

**Severe Publication Bias Contributes to Illusory  
Sleep Consolidation in the Motor Sequence Learning Literature**

Timothy C. Rickard,<sup>1</sup> Steven C. Pan,<sup>2</sup> and Mohan W. Gupta<sup>1</sup>


<sup>1</sup>University of California, San Diego

<sup>2</sup>University of California, Los Angeles

The final version of this article was accepted for publication in the *Journal of Experimental Psychology: Learning, Memory, and Cognition* on August 21, 2021. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version will be available, upon publication, via the journal and its website.

The article is copyrighted by the American Psychological Association or one of its allied publishers. It is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Author Note**

Steven C. Pan  <https://orcid.org/0000-0001-9080-5651>

Mohan W. Gupta  <https://orcid.org/0000-0001-8632-592X>

## Abstract

We explored the possibility that publication bias in the explicit motor sequence learning literature significantly inflates estimates of the sleep-specific performance gains, potentially leading researchers to falsely conclude that there is sleep-specific neural consolidation of that skill. We applied PET-PEESE weighted regression analyses to the 88 effect sizes that were included in a recent, comprehensive literature review. Basic PET analysis indicated pronounced publication bias; i.e., the effect sizes were strongly predicted by their standard error. When predictor variables that have been shown to both moderate the sleep gain effect and reduce unaccounted for effect size heterogeneity were included in that analysis, evidence for publication bias remained strong; the estimated post-sleep gain was negative, raising the possibility of forgetting rather than facilitation, and it was statistically indistinguishable from the estimated post-wake gain. In a qualitative review of a smaller group of more recent studies we observed that (1) small sample sizes – a major factor behind the publication bias in the earlier literature – are still the norm, (2) use of demonstrably flawed experimental design and analysis remains prevalent, and (3) when authors conclude in favor of sleep-specific consolidation, they do not cite the papers in which those methodological flaws have been demonstrated. Recommendations are made for reducing publication bias in future work on this topic.

*Keywords:* publication bias, sleep, motor sequence learning, finger-tapping, finger-thumb

## Severe Publication Bias Contributes to Illusory

### Sleep Consolidation in the Motor Sequence Learning Literature

Over the last two decades, there has been substantial research interest in the role that sleep may play in consolidation of motor learning, and in particular explicit motor sequence learning. The typical experiment in that literature<sup>1</sup> involves either a keyboard finger-tapping task or an analogous finger-thumb opposition task wherein participants learn type or tap a repeating sequence. Participants are given multiple training blocks to acquire proficiency in the task (most commonly, 12 blocks), with the duration of each block varying across experiments (most typically, 30 s), and with a rest period between each block (typically also of 30 s). After training, there is a delay involving either sleep (in the form of a nap or a full night) or only wakefulness. On a subsequent test, there are two or more blocks of the same task, again interleaved with rest periods. Across that literature, the time of training and time of testing vary between morning and late evening, and the delay period between those sessions varies between a few hours (for studies involving naps) and 72 hours. The effect of the delay between sessions is measured by comparing performance averaged over the last few training blocks (the *pretest*) to performance averaged over the first few test blocks (the *posttest*).

Early papers using that experimental paradigm yielded a striking result: performance improved substantially between training and test sessions (by about 20%) if sleep occurred between those sessions (the *post-sleep gain*), but did not improve, or improved minimally, if only wakefulness occurred between sessions (constituting a *relative sleep gain* effect, wherein sleep between sessions yields more performance improvement than does wakefulness).<sup>1-3</sup> By 2008, at least a dozen papers had been published showing those two effects, jointly yielding more than 6,300 citations to date. A multitude of similar papers have been published since.

Almost ubiquitously in those papers, observed sleep gain effects have been interpreted to reflect a sleep-specific consolidation process.

In several more recent papers, however, evidence has been advanced that the empirical post-sleep gain effect does not reflect a consolidation process, but rather is an artifact of several experimental and data analytic confounds. An early example published in 2008<sup>4</sup> demonstrated that the post-sleep gain is eliminated when (1) the duration of training blocks between breaks is reduced from the typical 30 s to 10 s, minimizing the accumulation of block-level reactive inhibition, as well as more general task fatigue, factors can inflate the post-sleep gain estimate, (2) a 24-hr delay is used, thus holding time of training and testing constant and eliminating possible time-of-day confounds (e.g., effects of circadian rhythms on performance), and (3) the post-sleep gain effect is assessed using a learning curve continuity analysis rather than the nearly ubiquitous pretest-posttest difference score, which can inflate the post-sleep gain estimate due to averaging over online learning. Several other authors have reached similar conclusions using related approaches to minimize confounds<sup>5-7</sup>; and similar findings have been reported for the case of implicit motor sequence learning.<sup>8,9</sup> In a comprehensive meta-analysis published in 2015 that included 88 effect sizes drawn from 34 studies,<sup>10</sup> we demonstrated that earlier empirical results<sup>4</sup> are actually predicted by the literature when confounding experimental design and analysis factors are statistically adjusted for using meta-regression.

