then there would be no evidence of heterogeneity in the sample and no need to conduct meta-regression analyses. As it stands, the aggregate effect size appears to be strongly moderated by one or more design, analysis, subject population, or other factors that vary at both the paper and the groups-within-paper level, motivating the following meta-regression analyses.

Single predictor fits. Table 2 lists the results for the primary and secondary predictor variables when each was introduced into the random effects model in isolation, or, in the case of time of training and time of testing, when the linear and quadratic components were jointly fitted in isolation.

As predicted by sleep consolidation theory, the estimated sleep status (i.e., relative gain) effect size was large and highly significant: $d = -0.64$ (i.e., post-delay gain effect size was 0.64 smaller for wake than for sleep groups). For sleep groups (combined full night and nap) there was a large and highly significant post-sleep gain ($d = 1.0$, $df = 19.0$, $p < 0.0001$, CI: 0.75, 1.25), whereas for waking groups there was a much smaller but still significant gain ($d = 0.36$, $df = 12.9$, $p = 0.03$, CI: 0.035, 0.68). The latter result is not predicted by the sleep-specific enhancement account and suggests that at least some factors underlying the post-delay gain are common to both wake and sleep groups.

Data averaging also significantly predicted effect size (at $\alpha = 0.05$). The predicted range from the smallest (25 s) to the largest (120 s) amount of averaging among the groups was large, $d = 0.74$, suggesting that the majority of the post-delay gain in this literature may reflect a data averaging artifact. There were also significant effects of time of testing, time of testing squared, and elderly status. In the latter case, however, the small degrees of freedom limit interpretation (Tipton, in press). None of the other primary predictors approached significance in the single predictor fits. Among the set of secondary predictors fitted in isolation, only elderly status was statistically significant, although there were again too few degrees of freedom to support strong inference.

Among all of the single predictor fits described above, the value of τ^2 remained high (≥ 0.19), as did ω^2 (≥ 0.11), indicating that more than one moderating variable underlies the heterogeneity across studies and groups. That fact, combined with the generally low multi-collinearity among primary predictor variables, suggests that statistical power to detect moderating effects will be increased through meta-regression involving multiple simultaneous predictors.

Simultaneous fits of planned and individually significant predictors. As the initial step in this analysis, we simultaneously fitted the nine primary predictor variables (sleep status, data averaging, performance duration, break duration, training duration, linear time of training, quadratic time of training, linear time of testing, and quadratic time of testing), plus elderly status, which uniquely among the secondary predictors exhibited a potent influence on effect size. The results are shown in Table 3. Significant predictors were sleep status, data averaging, training duration, the linear and quadratic components of time of testing, and elderly status.

Performance duration exhibited a trend in the hypothesized duration, although there were too few degrees of freedom to support strong inference.

We refined that model using an iterative elimination strategy wherein the least significant predictor was removed on each iteration of the model fit (e.g., Van den Bussche, Noortgate, & Reynvoet, 2009). The resulting model, which we will refer to as the *final working model* (i.e., our final iteration of a model that is subject to future refinement) is summarized in Table 4. The regression coefficients for the surviving predictors were generally larger in the final working model than in the individual predictor fits. A notable exception is the sleep status (relative gain) predictor, for which the regression coefficient in the final working model was less than half that for the individual fits. It appears that some of the variance associated with sleep status is shared with one or more other predictors, a possibility that is examined as part of the secondary analysis of relative gain.

The estimated value of τ^2 was markedly reduced in the final working model, from 0.27 when no predictors were in the model to 0.08, suggesting that most, but perhaps not all, of the between studies heterogeneity is accounted for by the model. The within-paper residual variance, ω^2, was reduced to zero.

Empirical Implications for Post-Sleep Gain

We next explored implications of the final working model for the theory of sleep-based enhancement. The core phenomenon supporting that theory is the empirical post-sleep gain that is observed in most studies. The model allowed us to estimate: (1) the extent to which the magnitude of the post-sleep gain can jointly explained by the identified predictor variables, and (2) whether the post-sleep gain survives after adjusting for the confounding influences of those variables.

Ninety-five percent confidence intervals for the post-sleep gain are plotted in Figure 4 for representative values of three of the primary predictor variables in the final working model: *data averaging* in the calculation of pre and post-test scores, *performance duration* per performance-break cycle, and combined linear and quadratic components of *time of testing*. The plots represent predictions for sleep groups, non-elderly subjects (as elderly subjects did not exhibit a post-sleep gain) and for the modal case in which there is 360 s of training. It is important to note that these plots represent *predictions* based on the final working model fit to all 88 groups, and not the data values themselves. Also, the width of the confidence intervals across the panels varies depending on whether the values of the moderating variables are densely vs. sparsely represented in the sample. For example, the majority of groups in the sample involved 30 s performance duration per cycle (Panels c and d), yielding relatively narrow intervals, whereas 10 s per cycle (Panels a and b) is a more sparsely represented minimum value of that variable, yielding wider intervals and lower confidence.

In each panel, the pronounced moderating influence of time of testing (joint linear and quadratic terms) on post-sleep gain is apparent. The largest gain estimates are in the early afternoon with progressive and substantial drop-off toward the early morning and late evening. Panel a depicts estimates when there is zero data averaging (i.e., for the important hypothetical case in which there is no pre- or post-test averaging and thus minimal online learning confound in the gain score values) and for the minimum performance duration in the sample (10.0 s; corresponding to the lowest hypothesized reactive inhibition confound in the sample). Across the full range of time of testing in that panel, there is no statistically significant post-sleep gain. Rather, there is a significant performance worsening after sleep when testing occurs in the

morning or evening. Even in the early afternoon, the confidence intervals extend only slightly above zero.

Panel b illustrates the profound effect of data averaging on sleep gain. When averaging is set to its modal value of 60 s, and all other variables are set to the same values as in Panel a, the mean effect size is shifted upward by $d = 0.80$ at the roughly 2 pm performance peak. In Panel c, averaging is again set at 60 s, but performance duration per block is increased to its maximum (and modal) value among the groups of 30 s. There is a substantial increase ($d = 0.63$ at 2 pm) in the estimated post-sleep gain relative to Panel b, illustrating the predicted reactive inhibition effect. Finally, predictions for the jointly extreme values of averaging (120 s) and performance duration (30 s) are shown in Panel d.

Overall, Figure 4 illustrates a remarkable degree of joint predictive power of time of testing, data averaging, and performance duration per cycle. From the smallest point estimate (the 10 pm prediction in Panel a) to the largest (the 2 pm prediction in Panel d), the change in the estimated effect size is 3.1, a value that exceeds the traditional criteria for a large effect size for *d* (Cohen, 1988) by a factor of nearly 4.0. Further, when the confounding influences of data averaging and reactive inhibition are minimized (Panel a), the literature actually predicts, contrary to widely held theory, that there is minimal or no post-sleep gain.

Matched Analysis of Relative Gain

In the final working model described above, the effect of sleep status, an index of relative gain, was statistically significant, but modest. However, the experimental matching of all 23 pairs of wake and sleep groups was ignored in that analysis, and a large number of sleep groups with no matching wake groups were also included. Those factors could have impacted both the magnitude and the significance level of the estimated relative gain. Here we report analyses

limited to the experimentally matched sleep-wake groups, analyses which should yield a more veridical estimate of the relative gain effect. By focusing on only the 23 matched groups, we were also able to investigate whether differences in experimental design or other factors influence the magnitude of relative gain.

To perform these analyses, we calculated the effect size, $d_{relative}$, corresponding to the difference between the mean gain score for a sleep group and the mean gain score for the wake group. Treating the sleep and wake groups for each pair as being independent,

$$d_{relative} = (\bar{\bar{X}}_{sleep} - \bar{\bar{X}}_{wake})/S_p,$$

where S_p is square root of the pooled variance.

As shown by Viechtbauer (2007), the sampling variability of $d_{relative}$ can be approximated by,

$$Sv_{relative} = 1/q + d^2/(2m),$$

where $q = (n_{sleep}n_{wake})/(n_{sleep} + n_{wake})$ and $m = n_{sleep} + n_{wake} - 2$. The 95% confidence interval for each effect size can then be calculated as,

$d_{relative} \pm t_{95\%}(Sv_{relative})^{.5}$, with m degrees of freedom.

Among the 23 sleep-wake group pairs, 16 involved independent samples for the two groups. For the remaining seven pairs the samples were dependent; the same subjects were used in the sleep and wake groups in experimental sessions run on separate days. Following Morris & DeShon (2002), in analyses involving a mixture of independent and dependent groups, effect sizes for all groups should involve the same metric (i.e., the same type of variability measure). Our treatment of all group pairs as being independent using the equations above achieves that goal. It should be noted, however, that this approach may result in somewhat inflated measures of sampling variability for the seven within-subjects groups.

The estimated weighted mean of $d_{relative}$ over the 23 sleep-wake pairs, based on the same random effects meta analytical method that was used for the primary meta-analyses, was 0.44, p = .018, CI: 0.09, 0.79. That effect is somewhat larger than the relative gain effect in the final working model (i.e., the sleep status effect; d = 0.29), and it can be viewed as a more veridical estimate of the effect. Because data averaging, training duration, and performance duration per cycle were exactly equated for all wake and sleep pairs, those variables were not expected to, and in fact did not, significantly predict the magnitude of the relative gain (all ps > .68). Stated differently, there no significant interactions between sleep-status and the other variables. However, statistical power to detect those effects in this relatively small data set may be limited. One additional factor, however, may have important moderating effects on relative gain: experimental design.

Relative Gain as a Function of Experimental Design

Among the 23 matched groups we identified four distinct experimental designs, as illustrated in Table 5. Each design differs from the others along one of four distinct dimensions: (1) whether there is sleep deprivation for the wake group (*deprivation design*), (2) whether the sleep group involves a nap rather than a full night of sleep (*nap design*), (3) whether the time of training and testing were different for the wake and sleep groups, with delay interval held constant (*varied time design*), and (4) whether the delay between training and testing was different for the wake and sleep groups, with time of day for training and testing held constant (*varied delay design*).

Each design has strengths and potential weaknesses. The *deprivation design* involves training and testing both groups at the same time, depriving the wake group of sleep the first night after training, and testing both groups after one or two nights of recovery sleep. That design

controls for both circadian effects and homeostatic sleep drive effects. Its weakness is that any observed relative gain could reflect either a sleep-specific consolidation effect during the first night for the sleep group or impairment of non-sleep specific consolidation during the night of sleep deprivation for the wake group, due to stress or other factors (e.g. Gais, Plihal, Wagner, & Born, 2000). The *nap design* typically involves training and testing of both groups at the same time. The nap group is allowed to nap (usually for no more than 90 min) whereas the control (wake) group is not. This design controls for circadian rhythm effects. A potential weakness is that a nap may partially resolve homeostatic sleep drive that has accumulated since awakening, resulting in demonstrably improved alertness and cognitive performance that is non-specific to recently trained tasks (e.g., Brooks & Lack, 2006; note however that experimental results appear to depend on duration of nap, sleep stages involved, and delay between awakening and testing). In the *varied time design*, the wake group is trained in the morning, the sleep group is trained at night, and both groups are tested after the same delay interval (typically 8 or 12 hrs). Like the deprivation and nap designs, it controls for delay interval between training and testing. A weakness is that the design does not control for time of training or testing in the wake and sleep groups, and it is thus vulnerable to both circadian and homeostatic confounds. Finally, the *varied delay design* involves training of both groups at the same time, testing the wake group several hrs later on the same day, and testing of the sleep group 24 hrs (or a multiple of 24 hrs) after the wake group is tested. It fully controls for both circadian and homeostatic confounds. A potential weakness is that the delay between training and the test is greater for the sleep than for the wake group, possibly resulting in greater forgetting for the sleep group that could counteract any sleep consolidation effect.

Among the included groups, there were nine varied time experiments, one deprivation experiment, eleven nap experiments, and two varied delay experiments (see Table 1). One of the varied time experiments (Tucker et al., 2011) also involved a varied delay, although not the 24 hr delay difference for test time that is definitional for that group. Further, that design matched closely to the other varied time designs with respect to time of testing (9 pm for the wake group and 9 am for the sleep group). By our definitions, then, the most appropriate design category for that study is varied time. One of the varied delay experiments (extracted from Ashtamker & Karni, 2013) involved a 22.5 hr rather than 24 hr delay difference between wake and sleep groups, but nevertheless best matched the varied delay design.

