Time for Considering the Possibility That Sleep Plays No Unique Role in Motor Memory Consolidation: Reply to Adi-Japha and Karni (2016)

Timothy C. Rickard and Steven C. Pan
University of California, San Diego

The hypothesis that sleep makes a unique contribution to motor memory consolidation has been debated in recent years. In the target article (Pan & Rickard, 2015), we reported results of a comprehensive meta-analysis of the explicit motor sequence learning literature in which evidence was evaluated for both enhanced performance after sleep and stabilization after sleep. After accounting for confounding variables, we found no compelling evidence for either empirical phenomenon, and hence no compelling evidence for sleep-specific consolidation. In their comment, Adi-Japha and Karni (2016) critiqued the target article on three primary grounds: (a) our unrealistic (in their view) assumption that, if sleep-specific consolidation occurs, it is mechanistically unitary across all variants of the motor sequence experiments included in the meta-analysis, (b) our inclusion of child groups, which they believe may have resulted in an underestimation of sleep effects among adult groups, and (c) our inclusion of several experiments with atypical experimental designs, which may have introduced unaccounted for heterogeneity. In this reply we address each of those potentially legitimate concerns. We show that the metaregression allowed for tests of multiple candidate variables that could engender separate consolidation mechanisms, yielding no behavioral evidence for it. We also show through reanalysis that the inclusion of child groups had virtually no impact on the parameter estimates among adults, and that the inclusion of experiments with atypical designs did not materially influence parameter estimates.

Keywords: sleep consolidation, motor skills, motor sequence learning, sleep enhancement, procedural memory

For over 20 years, the explicit motor sequence learning task has been at the heart of a growing literature on the role of sleep in the consolidation of procedural memory. The most common version of that task involves repeated finger tapping of a deterministic multikey sequence, and the primary performance measure is either rate of sequence completion or tapping response time (reaction time). Numerous studies (e.g., Fischer, Hallschmid, Elsner, & Born, 2002; Korman, Raz, Flash, & Karni, 2003; Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002) have reported that sleep between training and test sessions results in performance enhancement, whereas an awake period does not. As a result, many researchers have come to regard as fact the claim that a sleep-specific consolidation mechanism enhances explicit motor skill learning.

However, Rickard, Cai, Rieth, Jones, and Ard (2008) demonstrated that at least one confound related to experimental design, the buildup of reactive inhibition (i.e., worsening of performance during each continuous training block, becoming more pronounced over the course of training but dissipating between sessions), and two data analysis artifacts (use of training minus test phase difference scores, and calculation of those scores over performance periods of up to 240 s), may drive all of the sleep enhancement effect. When those confounds were addressed, no enhancement was observed. An alternative consolidation hypothesis—that sleep uniquely stabilizes learning (without enhancing it) – has met with mixed results, a topic that we consider in more detail below.

Given our findings, we expected that researchers would reevaluate the sleep consolidation hypothesis for explicit motor learning, or at least modify experimental designs and analytic techniques to mitigate confounds. However, such improvements rarely came to fruition prior to publication of the 2015 target article (for partial exceptions see Adi-Japha, Badir, Dorfberger, & Karni, 2014; Brawn, Penn, Nusbaum, & Margoliash, 2010). In fact, since 2008, at least 50 studies have been published with little or no change in either experimental design or data analysis methods.

Key Findings of the Meta-Regression Analyses in Pan and Rickard (2015)

In the Pan and Rickard (2015) target article we sought to determine whether the confounding factors observed in our prior studies also drive the results in the broader literature, and whether a number of other candidate moderating variables may also influence results. We performed hierarchical random effects metare-
gression analysis on 88 standardized effect sizes (Cohen’s d) collected from a comprehensive review of the explicit motor learning literature. Performance was defined in terms of postdelay gain, which is the difference between the mean RT or performance rate at the end of the training session and the mean RT or performance rate on the test (for RTs, RT\text{train}−RT\text{test}). For sleep groups this was designated as the postsleep gain, and for awake groups the postawake gain; the relative sleep gain was the difference between sleep and awake groups. These terms referred only to empirical observations and not to theoretical interpretation.