Although that work speaks against the hypothesis of a sleep consolidation process that enhances learning, the issue remains controversial; whereas some more recent studies acknowledge the foregoing evidence,<sup>6,7,11-13</sup> most authors still link the empirical post-sleep gain effect to sleep-dependent consolidation.<sup>14-20</sup>

In addition to the post-sleep gain, the relative sleep gain effect is often observed, and it is interpreted as evidence of more consolidation during sleep than during wakefulness. As pointed

out in the 2015 meta-analysis (<sup>10</sup>; see Figure 5 of that article), however, the relative gain effect has been consistently observed in only one of four experimental designs (namely, a *varied time* design involving a PM training – sleep – AM testing group and a PM training – wakefulness – AM testing group). In their regression model, that design was demonstrated to be vulnerable to a time of testing confound. Yet, many researchers continue to interpret the relative gain effect as unambiguous evidence for sleep-specific consolidation.<sup>12,15,16,21,22</sup>

### **Publication Bias, Sleep, and Motor Learning**

Defined broadly, publication bias occurs when studies that produce statistical evidence for a hypothesized phenomenon are more likely to be published than are those that do not. Several factors underlying publication bias have been discussed in the literature.<sup>23–25</sup> Chiefly among those factors for current purposes are (1) *publication criteria*, such as a cutoff *p*-value, that can block publication of non-significant results, and (2) *reporting bias*, wherein researchers decide not to pursue publication when the observed effect is null or in the opposite direction of that expected in the literature. In neither of those cases is the unpublished study necessarily flawed. Rather, the study results may reflect natural sampling variability. The expected consequence of publication bias in the literature is an inflated average effect size in the expected direction.

The likelihood of publication bias in a given literature is largely determined by the relation between true effect size and the standard error (*SE*) of the sample effect size. If there is a true effect, and if *SE* is small relative to that effect (if statistical power is high), then most studies will yield statistically significant results, facilitating publication. Relatively few studies will remain unpublished, and the mean effect size in the literature can approximate the true effect size. If, in contrast, *SEs* are large relative to the true effect size (i.e., power is low), then fewer studies will yield statistically significant results, and publication bias can be substantial.

An important characteristic of the sleep and motor sequence learning literature that is relevant to publication bias – but that has received no prior attention – is the generally small participant sample sizes. The distribution of those sample sizes, based on the 88 groups that were included in the 2015 meta-analysis, is shown in Figure 1. For most of that literature, the sample size is small (*median* = 15 per group), relative to both psychological research literature at large (*median* = 40) and current recommendations for best practice.<sup>26,27</sup> Because small samples yield relatively large *SEs*, there would appear to be a substantial risk of publication bias in the sleep and motor learning literature. We explore that possibility in the current paper. Specifically, we estimate the effect of publication bias using the Precision Effect Test and Precision Effect Test with Standard Errors (PET-PEESE) regression method,<sup>28,29</sup> both without and with inclusion of previously identified moderating variables that have been shown to account for the bulk of the effect size heterogeneity in that literature.<sup>10</sup>

## Method

### Meta-Analytic Procedures

In any experimental literature, there will be variability in *SE* over studies, because of variability in both sample size and random error. If there is publication bias, then studies with large *SEs* and small effect sizes, or with significant negative effects, are the most likely to be absent from the literature. If a positive effect is expected in a literature (as in the current case for both post-sleep and relative gain), then publication bias is expected to manifest as a positive slope in a plot of effect size (*Cohen's d* in the current case; y-axis) vs. a measure of statistical error, such as the standard error of the effect size (*SE*; x-axis). As preview, see Figure 2. PET-PEESE analysis captures that effect quantitatively using weighted least squares (WLS) regression of either *d* on *SE* (PET) or *d* on the sampling variability ( $SE^2$ ; PEESE). In both cases, the weighting variable is  $1/SE^2$ . The basic equation for PET is,

$$d = \beta_0 + \beta_1 SE_i + \varepsilon_i, \quad (1)$$

and for PEESE is,

$$d = \beta_0 + \beta_1 SE_i^2 + \varepsilon_i, \quad (2)$$

where  $\beta_1$  is the slope estimate (measuring the severity of publication bias, where a slope of zero indicates no bias),  $SE_i$  is the standard error for the  $i$ th effect size,  $SE_i^2$  is the corresponding sampling variability, and  $\varepsilon_i$  is the residual error. The parameter  $\beta_0$  is the estimated true effect size after adjusting for publication bias (i.e., for a hypothetical experiment with a very large sample size and hence negligible sampling variability).

PET-PEESE analysis involves three steps (in some papers, those three steps are referred to as FAT-PET-PEESE, although the first two steps involve inference based on the PET equation). First, the PET equation is fitted to the data and a significance test (at  $\alpha = .05$ ) is performed on the  $\beta_1$  estimate. If that test rejects the null, then publication bias is inferred. In that case, a significance test is performed on  $\beta_0$ . If that test does not reject the null hypothesis, then there no compelling evidence that the true effect size is difference from zero after correcting for publication bias, and the procedure is terminated. If that test does reject the null, then the true effect after adjusting for publication bias is estimated by the parameter  $\beta_0$  in a PEESE fit (Equation 2).

If candidate moderating variables are included in the regression (Stanley & Doucouliagos, 2014), then the respective equations are,

$$d = \beta_0 + \beta_1 SE_i + \sum_k \alpha_k z_k + \varepsilon_i, \quad (3)$$

and,

$$d = \beta_0 + \beta_1 SE_i^2 + \sum_k \alpha_k z_k + \varepsilon_i, \quad (4)$$

where  $z_k$  is the  $k^{\text{th}}$  moderator variable and  $\alpha_k$  is the corresponding effect size estimate. The three

PET-PEESE steps are identical to those for the case of no moderating variables.