To explore the effect of experimental design, we constructed a forest plot of the relative gain confidence intervals for all 23 sleep-wake pairs, ordered by experimental design, as shown in Figure 5. In most cases, statistical inference based on the confidence intervals matches that based on the hypothesis tests that were conducted in the papers (when reported); that is, if a two-tailed *t*-test in the paper describing the experiment had rejected the null hypothesis at $\alpha = 0.05$, then the corresponding confidence interval in Figure 5 did not include zero, and vice versa. There were some exceptions, typically cases in which the null hypothesis was rejected in the paper but zero was nevertheless marginally within the confidence interval. Those discrepancies likely reflect the fact that the confidence intervals on effect sizes are approximate and become less precise for small samples (e.g., Viechtbauer, 2007).

The forest plot suggests that experimental design is a potent factor in determining effect size. Only the varied time design has to date yielded a consistent and robust relative sleep gain effect. There is a slight trend toward a relative gain effect for the nap design (random effects analysis limited to nap groups yielded $p = 0.29$), but that trend is driven primarily by two of the

11 sleep-wake pairs. There is virtually no evidence of a relative gain effect for either the deprivation or varied delay designs.

Influence of Time of Testing and Delay

Based on the final working model, an important factor that may influence the observed relative gain in at least one experimental design is time of testing for the sleep vs. wake groups of each pair. In particular, the varied time design appears to be vulnerable to a time of testing confound. The mean time of testing for the varied time experiments was 9:03 pm for the wake groups and 9:36 am for the sleep groups, a difference of about 12.5 hrs. In contrast, for the other three designs the mean times of testing for the wake and sleep groups were identical, or nearly so (deprivation: 1:30 pm and 1:30 pm; nap: 4:49 pm and 4:49 pm; varied delay: 3:15 pm and 2:30 pm). We can estimate the degree to which the relative gain effect for each design might be due to time of testing by using the linear and quadratic regression coefficients from the final working model fitted to all groups (see Table 4). As shown in Figure 6, the predicted time of testing effect is substantial for the varied time design but, as expected, is negligible for the other three designs.

The current analyses also allow us to explore the potential confounding influence of delay in the varied delay design. The wake group in that design would be trained and tested on the same day, with a delay between tests of several hours. The sleep group is tested 24 hrs after the same group. Forgetting after sleep and during the waking hours prior to the test session may offset sleep consolidation effects in that design, resulting in an underestimation of the relative gain. The current results suggest, however, that the confounding influence of the different delay periods in that design is minimal. There was no trend toward an effect of delay (range: 8 to 72 hrs) when that predictor was fitted in isolation in the primary analyses (see Table 2). Further, when delay was added to the final working model, its point estimate approached zero (0.0002),

corresponding to an expected effect size change over the 24-hr delay difference between wake and sleep groups of only 0.005. Thus delay, the most obvious potential confounding factor in the varied delay design, may have negligible influence on the observed relative gain.

In summary, the current analyses limited to the 23 matched sleep-wake groups confirms the overall empirical relative gain effect. The results also raise the possibilities, however, that relative gain effects may be conditional on experimental design and that the one design that exhibits a consistent and robust gain effect (varied time) may be susceptible to a time of testing confound.

Discussion

The discussion is organized into four sections. First, we interpret results for each of the identified non-consolidation predictors, weigh evidence for them based jointly on prior experimental work and the current meta-analytical results, and consider what effect those predictors might have, if any, on sleep consolidation processes. Second, we draw theoretical conclusions about the nature of sleep consolidation in motor learning with respect to both the sleep-based enhancement hypothesis and the stabilization of learning hypothesis. Third, we describe a continuity test based on curve fitting that circumvents the need to compute pre-post difference scores, fully eliminates confounds due to averaging over online learning, and can clarify interpretation of duration of training effects in future work. Fourth, we make recommendations for experimental design and data analysis for future work in this area. Finally, we note implications of our behavioral findings for correlations between sleep gain and electrophysiological measures that have sometimes been observed in this literature.

Interpretation of Variables Moderating the Post-Delay Gain

Beyond sleep status, five variables were identified as predictors of post-delay and (or) relative gain effects: *data averaging*, *performance duration*, *time of testing*, *training duration*, and *elderly status*. Each is discussed further below.

Data averaging. Given the well-established properties of learning curves, data averaging in calculation of pre- and post-test score must, to some extent, influence the magnitude of the post-delay gain. The current results show, for the first time, that the confounding effect of that averaging is substantial. Indeed, the averaging artifact alone appears to account for the majority of the post-delay gain effect. It would be straightforward to show the same effects within most data sets in the literature. Consider, for example, the massed and spaced practice groups in Figure 2**.** Because performance in both groups decreases monotonically, greater averaging would clearly result in larger gain scores. In our view, the only viable way to resolve the averaging problem in future investigation of post-sleep gain effects is to employ curve fitting rather than pre-post difference scores to estimate gain effects, a topic that will be discussed in a section below.

Duration of training. Training duration exhibited the hypothesized negative correlation with the post-delay gain. Also, from a purely rational perspective, the negative correlation must hold in the population given only the very well-supported assumption that performance improvement is a monotonically decreasing function of practice. Only in long training sessions is that assumption potentially wrong (due to massive fatigue build-up), and there is no systematic evidence that it fails in this literature. The negative correlation may also reflect a greater degree of sleep consolidation in short vs. long training duration designs, a topic to which we will also return in the section on curve fitting below.

Time of testing. The meta-analytical results suggest that time of testing, but not time of training, has a large influence on the post-delay gain. With respect to training, that conclusion is consistent with the null effects of morning vs. evening training times that have been reported in a number of studies (Albouy et al., 2013; Brawn et al., 2010; Doyon et al., 2009; Korman et al., 2003). Given that selective time of testing effects have not previously been demonstrated, however, it will be important to experimentally confirm that effect and to more fully tease apart the effects of time of training and testing on not only gain scores but also performance during the training and test sessions.

It is an open question why time of testing would have a more potent effect on performance than does time of training. One possibility is that situational factors that are unique to the training session may allow subjects to maintain a high level of alertness and motivation at all times of day, largely overriding or suppressing circadian and homeostatic influences. During the training session, subjects may have evaluation apprehension about being in a novel lab setting. The task is also novel, initially requiring executive processes and possibly declarative memory engagement. During training subjects undergo rapid and presumably reinforcing learning. In our lab, it has not been unusual for subjects to spontaneously comment that the training session was mildly “fun” or “game-like.” None of those subject mood states or experiences is likely to be present to the same extent during the test session: the lab context is familiar, performance improvements are smaller, and the task is presumably less engaging. Under those conditions, the effects of circadian and homeostatic factors may be stronger. Consistent with that account, Hull et al. (2003) demonstrated that alertness and motivation can exert influences on task performance that are separable from both circadian and homeostatic

effects. Their data also suggest that high alertness levels may decrease the influence of those factors on performance.

Whatever its basis, a selective time of testing effect raises an important question regarding how to the theoretically assess gain scores effects. At what time of day should the test be given such that the observed post-sleep gain score is not confounded by circadian or homeostatic influences? A conclusive answer awaits theory development. Above, we have implicitly suggested one hypothesis, namely that (1) during training (but not testing), subjects are able to suppress negative circadian and homeostatic effects, yielding near optimal performance throughout the day, and (2) during testing, optimal performance occurs in the early afternoon, when joint circadian and homeostatic effects may be most favorable to task performance. If correct, that account implies that the potential time of testing confound in the observed gain score will be minimized then testing occurs in the early afternoon.

Reactive inhibition. Consistent with the hypothesis that reactive inhibition can have a confounding influence on the post-delay gain, longer performance durations within each performance-break cycle predicted a larger post-delay gain. Inference regarding that finding are qualified, however, by the low degrees of freedom for performance duration in the final working model fit. In essence, the common use of 30 s performance durations in the literature limits inference at the meta-analytical level. A strong case can be nevertheless be made for a potent influence of reactive inhibition on gain scores given the two randomized experiments on that topic that were discussed earlier: Rickard et al. (2008) and Brawn et al. (2010).

The current results suggest that duration of performance within each cycle drives the reactive inhibition effect, whereas duration of the break may not. The latter result, however, likely reflects the limited range of break durations in the sample. As duration of break

approaches zero, it must have an influence on the degree to which reactive inhibition is resolved between performance blocks. At the other extreme is the Brawn et al. (2010) demonstration that a five min break between training and the pre-test substantially reduces the performance difference between massed and spaced conditions.

Elderly status. For elderly subjects, there is a marked reduction in the magnitude of the post-delay gain that virtually eliminates the post-sleep gain effect, although a relative gain effect is observed in most cases. That result raises the possibility that sleep-based enhancement, but not sleep-based stabilization, is impaired in the elderly. An alternative account, however, was advanced by Tucker et al. (2011). They observed that elderly subjects exhibit markedly worse performance than do young subjects on the first few blocks of the post-test (the traditional post-test period used for analysis) in both wake and sleep conditions. However, rapid performance improvement after those initial test blocks restores performance to levels that, relative to the pre-test, are statistically indistinguishable from that of young subjects. A highly analogous pattern is also apparent in Fogel et al. (2014). Those results raise the possibility that elderly subjects have an unusually long warm-up period in the test session, but that any effect that sleep may have on motor sequence performance occurs equivalently for the young and the elderly.

Do the Moderating Variables Causally Influence Sleep Consolidation?

There are two conceptually distinct accounts of the effect of each of the moderating variables discussed above. One possibility is that their effects are independent of any sleep-specific consolidation mechanism. Alternatively, part or all of the effect of a given variable may reflect a direct influence on the magnitude of sleep-based enhancement. Further consideration supports the former account for the data averaging, time of testing, and performance duration (indexing reactive inhibition) predictors. Data averaging occurs after data are collected, is

unknown to subjects, and thus cannot moderate consolidation processes. Similarly, because testing occurs after sleep, time of testing cannot plausibly have a causal influence on sleep consolidation processes (although time of training in principle could have). Reactive inhibition effects also appear to be largely independent of any sleep consolidation that may occur. In Rickard et al. (2008, Experiment 1), the within-block reactive inhibition effects over both training and testing sessions were indistinguishable for wake and sleep groups, suggesting that sleep played no special role in resolving reactive inhibition. Further, Brawn et al. (2010) showed that an awake post-training break of only 5 min was sufficient to eliminate most of the performance differences between groups trained under massed and spaced conditions, suggesting that reactive inhibition effects can resolve relatively quickly and before sleep onset.

In contrast to those variables, it is plausible that training duration moderates the magnitude of sleep-based enhancement. Short duration training will generally allow more opportunities for additional learning (i.e., achieved skill after short duration training will be further from asymptote compared to long duration training) and the magnitude of sleep-based enhancement may depend on the amount of new learning that is possible. Alternatively, as we hypothesized earlier, pre- and post-test averaging over a steeper section of the learning curve in short duration training designs may exacerbate online learning confounds relative to long duration training. Those competing accounts of training duration can be tested in future work by using the continuity analysis described next.

Eliminating online learning confounds using a continuity test. With respect to testing the sleep-based enhancement hypothesis, the data averaging confound can profoundly bias results. How can that problem be resolved? A strategy of minimizing data averaging (e.g., to the last 10 s of training and the first 10 s of test) would mitigate but not fully eliminate the online

learning problem. Averaging over short durations also has the negative consequence of relatively low statistical power in the gain score test. It can also be highly sensitive to transient performance patterns, such as warm-up effects on initial test blocks. Addressing warm-up effects by eliminating data from the first test block(s) prior to calculation of pre-post gain scores (e.g., Fischer, Hallschmid, Elsner, & Born, 2002) is ill-advised because it may exacerbate the online learning confound in the pre-post difference scores (i.e., learning make occur on warm-up blocks that facilitates subsequent performance).