In the final working model (see Table 4 of Pan & Rickard, 2015), seven variables were identified that, when entered in the metaregression analysis simultaneously, constituted statistically significant moderators of the postdelay gain: sleep-status (awake vs. sleep), amount of data averaging (in calculation of the gain score), training block duration (an index of reactive inhibition), performance duration, time of testing, time of testing squared, and elderly status. That model allowed us to test for postsleep gain, postawake gain, and relative sleep gain while statistically adjusting for potential effects of nonconsolidation-related moderating variables. In line with our 2008 work, there was no evidence for a positive postsleep gain (nor a postawake gain) when confounding moderator effects were accounted for. Two subsequent empirical papers (Backhaus et al., 2016; Nettersheim, Hallschmid, Born, & Diekelmann, 2015) have yielded similar conclusions. For analogous results in implicit motor learning, see also Keisler, Ashe, & Willingham, 2007; Nemeth, Janacek, Londe, Ullman, Howard, & Howard, 2010; Hallgató, Györi-Dani, Pekár, Janacek, & Nemeth, 2013.

The final working model did yield a modest relative sleep gain effect (sleep-status variable: d = 0.29). That result was confirmed in a supplemental analysis of relative sleep gain on the 23 studies with matched awake and sleep groups (d = 0.44). Theoretically, those results suggest that sleep may stabilize learning, even if it does not enhance it. However, of the four matched experimental designs used to study relative sleep gain (varied time, varied delay, deprivation, and nap; for details see the caption of Figure 1), only one (varied time) exhibited a consistent and statistically significant relative sleep gain (and, as discussed in the target article, it is uniquely subject to circadian confounds).

Response to Adi-Japha and Karni’s (2016)

General Comments

In their comment, Adi-Japha and Karni (2016) raised several concerns that they felt may substantially compromise our inferences from the target article. They first expressed their opinion that motor memory consolidation mechanisms are likely to differ— with respect to both neural processes and behavioral magnitude—as a function of multiple factors, including not only sleep status, but also developmental stage, task type (finger-keyboard vs. finger-thumb), amount of training, and other secondary properties of the task or procedures (such as bimanual vs. unimanual finger tapping). This view is exemplified in their statement, “We argue that different task demands, task conditions, and developmental differences should be considered a priori . . . ” (p. 568). They viewed those differences as being traceable to “complex biological processes generating memories after initial encoding” (p. 569), although they did not provide evidence that directly supports that claim. We will refer to that view as the consolidation complexity hypothesis for explicit motor sequence learning.

From that point of view, Adi-Japha and Karni (2016) critiqued our null assumption that a single consolidation process operates across all of the included studies, studies that varied with respect to each of the above noted properties, among others. It is correct that our analyses were conducted under the null hypothesis of broad homogeneity in consolidation processes. However, our focus in the target article was on testing multiple variables as possible moderators of the postsleep and relative sleep gain effects, and in determining whether such sleep effects remain after confounding influences of those variables are adjusted for. Hence, contrary to Adi-Japha and Karni’s claim, we performed exactly the type of analyses (within a metaanalytical framework) that are needed to test the consolidation complexity hypothesis over multiple candidate moderating factors.

Our analyses yielded no evidence for the consolidation complexity hypothesis. None of the variables that might plausibly yield qualitatively different consolidation processes approached statistical significance, including time slept, nap status, task type, child status, and delay interval. Adi-Japha and Karni (2016) highlighted task-type (finger-keyboard vs. finger-thumb) as one variable that might yield distinct consolidation processes. That variable did not approach statistical significance, however, either when considered as a singleton predictor (p = .19) or when added to the final working model (p = .16). Adi-Japha and Karni criticized us for “expecting that this factor would emerge from the analyses rather than addressing the data separately for each task version” (p. 570). It is not clear to us, however, how one could determine whether gain effects are statistically dependent on task version by analyzing the data separately for each task version. Furthermore, as highlighted in the target article, the effects of variables that survived in the final working model can be explained without appeal to differences in underlying consolidation processes (and indeed, without appeal to any consolidation process at all).

The lack of support for the consolidation complexity hypothesis in our meta-analysis does not, of course, conclusively disprove that hypothesis. Effects may have been missed in some cases due to limited statistical power. There may also be neural differences in consolidation processes that yield similar behavioral effects. Nevertheless, following the parsimony heuristic, the burden of proof now rests with those who wish to advance the complexity account.