Recent work<sup>29,30</sup> has shown that the PET-PEESE method generally yields valid results when there is effect size *homogeneity*, such that all sample effect sizes are random deviates from a *single* true effect size (i.e., the *fixed effect* case). However, in the more likely case of effect size *heterogeneity*, wherein subsets of the sample effect sizes are random deviations from *different* true effect sizes (i.e., the *random effects* case), results using Equations 1 and 2 can be biased. In the 2015 meta-analysis,<sup>10</sup> substantial heterogeneity, both between papers and over effect sizes within paper, was observed in a hierarchical random effects (HRE) analyses of the same dataset that is analyzed here for possible publication bias. Therefore, application of Equations 1 or 2, although a useful starting point, may not yield trustworthy results regarding publication bias.

However, the 2015 meta-analysis identified seven variables that can moderate the post-delay gain effect (see Table 1). When all of those variables were simultaneously included in a HRE “final working model,” the between-paper heterogeneity was substantially reduced (from  $\tau^2 = .27$  to  $.08$ ), and the within-paper heterogeneity ( $\omega^2$ ) was eliminated (falling from  $.13$  to zero). Hence, by applying Equation 3 and 4 with those moderating variables included, the potential for an unaccounted-for heterogeneity influence on the PET-PEESE publication bias estimates is greatly reduced.

Based on simulation results in a recent methodological article,<sup>30</sup> two other aspects of the current data context favor valid inference using PET-PEESE. First, the identified moderating variables reduced the within-paper heterogeneity estimate ( $\omega^2$ ) to zero in the 2015 meta-analysis. In that case, potential bias in the PET-PEESE estimates is reduced. Second, that bias is further reduced when publication bias, if present, involves suppression of non-significant results rather than suppression of statistically significant, but negative (relative to expectation) results. Given the multiple confounding factors that have been shown to cumulatively inflate sleep-gain



estimates (data averaging, duration of training, duration of each training block, time of testing), in this literature the latter type of publication bias is much less likely than is the former.

## **Dataset**

Data used in the primary analyses correspond exactly to the data analyzed in the 2015 meta-analysis (see their Table 1). That dataset, which is publicly archived on the Open Science Framework (<https://osf.io/u2c8s>), encompasses 88 experimental groups (88 effect sizes) from 34 studies. Sixty-six of those were sleep groups (where sleep intervened between training and test sessions) and the remaining 23 were wake groups. Those data are comprehensive of that literature at the time, after applying well-justified exclusion criteria. Use of that dataset for the current analyses facilitates direct and instructive comparison between the results of the *HRE* analyses in the 2015 meta-analysis and the current fixed effects WLS regression analyses (including the *PET* and *PEESE* analyses). In the final Results section of this paper, we review the smaller, more recent literature.

## **Adjustment of Moderator Variable Intercepts**

In the current analyses, the overall model intercept ( $\beta_0$ ) estimates the post-sleep gain effect when the confounding influence of any included moderating variables is adjusted for to minimize their confounding influence. To achieve that goal, the numerical values of some of those moderating variables were shifted (by adding or subtracting a constant value) from their experimentally recorded values, such that the value of zero for each variable corresponded to the case of minimal confounding influence or, for dichotomous predictors, to the more appropriate or important category level. Note that these shifts play no role in either the moderator variable parameter estimates or their sampling error. Rather, they affect only the overall model intercept estimate,  $\beta_0$ . For further details about the moderating variables, see the 2015 article<sup>10</sup>).

For one of the continuous variables, *data averaging*, the natural zero point (no averaging)

minimizes the confounding influence, so no shift was needed. For the remaining continuous predictors, a shift by subtraction of a constant was necessary. For *performance duration* per block, 10 s is the minimal value in the literature, and shifting the intercept by that amount should minimize the confounding effect of reactive inhibition and general training at the regression intercept.<sup>4</sup> To create the new intercept for that variable, 10 s was subtracted from the *performance duration* value for each of the 88 experiments. A *training duration* of 360 s is by far the most common in the literature, and by the end of training in that case, task performance improvements per block are at their lowest level. As such, any performance gain between the end of training and the beginning of the test that is due to continuing *online learning* should be negligible, virtually eliminating factor as a confound. Training duration for each group was thus adjusted by subtracting 360 s.

The effect of *time of testing* is more a property of performance than it is a methodological confound, and as such predicting the post-sleep gain when time of testing yielding to minimal predicted gain is not appropriate. Given our working hypothesis that there is no consolidation-based sleep-gain effect, we instead we did the opposite. We set the intercept to represent the gain effect when time of testing produces its largest positive effect on sleep gain (2 pm; 14:00 hours). To test for the sleep gain intercept at that most favorable test time, we subtracted 14 hours from time of testing for each experiment, such that a test at 2 pm corresponded to the intercept for that moderator. Given that shift, if the post-sleep gain estimate ( $\beta_0$ ) in the following analyses is non-significant or is negative, then we can infer that the gain estimate is also non-significant or negative at any other time of testing.

Two dichotomous variables were also included as predictors. First, the *sleep status* predictor was assigned a value of 0 for sleep groups and 1 for wake groups. Hence, the regression intercept,  $\beta_0$ , is the estimated *post-sleep gain* effect size. Thus, the *sleep status*

estimate is the predicted *relative gain* effect size (taking a negative value when the post-sleep gain is larger than the post-wake gain). Second, the *elderly status* predictor was set to zero for groups with non-elderly participants and to one for groups with elderly participants. Elderly groups often exhibit little evidence for an immediate post-sleep gain (c.f., <sup>31,32</sup>). Hence, setting the elderly status variable to zero for non-elderly groups maximizes the possibility of observing a positive gain at the model intercept. In summary, the intercept for the later described analysis that includes all moderator variables corresponds to the estimated post-sleep gain effect for the following case: sleep groups, non-elderly participants, zero data averaging, block duration of 10 s, training duration of 360 s, and 2 pm time of testing. Model predictions for any other combination of the variable values can be calculated using the reported regression coefficients.