Fortunately, online learning confounds can be fully eliminated in future work by abandoning the use of pre-post difference scores and instead using curve fitting procedures to compare performance levels prior to and after the delay. Two curve fitting approaches are potentially viable. First, an appropriate empirical function, such as the three-parameter power function, can be fitted to training data for each subject and the gain score analysis can be based on the difference between the predicted post-test performance (based on extrapolation of the training data fit) and the observed performance on one of more test blocks (for a recent application of that approach see Adi-Japha, Badir, Dorfberger, & Karni, 2014). Under ideal conditions, that approach can fully resolve the online learning confound due to data averaging. It has the potential weakness, however, that subject-level data is often noisy and cannot be assumed to follow an exact curve (e.g., Rickard, 2004), potentially leading to inaccurate extrapolations.

Alternatively, statistical inference on data averaged over subjects is possible using a continuity test; that is, by fitting a learning curve simultaneously to training and test data, along with a continuity parameter that tests for an abrupt change in performance between the training and test sessions. Here, we introduce and apply one candidate approach to continuity testing in the hope of promoting future application.

The null case in the continuity test assumes no effect of the delay between the training and test sessions. As noted previously, performance improvement in that case should be well described across both training and test sessions by a smooth function with monotonically increasing performance improvement rate. Across a variety of tasks, the best fitting smooth function for data averaged over subjects is the three-parameter power function (Newell & Rosenbloom, 1981; see Heathcote et al., 2000, for discussion of exponential and hybrid exponential-power functions as alternatives). For cases in which learning gives rise to decreases in the value of the dependent variable (e.g., mean latency per key press), the power function takes the form,

$$p = a + b*N^{-c},$$

where p is performance level, N is the trial or block number, a is asymptotic performance, b is the amount of improvement that is possible with hypothetically infinite practice, and c is the non-linear improvement rate.

The continuity test involves fitting the selected smooth function across training and test sessions, along with a discontinuity parameter between the last training session data point and the first test session data point. For the power function, the most straightforward equation for the continuity test is:

$$p = k*(a + b*N^{-c}) + (1-k)*(a + b*N^{-c} - y), \quad (2)$$

where k takes a value of 1 for training session data and 0 for the test session data and y is the discontinuity parameter. If the estimate of y is small and non-significant, then a single smooth function – the three-parameter power function in this example – is sufficient to explain the performance curve across both the training and test sessions, and no effects of the delay between sessions (sleep related or otherwise) are indicated. If, on the other hand, y accounts for a

statistically significant portion of the variance in the model fit, then an effect of the delay on performance can be inferred and the value of *y* is a direct estimate of its magnitude.

Beyond eliminating online learning confounds, the continuity test has several advantages over other approaches. First, it eliminates the need to define an arbitrary range of data as constituting the pre- and post-tests. Second, the estimate of *y* is a purer measure of the post-delay gain than is a pre-post difference score, and it is more directly comparable across studies. Third, unlike averaged pre-post gain scores, the results of the continuity test should be minimally sensitive to a strategy of eliminating the first block of the test data to accommodate possible warm-up effects, provided that the true experimental block numbers are respected in the curve fitting. Fourth, because the continuity test is free of data averaging confounds and automatically accommodates differential performance improvement rates at the end of short vs. long duration training, it allows the two candidate accounts of the training duration effect to be disambiguated. If, for example, a continuity analysis based on a well-fitting practice function were to reveal systematic decreases in the magnitude of post-sleep gain (*y*) over increasing training duration, then the hypothesis that training duration moderates sleep-based enhancement would remain viable, whereas the alternative hypothesis advanced here, in which the training duration effect is solely a data averaging artifact, would not.

Finally, the continuity test can be statistically powerful. To illustrate, the power function continuity test was applied to the spaced practice data of Rickard et al. (2008; refer to the current Figure 2). Data from the first 10 s test block (block 37) in that group exhibit an apparent warm-up effect and thus were not fitted. The estimate for *y* was a non-significant -0.28 ms, CI: -2.4, 1.8. Hence, the continuity analysis is consistent with the conclusion of no post-sleep gain in the spaced group. Note also that the confidence interval is narrow, indicating high statistical power.

We also performed a continuity fit to the massed training data from Rickard et al. (2008), as shown in Figure 2. As for the fit to the spaced data, the first 10 s of test performance were eliminated to allow for possible warm-up effects. The discontinuity estimate was large and highly significant for that group: y = -28.0 ms, CI: -36.7, -17.3. The reactive inhibition effect for that group, however, suggests that the discontinuity is not due to sleep enhancement.

Summary of Implications for Sleep Consolidation in Motor Sequence Learning

The sleep-based enhancement hypothesis. A positive post-sleep gain is a necessary (but not sufficient) condition for inferring sleep-based enhancement of learning. The current results indicate that designs with an early afternoon test session are most conducive to observing that gain. Even then, however, our meta-analytical results are consistent with the hypothesis of no post-sleep gain when confounds due to online learning and reactive inhibition are eliminated or substantially reduced. The spaced training sleep group of Rickard et al. (2008), for which the test session occurred in the early afternoon and a continuity analysis has been applied (see Figure 2), is the only experiment to date in which all of those criteria have been met. Consistent with the meta-analyses, no trend suggesting sleep enhancement effect was observed. In a sense, then, the literature as a whole predicts the Rickard et al. results. We thus conclude that, to date, there is no compelling evidence for – nor even a discernable trend suggesting – that there is sleep-based enhancement in the domain of explicit motor sequence learning.

The sleep-based stabilization hypothesis. Both the main meta-analyses and the supplementary analysis of matched sleep-wake groups identified a statistically significant relative gain effect, suggesting that sleep may play a role in stabilization of learning rather than enhancement. Those analyses also raise the possibilities, however, that the magnitude of relative gain may depend on experimental design and that time of testing may constitute a serious

confound for the varied time design. We thus view the current literature as suggestive but not conclusive with respect to sleep-based stabilization.

Recommendations for Experimental Design and Data Analysis

Minimizing confounding influences on the post-sleep gain. If the research goal is to estimate the post-delay or post-sleep gain, then the following recommendations are offered:

1. Estimation of gain effects using pre-post difference scores should be abandoned in favor of curve fitting techniques. The continuity test is the preferred approach in our view because it is least likely to yield biased results (given that the continuity curve fits the averaged data well across both training and test sessions) and it yields a pure estimate of the magnitude of the post-delay gain.
2. The traditional design involving 30 s-30 s performance-break cycles should be abandoned given the evidence that it results in a reactive inhibition confound, and alternative designs with reduced performance duration per block used instead. One promising possibility is to switch to 10 s performance durations for each performance-break cycle. That design appears sufficient to eliminate at least the majority of the reactive inhibition effect (Brawn et al., 2010; Rickard et al., 2008). It also has the side benefit of producing less variable training performance (e.g., compare the massed and spaced groups in Figure 2), facilitating application of curve fitting techniques and likely improving statistical power. Should future studies adopt 10 s performance duration designs, confirmation that reactive inhibition is largely eliminated could be obtained through analysis of within-block reaction time patterns (see Rickard et al., 2008).

3. The potentially strong influence of time of testing on gain scores should be explicitly acknowledged as a factor that may limit theoretical inference. Based on our results, testing in the early afternoon will favor observation of post-sleep gain. If no gain is observed in that case, the results can be confidently interpreted as challenging the sleep-based enhancement hypothesis. Although our results suggest that time of training is not critical to the observed gain score, it is nevertheless prudent when focusing solely on the post-sleep gain and the sleep enhancement hypothesis to use a 24-hr delay, or some multiple thereof, that can fully equate training and testing sessions with respect to physiological circadian and homeostatic factors (although not necessarily alertness or motivation; see prior discussion of the time of testing effect).
4. To facilitate interpretation of results, scatter-plots of all block-level training and test data, along with the fitted performance curves, should be included.
5. To facilitate curve fitting, the test session should involve multiple blocks, a reasonable default being the same number of blocks as the training session. That design approach would also provide insight into possible differences in the shape of the practice curve during testing vs. training, differences that may have theoretical implications (e.g., Tucker et al., 2011).

Minimizing confounding influences on the relative sleep gain. We suggest that all of the design guidelines described above be followed in this case, where applicable, to minimize potential confounding influences and to provide veridical gain effect estimates for both wake and sleep groups. Although there is no evidence in the current analyses that factors such as reactive inhibition and training duration have differential impact for wake vs. sleep groups, small effects along those lines may not have been detected.

The most commonly used design to explore relative gain, the varied time of day design, should in our view be abandoned given the current evidence that differences in time of testing for wake and sleep groups in that design may account for at least a large portion of the relative gain effect (see Figure 6). Going forward, the nap and varied delay designs appear to be preferable. Both designs fully equate time of training and time of testing for the wake and sleep groups, effectively eliminating the possibility that circadian rhythms can differentially influence gain scores for wake and sleep groups.

One limitation of nap studies as conducted to date is that they have not tested for the possibility that the relative gain (when observed) is the result of improvement of general cognitive function following a nap (i.e., due to partial resolution of homeostatic sleep drive during the nap) as opposed to sleep-based consolidation of prior learning. It may be possible to modify the design to accommodate that possibility, however. One approach would be to introduce a secondary motor task, in the test session only, for both wake and sleep groups (such a task could also be used in other designs, but is particularly relevant to the nap design issues discussed here). If performance on the secondary task is better for the nap than for the wake control groups, then the theoretical interpretation of any observed relative gain effect for the primary task is unclear. The critical statistical test for relative sleep gain in that case would require some form of adjustment for the sleep group's advantage on the secondary task. A potential complication of the secondary task approach is that, as implied by the current meta-analyses, homeostatic effects in the test session may be smaller for the novel secondary task than for the non-novel primary task (e.g., subject may be more alert during the secondary task; see discussion of time of testing effects). As such, the sleep-wake difference for the secondary task may be attenuated, leading to overestimated relative gain estimates even after secondary task

performance is used to adjust the relative gain estimate. Other limitations of nap designs are that nap durations do not encompass the entire sleep stage cycle, and that hormonal concentrations during naps differ from those during a full night's sleep.

The varied delay design has the primary candidate weakness that the longer delay for the sleep group may yield greater forgetting that masks sleep consolidation effects. Based on the current results, however, the delay interval appears to have inconsequential effects on gain scores over a change of up to 72 hours, suggesting that the estimation of relative gain in that design is minimally influenced by delay interval. To assure that reactive inhibition effects have fully resolved between training and test session in the wake group, the delay for that group should at least be several hours.

Perhaps the most promising approach to reaching a strong theoretical conclusion about relative gain is to jointly pursue both the modified nap designs and the varied delay design, as well as any new designs that are plausibly free of major confounding influences. If none of those designs generates evidence for a relative gain effect, then the hypothesis of sleep-based stabilization of learning should be rejected. Conversely, if there proves to be converging evidence favoring a relative gain effect among all such designs, then sleep-specific consolidation in the form of stabilization would be well-supported. Divergent results from different designs, as currently exists for the varied time, nap, and varied delay designs, would invite deeper investigation of the underlying causal factors before any strong theoretical conclusions regarding relative gain effects would be warranted.

Implications for Electrophysiological Studies

Although our focus in this review is behavioral sleep consolidation, it is important to also consider implications of our findings for research on correlations between behavioral sleep gain

and electrophysiological patterns observed during sleep. Two aspects of sleep have been of most interest in that literature: stage 2 NREM sleep and sleep spindles.

In our sample of 34 papers, correlations between post-sleep gain and stage 2 NREM (measured either as duration or percent of total sleep) were reported for 10 independent nap and full night sleep groups across eight papers, with positive and statistically significant results (at $\alpha = 0.05$) for only four groups. Correlation coefficients were not reported or were reported as an inequality for six of the non-significant cases, precluding a more formal meta-analysis. Correlation analyses have also been reported for six additional groups in four studies that did not meet our inclusion criteria (Fischer et al., 2002; Fischer, Nitschke, Melchert, Erdmann, & Born, 2005; Holz, 2012; Wilhelm et al., 2011), none of which were statistically significant. We conclude that the evidence for a correlation between stage 2 NREM sleep and post-sleep gain in this literature is not yet compelling, a conclusion that is consistent with the current behavioral finding of no significant post-sleep gain after adjusting for confounding variables.