Responses to Specific Critiques in Adi-Japha and Karni’s (2016) Comment

Adi-Japha and Karni (2016) next focused on several specific factors that may have biased our conclusions. Their concerns were reasonable and were not directly addressed in the target article. As it turns out, however, none of them are supported empirically. Their most prominent criticism was that our inclusion of both adults and children in the metaregression analyses may have compromised the validity of the results for adults. They first addressed our preliminary, single predictor analyses, involving no adjustment for moderating variables. That analysis showed that, for data combined across adult and child groups, there was both a positive postsleep gain and, of particular interest to them, a positive postawake gain. They argued based on the literature that adults do not exhibit a positive
postawake gain, whereas children do, a potentially important pattern in the data that we may have overlooked. In support of that argument they performed hierarchical random effects meta-analysis, which showed that the postawake gain was no longer statistically significant when the four child awake groups were removed (leaving the 19 adult groups for the analysis). We confirmed that finding when strictly interpreted within the significance cutoff of \( p < .05 \). The effect of excluding the four child groups was, however, negligible in magnitude (for the full set of 23 awake groups, \( d = 0.38 \) and \( p = .03 \); for the 19 adult awake groups \( d = 0.32, p = .09 \)). Furthermore, the small increase in \( p \) value that occurs when the four child groups are removed likely reflects, to at least some extent, the reduction in the sample size from 23 to 19. Keeping in mind that there are innumerable ways in which the data set might be carved up to yield results that differ somewhat from those reported for the full data set, there is little force behind Adi-Japha and Karni’s argument that children and adult groups exhibit different postawake gain patterns.

More importantly, our final working model showed that there is neither a positive postsleep nor postawake gain when results are adjusted for confounding variables. That model is the more informative framework within which to interpret results from the primary analyses, and we work within it—along with the supplemental analysis of matched group relative gain—in the discussion below.

Adi-Japha and Karni (2016) also made the closely related point that, in prior work (Adi-Japha, Badir, Dorfberger, & Karni, 2014; Ashtamker & Karni, 2013; Dorfberger, Adi-Japha, & Karni, 2007; Wilhelm, Diekelmann, & Born, 2008), no relative sleep gain effect has been observed for children. Thus, they argued, our inclusion of child groups may have partially masked a relative sleep gain effect for adults in both the primary analysis (i.e., the sleep status effect) and the supplementary matched groups analyses. We were able to address that possibility by performing follow-up analyses in which the 12 groups that met Adi-Japha and Karni’s preferred definition of child status (no older than early adolescence) were removed.

---

**Table 1.** Forest plot of effect sizes for the 23 pairs of matched sleep-awake groups, grouped by the four experimental design types used in the literature: *varied time* (training and test occurring at different times of day for the sleep and awake groups, with the delay interval held constant for both groups), *deprivation* (sleep deprivation in the awake group), *nap* (a nap rather than a full night of sleep for the sleep group), and *varied delay* (different delay intervals for sleep and awake groups, with time of day for training and test held constant for both groups). Crosses (†) in the sleep versus awake group column indicate child groups, and error bars indicate 95% confidence intervals. Figure adapted from Pan and Rickard (2015; p. 825).

**Figure 1.**
and the final working model refitted for the adult groups. Results (see Table 1 of this article) were nearly identical to those for the full set of data shown in Table 4 of the target article, with regression coefficients and p values changing only marginally. Most importantly for current purposes, the sleep status effects were nearly identical: d = 0.26 in the analyses of the full data set described in Pan and Rickard (2015) and d = 0.22 in the current reanalyses that excluded children.

Similar results were obtained when we removed the four pairs of experimentally matched child awake-sleep group pairs from the full set of 23 matched awake-sleep groups in the supplementary relative sleep gain analyses. In the original analysis of all 23 matched pairs, the aggregate relative gain effect size estimate was d̂ = 0.44 (p = .018). In the reanalysis that excluded children, that estimate was d̂ = 0.54 (p = .013). Thus, although the increase in aggregate effect size with the child groups removed is numerically consistent with Adi-Japha and Karni’s (2016) hypothesis, it is a small-to-negligible difference by usual effect size standards.