### **Publication Bias Results**

Prior to investigating publication bias, we compared WLS regression results without the publication bias estimator (SE) to the 2015 meta-analysis' HRE results for the same dataset, with all of the previously identified moderating variables included in both cases (Table 1). The parameter estimates were virtually identical in the WLS and HRE analyses. That equivalence confirms that the moderating variables account for effect size heterogeneity to the same extent in the current WLS analysis as in the prior HRE analyses.

We next explored publication bias using PET-PEESE for three cases (see Table 2). In case 1, no moderating variables were included in the analysis (Equation 1). The PET results, depicted in Figure 2, suggest severe publication bias. There is (1) an absence of effect sizes in the lower right area of the figure (which is expected if there is publication bias), and (2) a robust effect of *SE* ( $p < .0001$ ), as indicated by the slope of the prediction line. The intercept estimate,  $\beta_0$ , was non-significantly negative. We next explored the case of only *sleep-status* as a moderator variable (Case 2 of Table 2), using Equation 3. The effect of *SE* was again substantial, and a

*sleep-status* (i.e., relative sleep gain) effect was observed:  $d = -.47$  ( $p = .001$ ).

Finally, we performed the PET-PEESE analysis with all moderating variables included, substantially reducing the unaccounted-for effect size heterogeneity. Results of the PET analysis are summarized in Case 3 of Table 2. Both *SE* and most of the previously identified moderator effects remained robust and statistically significant, the exceptions being *sleep-status* and the linear component of *time of testing*. Although the  $B_1$  estimate was reduced compared to Cases 1 and 2 (a result that is not surprising given the modest correlation among the moderating variables), it remained robust and highly significant ( $p < .0001$ ). Indeed, from the smallest to the largest *SE* value in the dataset (about .15 to .98), the  $B_1$  estimate of 1.95 predicts an increase in  $d$  of about 1.6; a very large effect. Hence, there is strong evidence for publication bias in the dataset. The estimated *sleep-status* effect is small and non-significant in this analysis ( $d = .17$ ,  $p = .16$ ), and the PEESE analyses yielded a  $\beta_0$  estimate of  $d = -.43$ ,  $p = .0007$ . Thus, when both the full set of previously identified moderator variables and *SE* are fitted simultaneously, the results suggest a negative post-sleep gain and a negligible relative gain; these results suggest that there may be *forgetting* of motor sequence skill during the delay between sessions, along with a minimal difference in the extent of that forgetting for wake and sleep groups.

### **Review of the More Recent Literature**

We found 12 papers that (1) were published after the cutoff date for inclusion in the 2015 meta-analysis, and (2) met the inclusion criteria used by those authors (see Table 3). Formal meta-analysis with moderator variables was not performed for those papers, because (a) for several of the papers it was not possible to extract the necessary statistics for individual experimental groups that are needed for that analysis, and (b) there was minimum variability in the values of the major predictor variables that we have previously identified, which, combined with the small set of papers, negated any possibility of accurately estimating their moderating

effect on performance, and (c) a formal meta-analysis to simply estimate the aggregate effect size for the post-sleep and relative gains would include only a subset of the papers, and we deemed that approach less useful than the alternative approach described below, which was applicable to all reported effects from all 12 papers.

For each paper, the authors' conclusions regarding both post-sleep gain and relative sleep gain (where applicable) were coded on a four-level ordinal scale which captured both the direction of the trend and statistical significance. Those levels are listed in Table 3 as "positive (s)," "positive (ns)," "negative (ns)," and "negative (s)." For the post-sleep gain, "positive" refers to better performance at the beginning of the test than at the end of training. For the relative sleep gain, "positive" refers to the finding of a pre-post difference scores that is larger for the sleep group than for the wake group. In the table, "s" refers to a statistically significant result at  $\alpha = .05$ , and "ns" refers to a non-significant result. In a few cases, no trend was specified for a null result, in which case the column entry is simply "ns."

Results for post-sleep gain were generally in-line with the earlier literature. A positive gain was observed in 18 of the 20 statistical tests (14 of which were statistically significant). For the relative gain, eight of 15 results were positive (six of them significant). With the exception of one significant negative case, remaining cases were either negative (ns) or (ns). In summary, for the case of a full night of sleep, the post-sleep and relative gain results for the newer studies are similar to those of the older studies. For the case of the nap experiments, most results in Table 3 were not significant, with one exception.<sup>33</sup> A similar pattern of mostly null results for nap studies is evident in the forest plot of a set of 11 experiments from the earlier literature (as in<sup>10</sup>; Table 5). Of note, all nap studies control for time of training and time of testing for the wake and nap groups.

Critical to interpretation of those results is whether the experimental design and data

analysis have been improved in the new studies. Unfortunately, in nearly all cases they were not. First, group sample sizes remain small (median = 13.6 subjects), perpetuating the conditions for likely publication bias as demonstrated in the current paper. Second, the previously identified experimental design confounds remain. In fact, in all of the studies listed in Table 3, the experimental design and data analysis were identical to, or nearly identical to, those that dominated in the earlier studies. Third, data analysis has not changed in the majority of papers. As in the earlier studies, it involves averaging of results across the last few training blocks to compute a pretest result, and over the first few test blocks to compute a posttest result, constituting a potential online learning confound that has been confirmed in our meta-analyses. In summary, all of the previously identified confounding factors are present. Thus, the results for newer studies do not contradict our conclusions based on the earlier studies.