Similarly, significant correlations between sleep spindles and post-sleep gain were observed for only four of 43 cases over 11 groups and nine papers. Further complicating interpretation, multiple correlations were performed for most of those groups using different measures of spindle activity (including density, spectral power, count, brain region, and quarter of the night). Spindle measures also varied in the use of change measures (with pre-training sleep data as a baseline). Given that no adjustments of significance levels to accommodate multiple comparisons were made in most of those studies, the evidence to date for a correlation between post-sleep gain and sleep spindle activity is not compelling.

If in the future robust correlations between some aspect of sleep and post sleep gain are established, theoretical interpretation will have to be reconciled with the current finding of no

behavioral sleep-based enhancement. A more plausible account of such findings, in light of the current results, would be that the electrophysiological correlations with post-sleep gain are a signature of sleep-based stabilization processes, or perhaps reflect brain activity that is somehow causally related to task training but is not related to consolidation of learning.

Conclusions

The claim that sleep plays a critical role in the enhancement of motor skill learning has, in the public eye and perhaps among most researchers, moved beyond hypothesis and toward accepted fact. Increasingly, the literature has treated sleep-based enhancement as a given and has focused on exploring the generalizability of that phenomenon and on more detailed theoretical accounts. The results of the current meta-analyses, however, reveal a potential weakness in the foundation for much of that work: when confounding variables that are independent of any possible sleep consolidation effect are factored out, there is no evidence in the literature for a performance gain that can be attributed to sleep.

Although our results do not preclude the possibility that sleep-based enhancement occurs in other motor domains (e.g., rotary pursuit, figure tracing, implicit sequence learning), it seems likely that the current results will generalize to those tasks, at least in part. Averaged pre- and post-test data are commonly used to assess post-sleep gain throughout the motor consolidation literature, likely resulting in some degree of contamination from online learning. The effects of time of testing, performance duration per cycle, and training duration effects are also strong candidates for generalization to other tasks.

Although sleep-based enhancement of learning was not supported, we did observe a relative gain effect that is consistent with sleep-based stabilization of learning, raising the possibility that sleep-specific consolidation has the same behavioral effect in the procedural

domain that it does in the declarative domain. Strong inference along those lines, however, should await further exploration of the possible influence of experiment design on the magnitude of the relative gain effect.

References

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

* Adi-Japha, E., Badir, R., Dorfberger, S., & Karni, A. (2014). A matter of time: Rapid motor memory stabilization in childhood. *Developmental Science, 17*(3), 424-433. doi:10.1111/desc.12132

* Albouy, G., Sterpenich, V., Vandewalle, G., Darsaud, A., Gais, S., Rauchs, G., Desseilles, M., Boly, M., Dang-Vu, T., Balteau, E., Degueldre, C., Philips, C., Luxan, A., & Maquet, P. (2013). Interaction between hippocampal and striatal systems predicts subsequent consolidation of motor sequence memory. *PloS One, 8*(3), e59490. doi:10.1371/journal.pone.0059490; 10.1371/journal.pone.0059490

Allen, P. A., Grabble, J., McCarthy, A., Bush, A. H., & Wallace, B. (2008). The early bird does not get the worm: Time-of-day effects on college student's basic cognitive processing. *The American Journal of Psychology, 21(2),* 551 – 564.

* Ashtamker, L., & Karni, A. (2013). Motor memory in childhood: Early expression of consolidation phase gains. *Neurobiology of Learning and Memory, 106,* 26-30.

* Barakat, M., Doyon, J., Debas, K., Vandewalle, G., Morin, A., Poirier, G., Martin, N., Lafortune, M., Karni, A., Ungerleider, L.G., Benali, H., & Carrier, J. (2011). Fast and slow spindle involvement in the consolidation of a new motor sequence. *Behavioural Brain Research, 217*(1), 117-121. doi:10.1016/j.bbr.2010.10.019

Blatter, K., & Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology & Behavior, 90*(2-3), 196-208. doi:10.1016/j.physbeh.2006.09.009

* Blischke, K., Erlacher, D., Kresin, H., Brueckner, S., & Malangré, A. (2008). Benefits of sleep in motor learning – prospects and limitations. *Journal of Human Kinetics, 20,* 23-36. doi: 10.2478/v10078-008-0015-9

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111. doi:10.1002/jrsm.12

* Brawn, T. P., Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2010). Consolidating the effects of waking and sleep on motor-sequence learning. *The Journal of Neuroscience, 30*(42), 13977-13982. doi:10.1523/JNEUROSCI.3295-10.2010

Brooks, A., & Lack, L. (2006). A brief afternoon nap following nocturnal sleep restriction: Which nap duration is most recuperative? *Sleep, 29*(6), 831-840.

* Cai, D. J., & Rickard, T. C. (2009). Reconsidering the role of sleep for motor memory. *Behavioral Neuroscience, 123*(6), 1153-1157. doi:10.1037/a0017672

* Cash, C. D. (2009). Effects of early and late rest intervals on performance and overnight consolidation of a keyboard sequence. *Journal of Research in Music Education, 57*(3), 252-266. doi:10.1177/0022429409343470

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale: Laurence Erlbaum Associates.

Diekelmann, S., & Born, J. (2007). One memory, two ways to consolidate? *Nature Neuroscience, 10*(9), 1085-1086. doi:10.1038/nn0907-1085

* Dorfberger, S., Adi-Japha, E., & Karni, A. (2007). Reduced susceptibility to interference in the consolidation of motor memory before adolescence. *PloS One, 2*(2), e240. doi:10.1371/journal.pone.0000240

* Dorfberger, S., Adi-Japha, E., & Karni, A. (2009). Sex differences in motor performance and motor learning in children and adolescents: An increasing male advantage in motor learning and consolidation phase gains. *Behavioural Brain Research, 198*(1), 165-171. doi:10.1016/j.bbr.2008.10.033

* Doyon, J., Korman, M., Morin, A., Dostie, V., Hadj Tahar, A., Benali, H., Karni, A., Ungerleider, L.G., & Carrier, J. (2009). Contribution of night and day sleep vs. simple passage of time to the consolidation of motor sequence and visuomotor adaptation learning. *Experimental Brain Research, 195*(1), 15-26. doi:10.1007/s00221-009-1748-y

* Feld, G. B., Wilhelm, I., Ma, Y., Groch, S., Binkofski, F., Mölle, M., & Born, J. (2013). Slow wave sleep induced by GABA agonist tiagabine fails to benefit memory consolidation. *Sleep, 36*(9), 1317-1326.

Fischer, S., Hallschmid, M., Elsner, A. L., & Born, J. (2002). Sleep forms memory for finger skills. *Proceedings of the National Academy of Sciences, 99*(18), 11987-11991. doi:10.1073/pnas.182178199

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 Neuroscience, 25*(49), 11248-11255. doi:10.1523/JNEUROSCI.1743-05.2005

* Fogel, S. M., Albouy, G., Vien, C., Popovicci, R., King, B. R., Hoge, R., Jbabdi, S., Benali, H., Karni, A., Maquet, P., Carrier, J., & Doyon, J. (2014). fMRI and sleep correlates of the age-related impairment in motor memory consolidation. *Human Brain Mapping, 35*(8), 3625-3645. doi:10.1002/hbm.22426

Gais, S., Plihal, W., Wagner, U., & Born, J. (2000). Early sleep triggers memory for early visual discrimination skills. *Nature Neuroscience, 3*(12), 1335-1339. doi:10.1038/81881

* Genzel, L., Dresler, M., Wehrle, R., Grozinger, M., & Steiger, A. (2009). Slow wave sleep and REM sleep awakenings do not affect sleep dependent memory consolidation. *Sleep, 32*(3), 302-310.

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.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. doi:10.1037/0033-2909.92.2.490

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. doi: 10.1002/jrsm.5

Holz, J., Piosczyk, H., Landmann, N., Feige, B., Spiegelhalder, K., Riemann, D., Nissen, C., & Voderholzer, U. (2012). The timing of learning before night-time sleep differentially affects declarative and procedural long-term memory consolidation in adolescents. *PloS One, 7*(7), e40963. doi:10.1371/journal.pone.0040963

Hotermans, C., Peigneux, P., de Noordhout, A. M., Moonen, G., & Maquet, P. (2006). Early boost and slow consolidation in motor skill learning. *Learning & Memory, 13*(5), 580-583. doi:10.1101/lm.239406

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

Hull, J. T., Wright Jr., K. P., Czeisler, C. A. (2003). The influence of subjective alertness and motivation on human performance independent of circadian and homeostatic regulation. *Journal of Biological Rhythms, 18,* 329-338.

Jewett, M. E., Wyatt, J. K., Ritz-De Cecco, A., Khalsa, S. B., Dijk, D. J., & Czeisler, C. A. (1999). Time course of sleep inertia dissipation in human performance and alertness. *Journal of Sleep Research, 8*(1), 1-8.

Keisler, A., Ashe, J., & Willingham, D. T. (2007). Time of day accounts for overnight improvement in sequence learning. *Learning & Memory, 14*(10), 669-672. doi:10.1101/lm.751807

Kleitman, N. (1933) Studies on the physiology of sleep: VIII. Diurnal variation in performance. *American Journal of Physiology, 104*, 449-456.

* Korman, M., Dagan, Y., & Karni, A. (in preparation). Motor learning consolidation gains in the elderly are under expressed unless a nap is afforded.

* Korman, M., Doyon, J., Doljansky, J., Carrier, J., Dagan, Y., & Karni, A. (2007). Daytime sleep condenses the time course of motor memory consolidation. *Nature Neuroscience, 10*(9), 1206-1213. doi:10.1038/nn1959

* Korman, M., Raz, N., Flash, T., & Karni, A. (2003). Multiple shifts in the representation of a motor sequence during the acquisition of skilled performance. *Proceedings of the National Academy of Sciences, 100*(21), 12492-12497. doi:10.1073/pnas.2035019100

* Kuriyama, K., Stickgold, R., & Walker, M. P. (2004). Sleep-dependent learning and motor-skill complexity. *Learning & Memory*, *11*(6), 705-713. doi:10.1101/lm.76304

* Marshall, L., Helgadottir, H., Molle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature, 444*(7119), 610-613. doi:10.1038/nature05278

* Mednick, S. C., Cai, D. J., Kanady, J., & Drummond, S. P. (2008). Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory. *Behavioural Brain Research, 193*(1), 79-86. doi:10.1016/j.bbr.2008.04.028

Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., & Wixted, J. T. (2011). An opportunistic theory of cellular and systems consolidation. *Trends in Neurosciences, 34*(10), 504-514. doi:10.1016/j.tins.2011.06.003; 10.1016/j.tins.2011.06.003

* Mednick, S. C., McDevitt, E. A., Walsh, J. K., Wamsley, E., Paulus, M., Kanady, J. C., & Drummond, S. P. (2013). The critical role of sleep spindles in hippocampal-dependent memory: A pharmacology study. *The Journal of Neuroscience, 33*(10), 4494-4504. doi:10.1523/JNEUROSCI.3127-12.2013; 10.1523/JNEUROSCI.3127-12.2013

Monk, T. H. (2005). The post-lunch dip in performance. *Clinics in Sports Medicine, 24*(2), e15-23, xi-xii. doi:10.1016/j.csm.2004.12.002

* Morin, A., Doyon, J., Dostie, V., Barakat, M., Hadj Tahar, A., Korman, M., Benali, H., Karni, A., Ungerleider, L.G., & Carrier, J. (2008). Motor sequence learning increases sleep spindles and fast frequencies in post-training sleep. *Sleep, 31*(8), 1149-1156.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125.

Nemeth, D., Janacsek, K., Londe, Z., Ullman, M. T., Howard, D. & Howard, J. H. Jr. (2009). Sleep has no critical role in implicit sequence learning in young and old adults. *Experimental Brain Research, 201*(2), 351-358. doi: 10.1007/s00221-009-2024-x

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

* Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PloS One, 2*(4), e341. doi:10.1371/journal.pone.0000341

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9*(4), 534-547. doi:10.1162/jocn.1997.9.4.534; 10.1162/jocn.1997.9.4.534

* Rasch, B., Buchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science, 315*(5817), 1426-1429. doi:10.1126/science.1138581

* Rasch, B., Pommer, J., Diekelmann, S., & Born, J. (2009). Pharmacological REM sleep suppression paradoxically improves rather than impairs skill memory. *Nature Neuroscience, 12*(4), 396-397. doi:10.1038/nn.2206

Rickard. T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 65–82.