A deeper understanding of the matched relative sleep gain effects, and of the influence of the child groups, can be achieved through inspection of the forest plot of the relative sleep gain for the full set of 23 matched awake-sleep pairs (See Figure 1; reproduced from Figure 5 in the target article). The four child matched pairs are indicated by crosses in the column titled “Sleep vs. awake group.” In the nap design, the relative gain scores for the two child matched pairs (both from Wilhelm, Metzkow-Meszaros, Knapp, & Born, 2012) were close to zero but also near the median of the distribution for the adult matched pairs (e.g., Doyon et al., 2009). Similarly, in the varied delay design, the relative gain score for the single child matched pair (Ashtamker & Karni, 2013) was near zero, and was similar to that of the adult matched pair (Cai & Rickard, 2009). The only case in which the child data differed from the adult data was the varied time design, wherein the relative gain score for the child matched pair (Wilhelm et al., 2008) was negative, and much smaller than that for most of the adult pairs.

Thus, with the exception of a single child matched pair in the varied time design, there is no evidence that the relative gain effects for the child groups differ meaningfully from those for adult groups. Furthermore, removal of the child groups did not alter our conclusion in the target article that there is statistical evidence for a relative sleep gain only for the varied time design. For the other three designs combined (deprivation, nap, and varied delay), the aggregate relative gain estimate was d̂ = 0.17 for the full set of 14 adult and child sleep-awake pairs, and d̂ = 0.16 when the three child sleep-awake pairs were removed. In neither case was the relative gain effect statistically significant (p = .21 and .34, respectively). Thus, although Adi-Japha and Karni were correct in their assertion of no apparent relative sleep gain for children, there was also no compelling evidence for a relative sleep gain for adults in three of the four types of matched experimental designs.

Adi-Japha and Karni (2016) also questioned our inclusion of eight groups from the Wilhelm et al. (2012) paper. In their view, the task used in that study was implicit rather than explicit. Based on our own inclusion criteria, they argued, those groups should have been excluded. We disagree. The key differences in the Wilhelm et al. task that are relevant to the explicit versus implicit distinction were that (a) as opposed to the more commonly used explicit motor sequence task, subjects were not explicitly told that the sequence repeated deterministically, and that (b) an eight keypress rather than a five keypress deterministic sequence was used. It is reasonable to expect that subjects noticed after the first few repetitions that the sequence repeated, at which point the task would, by definition, become explicit. Furthermore, that task clearly does not constitute an implicit learning task by standards in the literature, in which implicit learning is indexed by performance differences between random sequences and sequences that have a more complex underlying pattern that is extremely difficult if not impossible for subjects to deduce (e.g., Nemeth et al., 2010; Robertson, Pascual-Leone, & Press, 2004). Nevertheless, to address this critique, we refitted the final working model from the primary analysis after excluding all eight child and adult groups from the Wilhelm et al. study. Results were again nearly identical to the results for the full data set reported in the target article. The sleep status estimate was d̂ = 0.24.

Finally, Adi-Japha and Karni (2016) critiqued our inclusion of one group from Feld et al. (2013) and the same eight groups from Wilhelm et al. (2012) on the basis that end-of-block performance summary feedback was provided, along with four groups from Kuriyama, Stickgold, and Walker (2004) on the basis that the motor sequence task was bimanual. Once again, however, results for the final working model were essentially unchanged when all of those groups were removed (for the sleep status variable: d̂ = 0.23).

Concluding Remarks

Adi-Japha and Karni (2016) make an important contribution in pointing out potential limitations in the target article analyses. Although it turned out that none of their concerns compromised our original conclusions, results could have been otherwise. We were also pleased that, in the last sentence of their comment, Adi-Japha and Karni agree with our conclusion that the field is in need of improved experimental design and data analytic techniques. As suggested by reviewers, we included in the target article a detailed list of experimental design and data analytic improvements that should help mitigate the risk of confounding influences in future work (Pan & Rickard, 2015, pp. 828–829). We fully agree with Adi-Japha and Karni that well-controlled (and in addition, large sample) experiments will be the ultimate adjudicators of hypotheses about the role that sleep may play in motor memory consolidation.