Note that one paper<sup>13</sup> in Table 3 compared performance across wake and sleep groups in a design involving 12 training and 12 test blocks, rather than only a few test blocks as in most studies. They did not observe a statistically significant relative gain effect in usual the comparison of the last few training blocks to the first few test blocks (noted in Table 3), but they did observe a relative sleep gain in the comparison of the last few training blocks to the *last* few test blocks, raising the intriguing possibility that sleep consolidation effects may not manifest immediately, but rather only after sufficient practice. However, over several other papers in which a large number of test blocks was administered,<sup>4-6,34</sup> there have been no reported patterns of increasing relative sleep gain over blocks.

Finally, we note a pattern of scholarly omission among the more recent papers in Table 3. Our first paper raising the specter of serious confounds in this literature was published in 2008, and two other papers making related points were published prior to 2011.<sup>5,35</sup> In contrast, all of the papers listed in Table 3 were published during or after 2014. Despite that time differential,

none of the prior work demonstrating confounds in the sleep and motor learning literature was discussed among the nine papers listed in Table 3 in which the authors concluded in favor of sleep consolidation. That pattern of absent citation is remarkable considering the virtual absence of direct challenges in the literature to the evidence for serious confounds, the sole exception to our knowledge being a critique of the 2015 meta-analysis<sup>21</sup> to which we responded effectively.<sup>36</sup> In contrast, among each of the three papers in Table 3 in which the authors did not observe the usual post-sleep or relative gain effects – and in which the prior work regarding confounds buttressed their conclusions – that prior work was cited and discussed.<sup>13,32,37</sup> It seems unlikely that an exact match between the authors’ conclusions and citation of that prior work would occur by chance.

### **Discussion**

In prior work, confounding design and analysis variables in the sleep and motor learning literature were identified.<sup>4,5,10,34–36</sup> In the current work, we have identified extensive publication bias as an additional factor that complicates interpretation. When we simultaneously accounted for both types of confounding factors in the current analyses, the evidence in the literature suggests no, or possibly negative, post-sleep gain, and negligible relative gain. Hence there is little behavioral evidence for sleep-specific consolidation.

In our prior meta-analysis, the relative gain effect remained statistically significant when the full set of moderators was included. When all of those moderators were included along with SE to account for publication bias in the current paper, the sleep-status effect size estimate decreased by about half and became non-significant. That result suggests that one source of publication bias is an underrepresentation in the literature of non-significant relative gain effects (i.e., a null difference for wake vs. sleep groups), most likely for small *n* studies. As a consequence, the average relative gain effect in the published literature would be inflated. That

mechanism for publication bias would also be expected to yield both inflated post-sleep gains effects in the literature and deflated post-wake gain effects. Publication bias is likely also present among studies involving only sleep groups.

### **The Pernicious Effect of Publication Bias and Recommendations**

Between the two issues summarized above, that of publication bias may be the more problematic. The previously identified design and analysis confounds do not call into question the replicability of published results. If those factors constituted the only sources of bias, then new studies involving improved design and analysis would allow for convergence on the true effects. The misleading influence of publication bias, on the other hand, may be difficult to eliminate unless the great majority of future, well-designed studies are published regardless of results, a goal that would require broad cooperation among both researchers and publication outlets.

Multiple articles provide general recommendations for the reduction of publication bias.<sup>25-27</sup> At a minimum, sample sizes of 40 or more, the higher end of the current literature, are needed, and a priori statistical power analysis should be adopted as a standard in this literature. For the motor sequence and other frequently used motor tasks, it should not be difficult to access existing datasets to obtain variance estimates that will support power analysis. Further, the apparent absence of sleep effects among studies that are relatively well-controlled for confounding factors motivates power analysis for a two-tailed test.

If the experiments are purely behavioral and involve an accessible population, samples of sufficient size to obtain high statistical power should be readily obtainable. It may not be feasible to collect large samples if neurophysiological measurements such as EEG or neuroimaging are involved. However, as a compromise, we recommend collection of a large sample for a behaviorally identical experiment that does not involve neurophysiological measures, with the



goal of publishing both experiments in the same paper. If possible, random assignment to neurophysiological and behavioral groups is preferred. Finally, authors should publish both null and contradictory results.<sup>38,39</sup> If such findings cannot be published in a primary journal, they should be published in one of the multiple outlets that expressly embraces publication of null results.

### **Limitations of the Current Work**

As with any application of meta-analysis with predictor variables, there may be patterns in the data that were not accounted for. In particular, although inclusion of the full set of moderating variables accounted for the majority of the heterogeneity among the 88 effects sizes that were analyzed both here and in the 2015 meta-analysis, a small portion of the between-paper heterogeneity remains unexplained. However, our conclusion that substantial publication bias exists in this literature is unlikely to be compromised by that fact. The publication bias effect ( $\beta_1$ ) was robust both when the PET analysis included no moderating variables (and hence none of the effect size heterogeneity was accounted for) and when it included the full set of moderators (and hence most of the effect size heterogeneity was accounted for). In both cases, the p-value for  $\beta_1$  was less than .0001. The fact that the  $\beta_1$  estimate remained robust when most of the effect size heterogeneity was accounted for suggests that it would also remain robust in the hypothetical case in which the smaller, remaining heterogeneity is accounted for.