* Rickard, T. C., Cai, D. J., Rieth, C. A., Jones, J., & Ard, M. C. (2008). Sleep does not enhance motor sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 834-842. doi:10.1037/0278-7393.34.4.834

Robertson, E. M., Pascual-Leone, A., & Miall, R. C. (2004). Current concepts in procedural consolidation. *Nature Reviews Neuroscience, 5*(7), 576-582. doi:10.1038/nrn1426

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York, NY: Russell Sage Foundation.

* Sheth, B. R., Janvelyan, D., & Khan, M. (2008). Practice makes imperfect: Restorative effects of sleep on motor learning. *PloS One, 3*(9), e3190. doi:10.1371/journal.pone.0003190

Smith, C. (2001). Sleep states and memory processes in humans: Procedural versus declarative memory systems. *Sleep Medicine Reviews, 5*(6), 491-506. doi:10.1053/smrv.2001.0164

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature, 437*(7063), 1272-1278. doi:10.1038/nature04286

Strube, M. J., & Hartmann, D. P. (1983). Meta-analysis: Techniques, applications, and functions. *Journal of Consulting and Clinical Psychology, 51*(1), 14-27.

* Sugawara, S. K., Tanaka, S., Okazaki, S., Watanabe, K., & Sadato, N. (2012). Social rewards enhance offline improvements in motor skill. *PloS One, 7*(11), e48174. doi:10.1371/journal.pone.0048174; 10.1371/journal.pone.0048174

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*(1), 13-30. doi: 10.1002/jrsm.1091

Tipton, E. (in press). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods,* doi:2014-14616-001

* Tucker, M., McKinley, S., & Stickgold, R. (2011). Sleep optimizes motor skill in older adults. *Journal of the American Geriatrics Society, 59*(4), 603-609. doi:10.1111/j.1532-5415.2011.03324.x; 10.1111/j.1532-5415.2011.03324.x

* Tucker, M. A., & Fishbein, W. (2009). The impact of sleep duration and subject intelligence on declarative and motor memory performance: How much is enough? *Journal of Sleep Research, 18*(3), 304-312. doi:10.1111/j.1365-2869.2009.00740.x; 10.1111/j.1365-2869.2009.00740.x

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135*(3), 452-477. doi:10.1037/a0015329; 10.1037/a0015329

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics, 32*(1), 39-60. doi:10.3102/1076998606298034

Walker, M. P. (2005). A refined model of sleep and the time course of memory formation. *The Behavioral and Brain Sciences, 28*(1), 51-64; discussion 64-104.

* 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*(1), 205-211.

* Walker, M. P., Brakefield, T., Seidman, J., Morgan, A., Hobson, J. A., & Stickgold, R. (2003). Sleep and the time course of motor skill learning. *Learning & Memory, 10*(4), 275-284. doi:10.1101/lm.58503

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron, 44*(1), 121-133. doi:10.1016/j.neuron.2004.08.031

* Wilhelm, I., Diekelmann, S., & Born, J. (2008). Sleep in children improves memory performance on declarative but not procedural tasks. *Learning & Memory, 15*(5), 373-377. doi:10.1101/lm.803708

Wilhelm, I., Diekelmann, S., Molzow, I., Ayoub, A., Molle, M., & Born, J. (2011). Sleep selectively enhances memory expected to be of future relevance. *The Journal of Neuroscience, 31*(5), 1563-1569. doi:10.1523/JNEUROSCI.3575-10.2011

* Wilhelm, I., Metzkow-Meszaros, M., Knapp, S., & Born, J. (2012). Sleep-dependent consolidation of procedural motor memories in children and adults: The pre-sleep level of

performance matters. *Developmental Science, 15*(4), 506-515. doi:10.1111/j.1467-7687.2012.01146.x; 10.1111/j.1467-7687.2012.01146.x

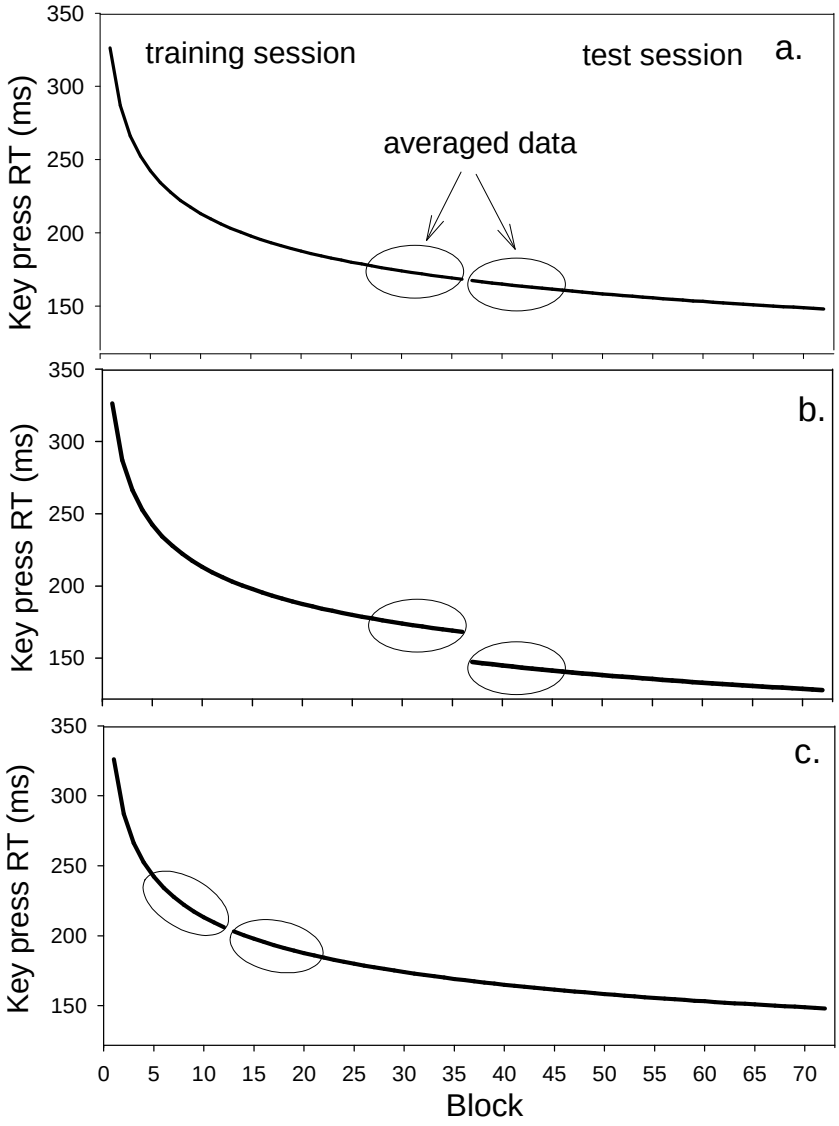

Figure 1. Hypothetical relations between practice block, session, and performance. Ovals represent the range of data used to find the pre- and post-test means. Panel a: performance curve if there is no effect of the delay between sessions. Panel b: performance curve if there is improvement due to the delay between sessions. Panel c: performance curve if there is no effect of the delay between sessions and a short training session.

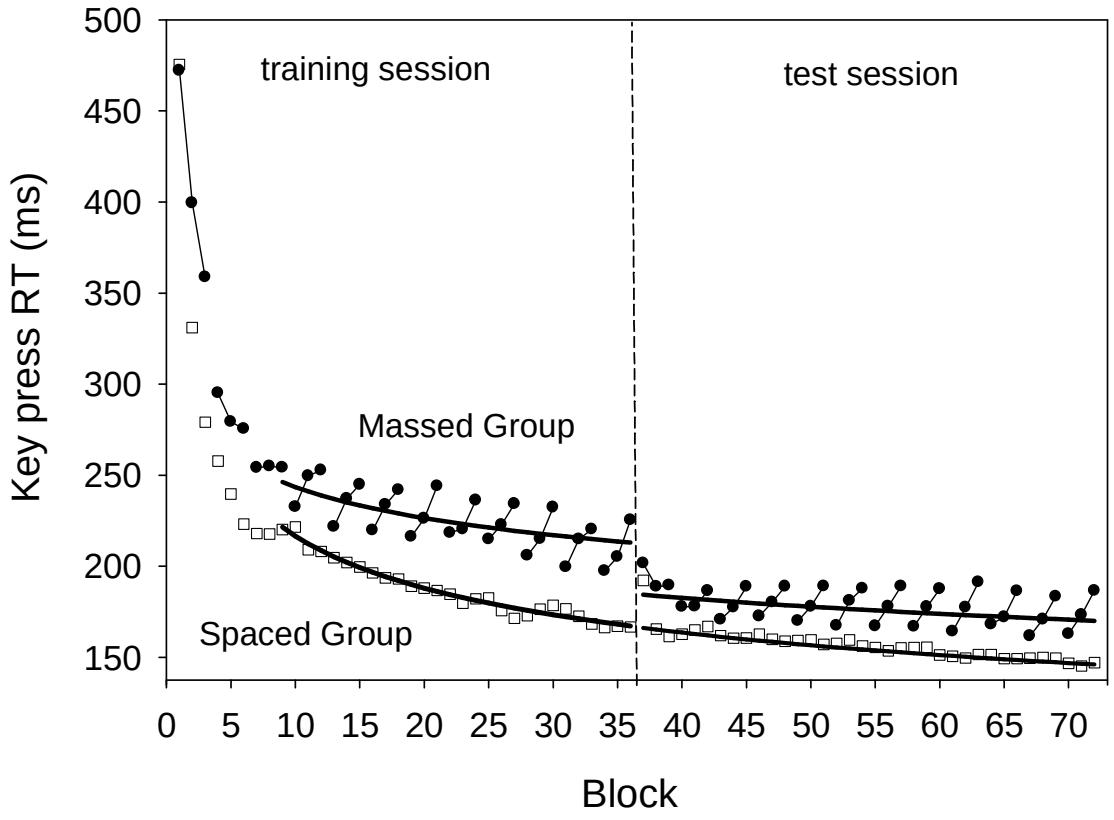

Figure 2. Performance as a function of group (massed vs. spaced), block, and session in Experiment 2 of Rickard et al. (2008). Each data point is the mean over a 10 s block. Each set of three linked circles for the massed group represent the three contiguous 10 s blocks within each 30 s performance period. Each performance period for the massed group was followed by a 30 s break. In the spaced group, there was a 30 s break between each 10 s block. Fitted lines for both massed and spaced groups are best fits of a three-parameter power function continuity test.

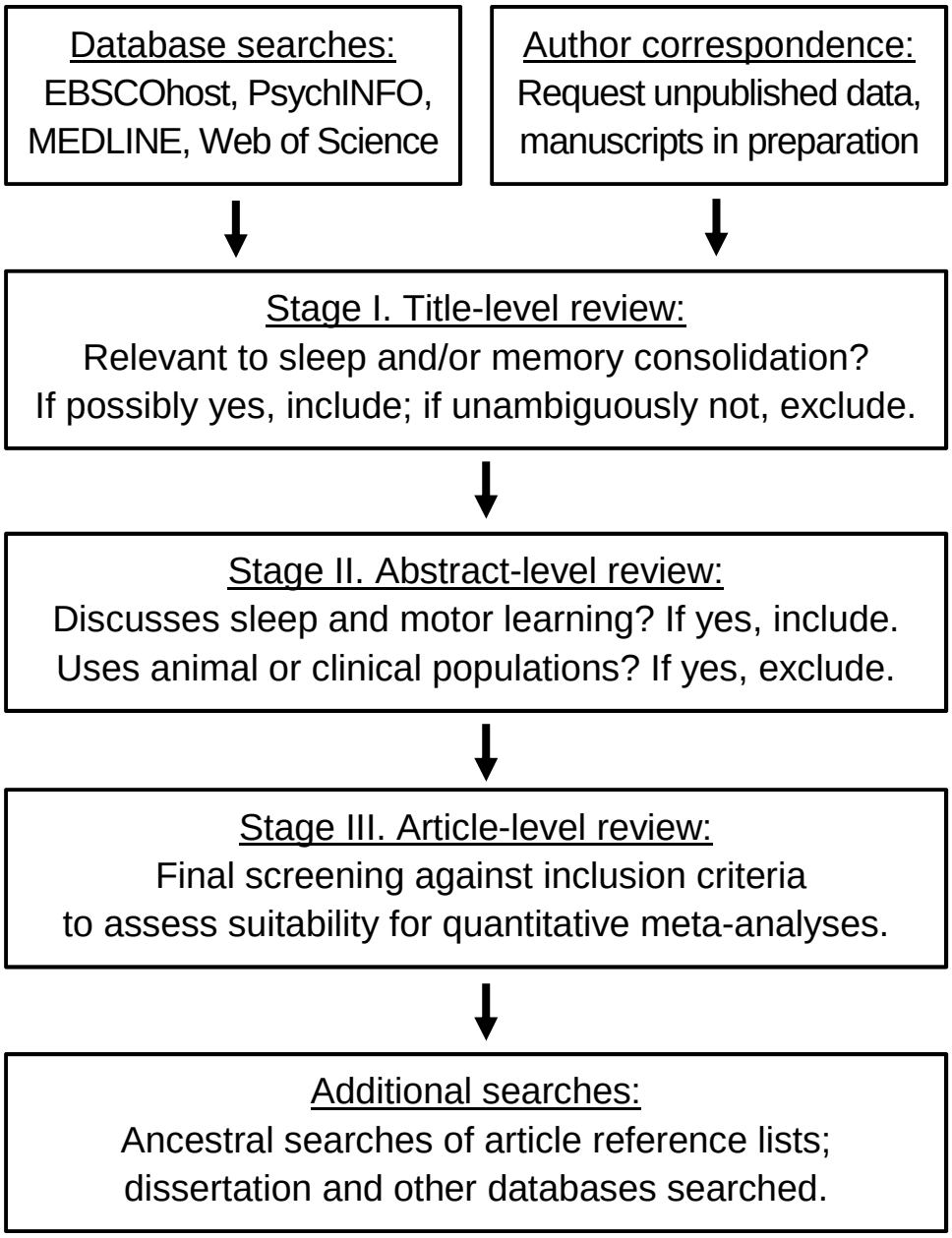

Figure 3. Flowchart of the literature search and selection process.

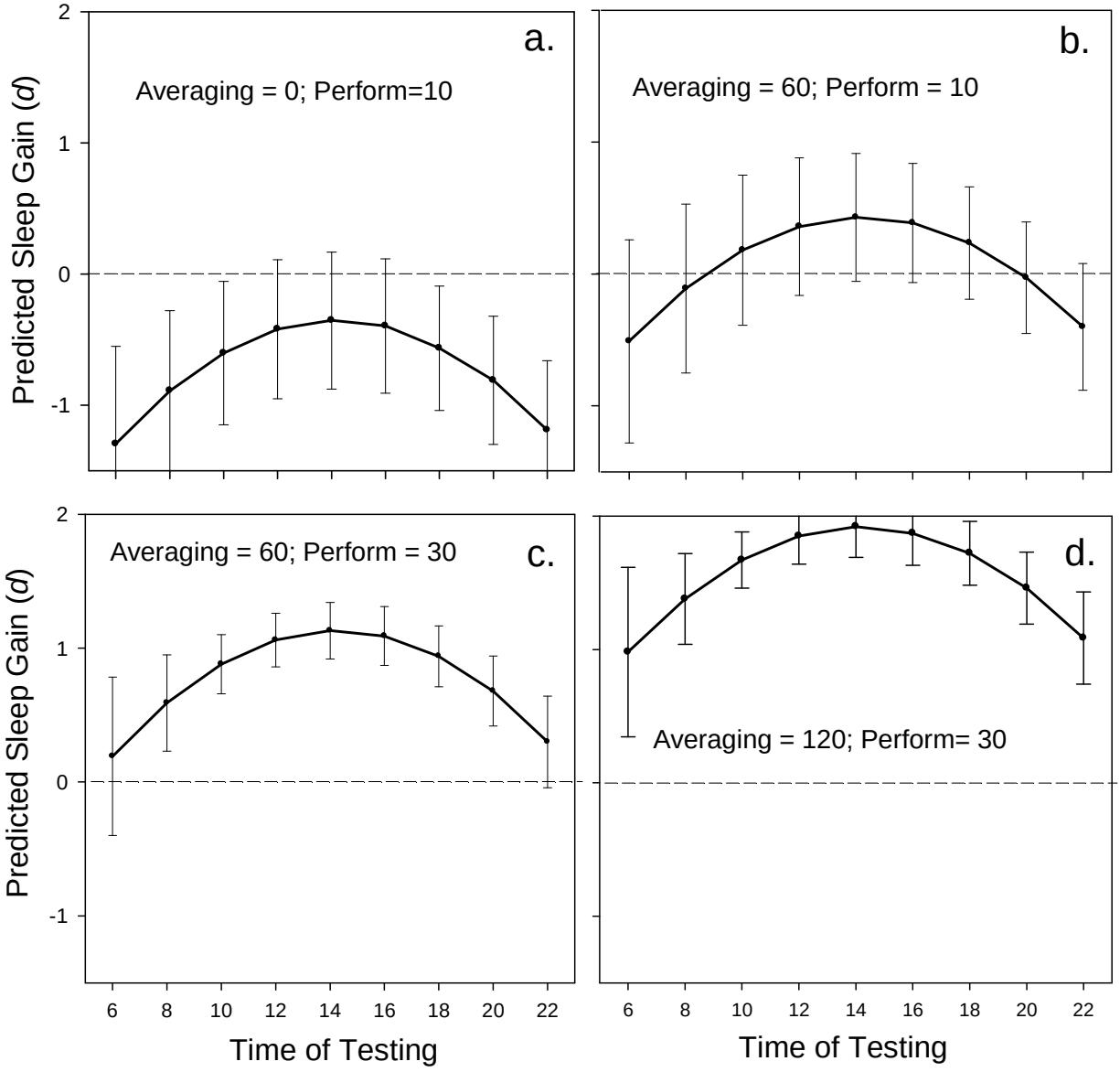

Figure 4. Predicted effect size for the post-sleep gain as a function of time of testing, data averaging, and performance duration per performance-break cycle (Perform). The predictions are based on the final working model fit to the data, and not on data values themselves.

Design	Reference	Sleep vs. wake group
Varied time	Walker et al., 2002	Group D vs. B
		Group E vs. C
	Korman et al., 2003	Exp. 3, overnight vs. over-day
	Blischke et al., 2008	Unguided, AMA vs. MAM
	Wilhelm et al., 2008	Children, sleep vs. wake*
		Adults, sleep vs. wake*
	Rickard et al., 2008	Exp. 1, sleep vs. awake
	Brawn et al., 2010	PM-spaced vs. AM-spaced
	Tucker et al., 2011	Elderly, sleep vs. wake*
Deprivation	Albouy et al., 2013	Sleep vs. sleep deprived
Nap	Nishida & Walker, 2007	Nap vs. no nap
	Korman et al., 2007	NapNoInt vs. NoNapNoInt
	Mednick et al., 2008	Nap vs. placebo
	Doyon et al., 2009	DaySleep vs. NoSleep
	Wilhelm et al., 2012	Children-low, sleep vs. wake*
		Children-med., sleep vs. wake*
		Adult-med., sleep vs. wake*
		Adult-high, sleep vs. wake*
	Fogel et al., 2014	Young, nap vs. no-nap
		Older, nap vs. no-nap
	Korman et al., in prep.	NapNoInt vs. NoNapNoInt
Varied delay	Cai & Rickard, 2009	1-night vs. wake
	Ashtamker & Karni, 2013	24 hr vs. 1 or 3 hr

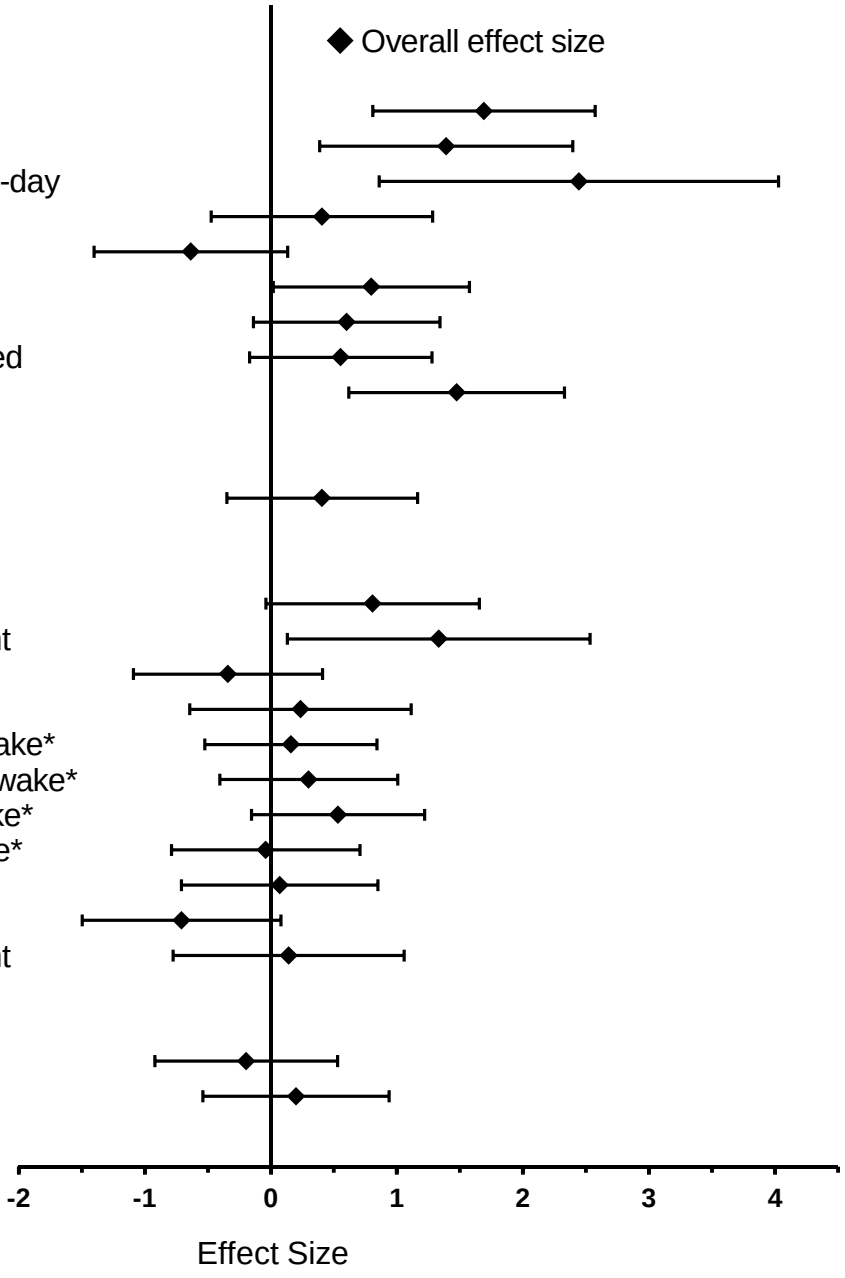

Figure 5. Forest plot of effect sizes for the 23 pairs of matched sleep-wake groups, with error bars displaying 95% confidence intervals. Asterisks represent correlated groups (i.e., same subjects in the sleep and wake groups).