With respect to the PEESE estimate for  $\beta_0$ , we should be more cautious. Although that estimate suggests forgetting between sessions for both wake and sleep groups, simulation results<sup>30</sup> show that residual, unaccounted for between-paper heterogeneity can inflate the magnitude of  $\beta_0$ . Stronger inference regarding the true post-sleep gain effect comes from a combination of the meta-analytical results and the results of experiments in the literature in which efforts have been made to reduce confounding influences.<sup>4-7,34</sup> Those experiments have

yielded non-significant post-sleep gain estimates that exhibit no positive trend. That combined evidence strongly suggests that there is no performance enhancing, sleep-specific consolidation for the case of motor-sequence learning.

We have focused here on behavioral results for humans. Our results do not speak directly to the possibility of sleep-based motor skill consolidation in other animals. The current results also do not address electrophysiological or neuroimaging results for humans. However, the neurophysiological evidence for sleep consolidation in the motor sequence task is mixed. In the 2015 meta-analysis, the EEG results of eight articles that were included in behavioral meta-analysis were reviewed. It was observed that null results predominated, despite the broad use of uncorrected multiple comparisons. In contrast, distinct activation patterns or brain areas have been correlated with empirical sleep gain effects in some more recent neurophysiological studies,<sup>40,41</sup> and those patterns have been interpreted as evidence for offline sleep consolidation processes. However, those studies have used the same problematic experimental designs as have the majority of behavioral studies. Various confounding factors that can differentially affect performance during training vs. a test phase, or for wake vs. sleep groups, can potentially account for the neurophysiological effects. Further, neurophysiological studies nearly always involve small sample sizes, raising the prospect of publication bias in that sub-literature. Finally, because there are apparently no behavioral post-sleep or relative gain effects when confounding factors that are unrelated to consolidation processes are corrected for, any argument that neurophysiological findings reflect consolidation processes is unpersuasive.

### **Conclusions**

We have presented evidence of severe publication bias in the sleep and motor sequence learning literature. We also showed that, in PET-PEESE analyses that estimated the combined effects publication bias and previously identified confounding factors, there was no compelling

evidence for sleep-specific consolidation of motor sequence learning, in the form of either a post-sleep gain or a relative gain. Going forward, improved experimental methods, along with efforts to increase sample size and report null or contradictory results, will be needed to reach consensus regarding of the role of sleep in human motor sequence learning.

## References

1. Walker MP, Brakefield T, Morgan A, Hobson JA, Stickgold R. Practice with Sleep Makes Perfect: Sleep-Dependent Motor Skill Learning. *Neuron*. 2002;35(1):205-211.  
doi:10.1016/S0896-6273(02)00746-8
2. Walker MP, Brakefield T, Seidman J, Morgan A, Hobson JA, Stickgold R. Sleep and the Time Course of Motor Skill Learning. *Learn Mem*. 2003;10(4):275-284.  
doi:10.1101/lm.58503
3. Korman M, Raz N, Flash T, Karni A. Multiple shifts in the representation of a motor sequence during the acquisition of skilled performance. *Proc Natl Acad Sci*. 2003;100(21):12492-12497. doi:10.1073/pnas.2035019100
4. Rickard TC, Cai DJ, Rieth CA, Jones J, Ard MC. Sleep does not enhance motor sequence learning. *J Exp Psychol Learn Mem Cogn*. 2008;34(4):834-842. doi:10.1037/0278-7393.34.4.834
5. Brawn TP, Fenn KM, Nusbaum HC, Margoliash D. Consolidating the Effects of Waking and Sleep on Motor-Sequence Learning. *J Neurosci*. 2010;30(42):13977-13982.  
doi:10.1523/JNEUROSCI.3295-10.2010
6. Landry S, Anderson C, Conduit R. The effects of sleep, wake activity and time-on-task on offline motor sequence learning. *Neurobiol Learn Mem*. 2016;127:56-63.  
doi:10.1016/j.nlm.2015.11.009
7. Nettersheim A, Hallschmid M, Born J, Diekelmann S. The Role of Sleep in Motor Sequence Consolidation: Stabilization Rather Than Enhancement. *J Neurosci*. 2015;35(17):6696-6702. doi:10.1523/JNEUROSCI.1236-14.2015

8. Nemeth D, Janacsek K, Londe Z, Ullman MT, Howard DV, Howard JH. Sleep has no critical role in implicit motor sequence learning in young and old adults. *Exp Brain Res.* 2010;201(2):351-358. doi:10.1007/s00221-009-2024-x
9. Simor P, Zavecz Z, Horváth K, et al. Deconstructing Procedural Memory: Different Learning Trajectories and Consolidation of Sequence and Statistical Learning. *Front Psychol.* 2019;9. doi:10.3389/fpsyg.2018.02708
10. Pan SC, Rickard TC. Sleep and motor learning: Is there room for consolidation? *Psychol Bull.* 2015;141(4):812-834. doi:10.1037/bul0000009
11. Borragán G, Urbain C, Schmitz R, Mary A, Peigneux P. Sleep and memory consolidation: motor performance and proactive interference effects in sequence learning. *Brain Cogn.* 2015;95:54-61. doi:10.1016/j.bandc.2015.01.011
12. Humiston GB, Wamsley EJ. A brief period of eyes-closed rest enhances motor skill consolidation. *Neurobiol Learn Mem.* 2018;155:1-6. doi:10.1016/j.nlm.2018.06.002
13. Maier JG, Piosczyk H, Holz J, et al. Brief periods of NREM sleep do not promote early offline gains but subsequent on-task performance in motor skill learning. *Neurobiol Learn Mem.* 2017;145:18-27. doi:10.1016/j.nlm.2017.08.006
14. Astill RG, Piantoni G, Raymann RJEM, et al. Sleep spindle and slow wave frequency reflect motor skill performance in primary school-age children. *Front Hum Neurosci.* 2014;8. doi:10.3389/fnhum.2014.00910
15. Breton J, Robertson EM. Dual enhancement mechanisms for overnight motor memory consolidation. *Nat Hum Behav.* 2017;1(6). doi:10.1038/s41562-017-0111
16. Bottary R, Sonni A, Wright D, Spencer RMC. Insufficient chunk concatenation may underlie changes in sleep-dependent consolidation of motor sequence learning in older adults. *Learn Mem.* 2016;23(9):455-459. doi:10.1101/lm.043042.116