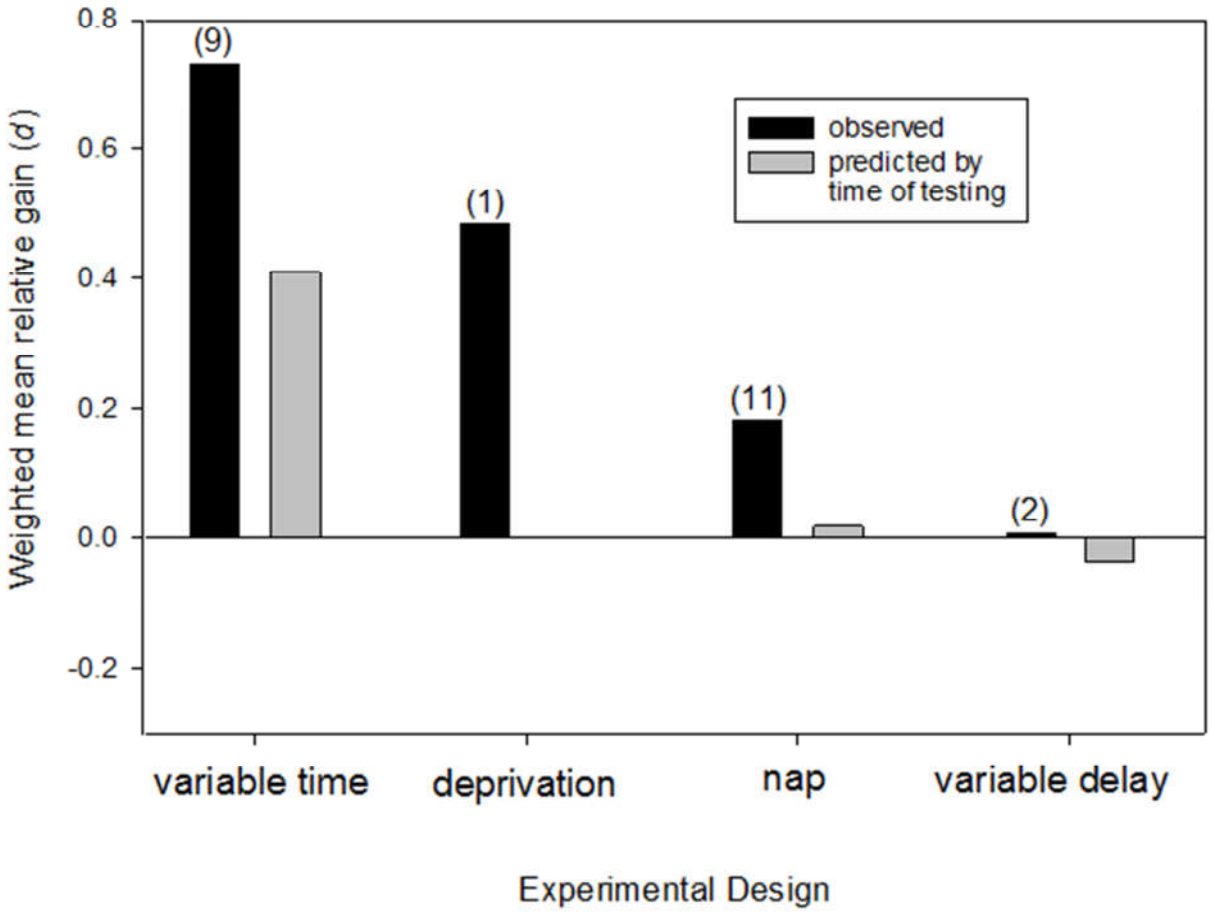

Figure 6. The mean relative gain effect size for each experimental design (black bars). Predicted effects due to time of testing are depicted by the gray bars. Numbers in parentheses above the bars are the number of experiments per design.

Table 1

List of Studies, Predictor Variables, and Effect Sizes

			Primary predictors							
Reference	Condition	*N*	Sleep status	Train time	Test time	Training duration	Perform.-break (s)	Data avg. (s)	*d*	*sv*
Walker et al., 2002	Group D	15	S	22	10	360	30-30	60	2.608	0.404
	Group B	15	W[a]	10	22	360	30-30	60	0.387	0.085
	Group E	12	S	22	10	360	30-30	60	1.501	0.252
	Group C	10	W[a]	10	22	360	30-30	60	0.474	0.149
Walker et al., 2003	Group 1	10	S	22	10	360	30-30	90	1.369	0.296
	Group 2	10	S	13	13	360	30-30	90	1.303	0.280
	Group 3	10	S	13	13	720	30-30	90	2.277	0.592
	Group 4	10	S	13	13	360	30-30	90	1.961	0.472
Korman et al., 2003	Exp. 1a	36	S	10	10	520	30-50	120	1.178	0.052
	Exp. 3, overnight	8	S	21.5	9.5	520	30-50	120	2.411	0.956
	Exp. 3, over-day	7	W[a]	12.5	21	520	30-50	120	0.431	0.247
Kuriyama et al., 2004	Group 1	15	S	13	13	360	30-30	90	1.609	0.202
	Group 2	15	S	13	13	360	30-30	90	1.864	0.244
	Group 3	13	S	13	13	360	30-30	90	1.445	0.215
	Group 4	14	S	13	13	360	30-30	90	1.189	0.159
Marshall et al., 2006	Sham	13	S	21.75	7.25	360	30-30	90	2.283	0.400
Dorfberger et al., 2007	Exp. 1, 9 yr. olds	21	S	10.5	10.5	600	30-20	120	1.729	0.144
	Exp. 1, 12 yr. olds	21	S	10.5	10.5	600	30-20	120	1.747	0.146
	Exp. 1, 17 yr. olds	20	S	10.5	10.5	600	30-20	120	1.504	0.130
Rasch et al., 2007	Exp. I, odor	18	S	22	7	360	30-30	90	0.657	0.079
	Exp. I, vehicle*	18	S	22	7	360	30-30	90	0.804	0.087
	Exp. II, odor	17	S	22	7	360	30-30	90	0.250	0.070
	Exp. II, vehicle*	17	S	22	7	360	30-30	90	0.502	0.077
	Exp. III, odor	17	S	22	7	360	30-30	90	0.796	0.093
	Exp. III, vehicle*	17	S	22	7	360	30-30	90	1.312	0.137
	Exp. IV, odor	18	S	22	7	360	30-30	90	0.340	0.072
	Exp. IV, vehicle*	18	S	22	7	360	30-30	90	0.530	0.073
Nishida and Walker, 2007	Nap	14	S	10	18	360	30-30	90	1.082	0.146
	No nap	12	W[c]	10	18	360	30-30	90	0.231	0.105
Korman et al., 2007	NoInt	9	S	9	9	520	30-50	120	1.670	0.448
	NapNoInt	8	S	12.5	21	520	30-50	120	1.517	0.484
	NoNapNoInt	8	W[c]	12.5	21	520	30-50	120	0.456	0.203
Blischke et al., 2008	AMA, unguided	11	S	20	8	360	30-30	60	0.829	0.166
	MAM, unguided	12	W[a]	8	20	360	30-30	60	0.505	0.119
Wilhelm et al., 2008	Children, sleep	15	S	20	8	360	30-30	90	0.553	0.092
	Children, wake*	15	W[a]	8	20	360	30-30	90	1.936	0.258
	Adults, sleep	15	S	22	8	360	30-30	90	1.211	0.148
	Adults, wake*	15	W[a]	8	22	360	30-30	90	0.563	0.093
Rickard et al., 2008	Exp. 1, sleep	16	S	22	10	360	30-30	30	0.245	0.084
	Exp. 1, awake	16	W[a]	10	22	360	30-30	30	-0.350	0.078
	Exp. 2, massed	54	S	13	13	360	30-30	30	0.814	0.026
	Exp. 2, spaced	55	S	13	13	360	10-30	30	-0.237	0.019
Morin et al., 2008	Motor sequence	13	S	21	9	360	30-30	90	1.783	0.280
Sheth et al., 2008	12 hr	44	S	23	11	360	30-30	30	0.535	0.029
	24 hr	11	S	11	11	360	30-30	30	-0.241	0.118

(table continues)

Table 1 *(continued)*

Reference	Condition	*N*	Primary predictors: Sleep status	Train time	Test time	Training duration	Perform.-break (s)	Data avg. (s)	*d*	*sv*
Mednick et al., 2008	Nap	13	S	9.5	16	360	30-30	60	1.448	0.216
	Placebo	18	W[c]	9.5	16	360	30-30	60	1.768	0.180
Dorfberger et al., 2009	Exp. 2, all groups	60	S	10	10	600	30-30	120	1.742	0.045
Genzel et al., 2009	Undisturbed	12	S	21	11	360	30-20	90	1.273	0.210
Rasch et al., 2009	Placebo	32	S	22.5	7	360	30-30	90	0.467	0.037
Doyon et al., 2009	DaySleep	10	S	9	21	520	30-30	120	0.727	0.176
	NoSleep	13	W[c]	9	21	520	30-30	120	0.438	0.104
	ImmDaySleep	9	S	12	20	520	30-30	120	1.523	0.398
	NightSleep	13	S	21	9	360	30-30	120	1.966	0.320
Tucker and Fishbein, 2009	Full night	13	S	23	7.5	354	29.4-30	88.5	1.201	0.177
Cash, 2009	No rest	12	S	20.75	8.75	360	30-30	90	0.981	0.166
Cai and Rickard, 2009	1-night	17	S	9.5	17.5	532	23.6-30	23.6	-0.375	0.082
	Wake	15	W[d]	9.5	17.5	511	23.6-30	23.6	-0.179	0.081
	2-night	11	S	9.5	17.5	575	23.6-30	23.9	0.028	0.113
Brawn et al., 2010	PM-spaced	14	S	21	9	440	10-30	60	0.270	0.088
	AM-spaced	20	W[a]	9	21	440	10-30	60	-0.532	0.065
Barakat et al., 2011	Motor sequence	12	S	21	9	360	30-30	90	1.051	0.175
Tucker et al., 2011	Elderly sleep	16	S	9	9	360	30-30	90	0.100	0.073
	Elderly wake*	16	W[a]	9	21	360	30-30	90	-1.288	0.145
	Young sleep	15	S	9	9	360	30-30	90	1.789	0.231
Wilhelm et al., 2012	Children-low	18	S	12.25	14.25	280	23.3-20	70.08	1.169	0.114
	Children-low*	18	W[c]	12.25	14.25	280	23.3-20	70.08	0.912	0.094
	Children-med.	17	S	12.25	14.25	192	18.9-20	57	0.749	0.090
	Children-med.*	17	W[c]	12.25	14.25	192	18.9-20	57	-0.058	0.067
	Adult-med.	18	S	12.25	14.25	28	14-20	42	1.358	0.132
	Adult-med.*	18	W[c]	12.25	14.25	27.2	13.6-20	42	0.957	0.097
	Adult-high	15	S	12.25	14.25	140	10.6-20	33.6	0.571	0.093
	Adult-high*	15	W[c]	12.25	14.25	140	10.6-20	33.6	0.818	0.110
Sugawara et al., 2012	No-praise	16	S	13.5	13.5	360	30-30	90	2.250	0.294
Mednick et al., 2013	Study 2, placebo	30	S	6	15	360	30-30	90	1.388	0.074
Albouy et al., 2013	Sleep	15	S	13.5	13.5	350	21-15	42	0.669	0.099
	Sleep deprived	15	W[b]	13.5	13.5	350	21-15	42	0.183	0.079
Feld et al., 2013	Placebo	12	S	21.75	19.5	360	30-30	90	1.167	0.192
Ashtamer and Karni, 2013	24 hr	10	S	11.5	11.5	624	30-30	120	0.992	0.189
	1 or 3 hr	30	W[d]	10.5	13	600	30-30	120	1.278	0.068
Adi-Japha et al., 2014	Children	20	S	10	10	600	30-30	120	2.057	0.194
	Adults	20	S	10	10	600	30-30	120	1.673	0.147
Fogel et al., 2014	Young, Nap	13	S	11	16	340.2	22.8-15	91.2	1.198	0.177
	Young, No-Nap	15	W[c]	11	16	340.2	22.8-15	91.2	0.589	0.094
	Older, Nap	14	S	11	16	525	34.8-15	139.2	-0.380	0.085
	Older, No-Nap	15	W[c]	11	16	525	34.8-15	139.2	0.404	0.093
Korman et al., in preparation	NapNoInt	11	S	12	20	520	30-50	120	-0.128	0.115
	NoNapNoInt	10	W[c]	12	20	520	30-50	120	-0.734	0.177

Notes: Entries in the condition column correspond to group labels in the respective papers. In the condition column, each wake group is listed immediately below its matched sleep group. Asterisks represent correlated groups (i.e., same subjects in two groups); each asterisked group is listed immediately below its corresponding correlated group. In the sleep status

column, superscripted letters for each matched pair of sleep-wake groups (placed on the sleep status indicator of the wake group) indicate the type of experimental design: a = varied time; b = deprivation; c = nap; d = varied delay. Sleep status: S = sleep group (full night or nap), W = wake group. Perform-break = duration of performance and break within each performance-break cycle.

Table 2

Individual Predictor Fits

Category	Predictor	k	*β*	*SE*	*df*	*p*	95% CI
Primary							
	Sleep status (wake = 1)	88	-0.62	0.17	15.8	0.002	-0.97, -0.26
	Averaging	88	0.0078	0.003	10.6	0.045	0.0002, 0.016
	Performance duration	88	0.032	0.020	3.3	0.20*	-0.028, 0.091
	Break duration	88	-0.002	0.014	5.4	0.91	-0.037, 0.034
	Training duration	88	0.0004	0.0008	4.2	0.64	-0.0018, 0.0026
	Train time	88	0.017	0.19	11.5	0.40	-0.025, 0.057
	Train time squared	88	-0.005	0.006	10.7	0.42	-0.0008, 0.012
	Test time	88	-0.056	0.024	12.0	0.04	-0.109, -0.002
	Test time squared	88	-0.002	0.0007	13.1	0.014	-0.0037, -0.0005
Secondary							
	Delay	88	0.0030	0.0071	4.62	0.69	-0.016, 0.022
	Child status (child = 1)	88	0.47	0.28	5.12	0.16	-0.26, 1.19
	Elderly status (elderly = 1)	88	-1.22	0.21	2.3	0.02*	-2.01, -0418
	Task type (thumb = 1)	88	0.42	0.30	8.1	0.19	-0.26, 1.10
	Nap status (nap = 1)	65	-0.68	0.17	5.9	0.710	-0.3223, 0.499
	Time slept	49	-0.071	0.17	6.9	0.69	-0.48, 0.34

Note: For dichotomous variables, the value of zero represents value that occurred most frequently and a value of one was used for the value that occurred less frequently. For example, for the sleep status variable, sleep groups (n = 65) were assigned a value of zero and wake groups (n = 23) were assigned a value of one. The level of the variable assigned a value of one is listed in parentheses beside the variable name. k = number of groups (effects sizes) available for each variable; β = regression coefficient; *SE* = standard error; *df* = adjusted degrees of freedom; *CI* = confidence interval. An asterisk indicates that the p-value is untrustworthy due to insufficient degrees of freedom (< 4). The time slept analysis excluded nap groups.

Table 3

Simultaneous Planned and Individually Significant Predictor Fits

Predictor	*β*	*SE*	*df*	*p*	95% CI
Sleep status	-0.259	0.094	7.6	0.026	-0.47, 0.041
Averaging	0.013	0.0022	6.3	0.001	0.008, 0.019
Training duration	-0.002	0.0006	6.6	0.026	-0.003, 0.0003
Performance duration	0.031	0.011	3.2	0.064*	-0.018, 0.026
Break duration	-0.0041	0.009	5.8	0.660	-0.018, 0.026
Train time	-0.052	0.103	5.6	0.636	-0.31, 0.21
Train time squared	0.001	0.003	6.2	0.715	-0.007, 0.009
Test time	0.381	0.114	11.2	0.006	0.13, 0.63
Test time squared	-0.014	0.003	11.1	0.002	-0.022, -0.006
Elderly status	-1.60	0.176	3.0	0.003*	-2.16, -1.05

Note: β = regression coefficient; *SE* = standard error; *df* = adjusted degrees of freedom; *CI* = confidence interval. An asterisk indicates that the p-value may be untrustworthy due to insufficient degrees of freedom (< 4).

Table 4

Final Working Model Fits

Predictor	β	*SE*	*df*	*p*	95% CI
Sleep status	-0.26	0.078	7.9	0.012	-0.44, -0.076
Averaging	0.013	0.002	5.8	<0.001	0.008, 0.019
Train duration	-0.001	0.0005	6.0	0.031	-0.0026, 0.0002
Performance duration	0.032	0.011	3.7	0.049*	0.0003, 0.063
Test time	0.39	0.098	9.1	0.003	0.175, 0.620
Test time squared	-0.014	0.003	11.3	<0.001	-0.021, -0.007
Elderly status	-1.61	0.195	3.1	0.009*	-4.86, -0.96

Note: β = regression coefficient; *SE* = standard error; *df* = adjusted degrees of freedom; *CI* = confidence interval. An asterisk indicates that the p-value may be untrustworthy due to insufficient degrees of freedom (< 4).

Table 5

Summary of Four Distinct Types of Sleep-Wake Comparison Designs

Design	Sleep deprivation in wake group	Type of sleep in sleep group	Controlled time of training/testing for sleep/wake groups	Controlled delay interval for both sleep/wake groups
Varied time	No	Full night	No	Yes
Deprivation	Yes	Full night	Yes	Yes
Nap	No	Nap	Yes	Yes
Varied delay	No	Full night	Yes	No

Appendix A

Reference	Condition	Task	Delay	Hours slept	Nap vs. full night	Child status	Elderly status
Walker et al., 2002	Group D	Finger-keyboard	12	7.6	Night	No	No
	Group B	Finger-keyboard	12	–	–	No	No
	Group E	Finger-keyboard	12	7.9	Night	No	No
	Group C	Finger-keyboard	12	–	–	No	No
Walker et al., 2003	Group 1	Finger-keyboard	12	7.8	Night	No	No
	Group 2	Finger-keyboard	24	7.8	Night	No	No
	Group 3	Finger-keyboard	24	7.8	Night	No	No
	Group 4	Finger-keyboard	72	7.8	Night	No	No
Korman et al., 2003	Exp. 1a	Finger-thumb	24	–	Night	No	No
	Exp. 3, overnight	Finger-thumb	12	–	Night	No	No
	Exp. 3, over-day	Finger-thumb	12	–	–	No	No
Kuriyama et al., 2004	Group 1	Finger-keyboard	24	7.6	Night	No	No
	Group 2	Finger-keyboard	24	7.6	Night	No	No
	Group 3	Finger-keyboard	24	7.6	Night	No	No
	Group 4	Finger-keyboard	24	7.6	Night	No	No
Marshall et al., 2006	Sham	Finger-keyboard	8	7.6	Night	No	No
Dorfberger et al., 2007	Exp. 1, 9 yr. olds	Finger-thumb	24	6	Night	Yes	No
	Exp. 1, 12 yr. olds	Finger-thumb	24	6	Night	Yes	No
	Exp. 1, 17 yr. olds	Finger-thumb	24	6	Night	Yes	No
Rasch et al., 2007	Exp. I, odor	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. I, vehicle*	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. II, odor	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. II, vehicle*	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. III, odor	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. III, vehicle*	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. IV, odor	Finger-keyboard	8.5	7.5	Night	No	No
	Exp. IV, vehicle*	Finger-keyboard	8.5	7.5	Night	No	No
Nishida and Walker, 2007	Nap	Finger-keyboard	8	1.1	Nap	No	No
	No nap	Finger-keyboard	8	–	–	No	No
Korman et al., 2007	NoInt	Finger-thumb	24	7	Night	No	No
	NapNoInt	Finger-thumb	8	1.3	Nap	No	No
	NoNapNoInt	Finger-thumb	8	–	–	No	No
Blischke et al., 2008	AMA, unguided	Finger-keyboard	12	8	Night	No	No
	MAM, unguided	Finger-keyboard	12	–	–	No	No
Wilhelm et al., 2008	Children, sleep	Finger-keyboard	12	9.5	Night	Yes	No
	Children, wake*	Finger-keyboard	12	–	–	Yes	No
	Adults, sleep	Finger-keyboard	12	6.9	Night	No	No
	Adults, wake*	Finger-keyboard	12	–	–	No	No
Rickard et al., 2008	Exp. 1, sleep	Finger-keyboard	12	6.3	Night	No	No
	Exp. 1, awake	Finger-keyboard	12	–	–	No	No
	Exp. 2, massed	Finger-keyboard	24	6.8	Night	No	No
	Exp. 2, spaced	Finger-keyboard	24	6.9	Night	No	No
Morin et al., 2008	Motor sequence	Finger-keyboard	12	7.5	Night	No	No
Sheth et al., 2008	12 hr	Finger-keyboard	12	7.2	Night	No	No
	24 hr	Finger-keyboard	24	7.6	Night	No	No

(appendix continues)

Appendix A *(continued)*

Reference	Condition	Task	Delay	Hours slept	Nap vs. full night	Child status	Elderly status
Mednick et al., 2008	Nap	Finger-keyboard	7	1.2	Nap	No	No
	Placebo	Finger-keyboard	7	–	–	No	No
Dorfberger et al., 2009	Exp. 2, all groups	Finger-thumb	24	6	Night	Yes	No
Genzel et al., 2009	Undisturbed	Finger-keyboard	62	6.7	Night	No	No
Rasch et al., 2009	Placebo	Finger-keyboard	32	7	Night	No	No
Doyon et al., 2009	DaySleep	Finger-keyboard	12	1.4	Nap	No	No
	NoSleep	Finger-keyboard	12	–	–	No	No
	ImmDaySleep	Finger-keyboard	8	1.5	Nap	No	No
	NightSleep	Finger-keyboard	12	7.5	Night	No	No
Tucker and Fishbein, 2009	Full night	Finger-keyboard	8.2	7.5	Night	No	No
Cash, 2009	No rest	Finger-keyboard	12	6.4	Night	No	No
Cai and Rickard, 2009	1-night	Finger-keyboard	32	7.5	Night	No	No
	Wake	Finger-keyboard	8	–	–	No	No
	2-night	Finger-keyboard	56	6.8	Night	No	No
Brawn et al., 2010	PM-spaced	Finger-keyboard	12	7.2	Night	No	No
	AM-spaced	Finger-keyboard	12	–	–	No	No
Barakat et al., 2011	Motor sequence	Finger-keyboard	12	7.5	Night	No	No
Tucker et al., 2011	Elderly sleep	Finger-keyboard	24	5.7	Night	No	Yes
	Elderly wake*	Finger-keyboard	12	–	–	No	Yes
	Young sleep	Finger-keyboard	24	6.1	Night	No	No
Wilhelm et al., 2012	Children-low	Button-box	2	1.1	Nap	Yes	No
	Children-low*	Button-box	2	–	–	Yes	No
	Children-med.	Button-box	2	1.1	Nap	Yes	No
	Children-med.*	Button-box	2	–	–	Yes	No
	Adult-med.	Button-box	2	1.2	Nap	No	No
	Adult-med.*	Button-box	2	–	–	No	No
	Adult-high	Button-box	2	1.1	Nap	No	No
	Adult-high*	Button-box	2	–	–	No	No
Sugawara et al., 2012	No-praise	Finger-keyboard	24	8	Night	No	No
Mednick et al., 2013	Study 2, placebo	Finger-keyboard	9	1.5	Nap	No	No
Albouy et al., 2013	Sleep	Finger-keyboard	72	–	Night	No	No
	Sleep deprived	Finger-keyboard	72	–	–	No	No
Feld et al., 2013	Placebo	Finger-keyboard	11	7.4	Night	No	No
Ashtamer and Karni, 2013	24 hr	Finger-thumb	24	8	Night	Yes	No
	1 or 3 hr	Finger-thumb	2	–	–	Yes	No
Adi-Japha et al., 2014	Children	Finger-thumb	24	6	Night	Yes	No
	Adults	Finger-thumb	24	6	Night	No	No
Fogel et al., 2014	Young, Nap	Finger-keyboard	5	1.2	Nap	No	No
	Young, No-Nap	Finger-keyboard	5	–	–	No	No
	Older, Nap	Finger-keyboard	5	0.8	Nap	No	Yes
	Older, No-Nap	Finger-keyboard	5	–	–	No	Yes
Korman et al., in preparation	NapNoInt	Finger-thumb	8	1.5	Nap	No	Yes
	NoNapNoInt	Finger-thumb	8	–	–	No	Yes

Notes: Entries in the condition column correspond to group labels in the respective papers. In the condition column, each wake group is listed immediately below its matched sleep group. In the hours slept column, a dash indicates a wake group or unreported sleep time.

Appendix B

Stata code used for the final working model:

robumeta d sleep status averaging training-duration performance-duration test-time test-time-squared elderly-status, study(paper) variance(sv) weighttype(hierarchical)