17. Diekelmann S. Neuroscience: Sleep, memories, and the brain. *Nat Hum Behav.* 2017;1(6):1-2. doi:10.1038/s41562-017-0124
18. Gregory MD, Agam Y, Selvadurai C, et al. Resting state connectivity immediately following learning correlates with subsequent sleep-dependent enhancement of motor task performance. *NeuroImage.* 2014;102(0 2):666-673. doi:10.1016/j.neuroimage.2014.08.044
19. Tucker MA, Nguyen N, Stickgold R. Experience Playing a Musical Instrument and Overnight Sleep Enhance Performance on a Sequential Typing Task. *PLOS ONE.* 2016;11(7):e0159608. doi:10.1371/journal.pone.0159608
20. Wamsley EJ, Hamilton K, Graveline Y, Manceor S, Parr E. Test Expectation Enhances Memory Consolidation across Both Sleep and Wake. *PLOS ONE.* 2016;11(10):e0165141. doi:10.1371/journal.pone.0165141
21. Adi-Japha E, Karni A. Time for considering constraints on procedural memory consolidation processes: Comment on Pan and Rickard (2015) with specific reference to developmental changes. *Psychol Bull.* 2016;142(5):568-571. doi:10.1037/bul0000048
22. King BR, Saucier P, Albouy G, et al. Cerebral Activation During Initial Motor Learning Forecasts Subsequent Sleep-Facilitated Memory Consolidation in Older Adults. *Cereb Cortex N Y N 1991.* 2017;27(2):1588-1601. doi:10.1093/cercor/bhv347
23. Ioannidis JPA. Why Most Published Research Findings Are False. *PLOS Med.* 2005;2(8):e124. doi:10.1371/journal.pmed.0020124
24. Collaboration OS. Estimating the reproducibility of psychological science. *Science.* 2015;349(6251). doi:10.1126/science.aac4716
25. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci.* 2011;22(11):1359-1366. doi:10.1177/0956797611417632

26. Asendorpf JB, Conner M, Fruyt FD, et al. Recommendations for Increasing Replicability in Psychology. *Eur J Personal*. 2013;27(2):108-119. doi:10.1002/per.1919
27. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376. doi:10.1038/nrn3475
28. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Res Synth Methods*. 2014;5(1):60-78. doi:10.1002/jrsm.1095
29. Stanley TD. Limitations of PET-PEESE and Other Meta-Analysis Methods: *Soc Psychol Personal Sci*. Published online February 28, 2017. doi:10.1177/1948550617693062
30. Alinaghi N, Reed WR. Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Res Synth Methods*. 2018;9(2):285-311. doi:10.1002/jrsm.1298
31. Tucker M, McKinley S, Stickgold R. Sleep optimizes motor skill in older adults. *J Am Geriatr Soc*. 2011;59(4):603-609. doi:10.1111/j.1532-5415.2011.03324.x
32. Backhaus W, Braaß H, Renné T, Krüger C, Gerloff C, Hummel FC. Daytime sleep has no effect on the time course of motor sequence and visuomotor adaptation learning. *Neurobiol Learn Mem*. 2016;131:147-154. doi:10.1016/j.nlm.2016.03.017
33. Vien C, Boré A, Lungu O, et al. Age-related white-matter correlates of motor sequence learning and consolidation. *Neurobiol Aging*. 2016;48:13-22. doi:10.1016/j.neurobiolaging.2016.08.006
34. Cai DJ, Rickard TC. Reconsidering the role of sleep for motor memory. *Behav Neurosci*. 2009;123(6):1153-1157. doi:10.1037/a0017672
35. Sheth BR, Janvelyan D, Khan M. Practice Makes Imperfect: Restorative Effects of Sleep on Motor Learning. *PLOS ONE*. 2008;3(9):e3190. doi:10.1371/journal.pone.0003190

36. Rickard TC, Pan SC. Time for considering the possibility that sleep plays no unique role in motor memory consolidation: Reply to Adi-Japha and Karni (2016). *Psychol Bull.* 2017;143(4):454-458. doi:10.1037/bul0000094
37. Backhaus W, Kempe S, Hummel FC. The effect of sleep on motor learning in the aging and stroke population - a systematic review. *Restor Neurol Neurosci.* 2015;34(1):153-164. doi:10.3233/RNN-150521
38. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. *Science.* 2014;345(6203):1502-1505. doi:10.1126/science.1255484
39. Engel M, Matosin N. Positives in negative results: when finding “nothing” means something. *The Conversation.* Accessed October 12, 2020.  
<http://theconversation.com/positives-in-negative-results-when-finding-nothing-means-something-26400>
40. Shanahan LK, Gjorgieva E, Paller KA, Kahnt T, Gottfried JA. Odor-evoked category reactivation in human ventromedial prefrontal cortex during sleep promotes memory consolidation. de Lange F, Behrens TE, eds. *eLife.* 2018;7:e39681. doi:10.7554/eLife.39681
41. Studte S, Bridger E, Mecklinger A. Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain Lang.* 2017;167:28-35. doi:10.1016/j.bandl.2016.03.003
42. Gudberg C, Wulff K, Johansen-Berg H. Sleep-dependent motor memory consolidation in older adults depends on task demands. *Neurobiol Aging.* 2015;36(3):1409-1416. doi:10.1016/j.neurobiolaging.2014.12.014



43. Cedernaes J, Sand F, Liethof L, et al. Learning and sleep-dependent consolidation of spatial and procedural memories are unaltered in young men under a fixed short sleep schedule. *Neurobiol Learn Mem.* 2016;131:87-94. doi:10.1016/j.nlm.2016.03.012
44. Fogel S, Albouy G, King BR, et al. Reactivation or transformation? Motor memory consolidation associated with cerebral activation time-locked to sleep spindles. *PLOS ONE.* 2017;12(4):e0174755. doi:10.1371/journal.pone.0174755

### **Figure Captions List**

Figure 1. Distribution of experimental group sample sizes in the published sleep and motor sequence literature.

Figure 2. Relationship between SE and effect size in the sleep and motor sequence literature as of 2015.

## Tables

*Table 1. Effect Size Predictors for the Sleep and Motor Sequence Literature*

Predictor	Description
Sleep status	wake vs. sleep group
Data averaging	Amount of data that was averaged to calculate pre-post performance gains between the training and test sessions.
Performance duration	Amount of time per training block
Training duration	Total time devoted to training
Time of testing	Time of day that the test session occurred (
Time of testing squared	Time of day that the test session occurred, squared
Elderly status	> 59 years of age

*Note.* Predictors drawn from the final working model of the 2015 meta-analysis.<sup>10</sup>

*Table 2. PET Results for the Full Dataset*

Analysis type	Predictor/intercept	Effect size estimate ( <i>d</i> )	<i>SE</i>	<i>p</i>
Case 1: No predictors	Intercept	-0.28	0.19	.15
	SE	3.05	0.59	<.0001
Case 2: Sleep status predictor included	Intercept	-0.19	0.18	.31
	SE	3.16	0.56	<.0001
	Sleep-status	-0.47	0.14	.0010
Case 3: All predictors included	Intercept	-0.70	0.14	<.0001
	SE	1.94	0.42	<.0001
	Sleep-status	-0.17	0.12	.16
	Data averaging (s)	0.0098	0.0018	<.0001
	Performance duration (s)	0.030	0.0071	<.0001
	Training duration (s)	-0.00093	0.00042	.032
	Time of testing (hrs)	-0.011	0.013	.39
	Time of testing squared (hrs)	-0.014	0.0020	<.0001
	Elderly-status	-1.41	0.17	<.0001

**Table 3. Recent Studies of Sleep and Explicit Motor Learning**

Reference	Mean <i>n</i> per group	Group(s)	Post-sleep gain	Relative gain
Astill et al. (2014) <sup>14</sup>	30	Children, 12-hr sleep vs. wake	positive (s)	positive (s)
Gregory et al. (2014) <sup>18</sup>	10	Children, 24-hr sleep group	positive (s)	—
		Adults, 12-hr sleep vs. wake	positive (s)	positive (s)
Gudberg et al. (2015) <sup>42</sup>	11.4	Adults, 24-hr sleep (with fMRI)	positive (s)	—
		Young adults, sleep vs. wake	positive (s)	positive (s)
Backhaus et al. (2016) <sup>32</sup>	11.7	Older adults, sleep vs. wake	positive (ns)	positive (ns)
		Adults, nap vs. wake	positive (ns)	negative (ns)
Backhaus et al. (2016) <sup>37</sup>	11	Adults, long nap vs. wake	positive (ns)	negative (ns)
		Older adults, nap vs. wake	negative (ns)	negative (ns)
Bottary et al. (2016) <sup>16</sup>	18.3	Older adults, long nap vs. wake	negative (ns)	negative (ns)
		Young adults, sleep vs. wake	positive (s)	positive (s)
Cedernaes et al. (2016) <sup>43</sup>	16	Older adults, sleep vs. wake	positive (ns)	positive (ns)
		Adults, 8.5-hr sleep ("NSS")	positive (s)	—
		Adults, 4.25-hr sleep ("SSS")	positive (s)	—
Tucker et al. (2016) <sup>19</sup>	10	Sleep vs. wake	positive (s)	positive (s)
Vien et al. (2016) <sup>33</sup>	14.3	Young, nap vs. No nap	positive (s)	positive (s)
		Old, nap vs. No nap	positive (s)	negative (s)
Wamsley et al. (2016) <sup>20</sup>	24.3	Sleep (combined) vs. wake (combined)	positive (s)	positive (ns)
Fogel et al. (2017) <sup>44</sup>	13	Sleep	positive (s)	—
Maier et al. (2017) <sup>13</sup>	18	Sleep (combined) vs. wake (combined)	positive (s)	positive (ns)