JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0000943

Overlearning of non-native speech sounds does not result in superior consolidation after a period of sleep

Pamela Fuhrmeister,[1,a] Garrett Smith,[2] and Emily B. Myers[1]

[1]*Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Drive, Storrs, Connecticut 06269, USA*

[2]*Universität Potsdam, Department of Linguistics, Haus 14, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany*

pamela.fuhrmeister@uconn.edu, gasmith@uni-potsdam.de, emily.myers@uconn.edu

Abstract: Recent studies suggest that sleep-mediated consolidation processes help adults learn non-native speech sounds. However, overnight improvement was not seen when participants learned in the morning, perhaps resulting from native-language interference. The current study trained participants to perceive the Hindi dental/retroflex contrast in the morning and tested whether increased training can lead to overnight improvement. Results showed overnight effects regardless of training amount. In contrast to previous studies, participants in this study heard sounds in limited contexts (i.e., one talker and one vowel context), corroborating other findings, suggesting that overnight improvement is seen in non-native phonetic learning when variability is limited.

[Editor: Martin Cooke] Pages: EL289–EL294

Received: 7 January 2020 **Accepted:** 4 March 2020 **Published Online:** 25 March 2020

1. Introduction

Learning to perceive speech sounds in a second language is difficult for adults, especially when second-language speech sounds are perceptually similar to native-language speech sounds (e.g., Best *et al.*, 2001). Recent work has shown that sleep can help learners consolidate perceptually learned non-native speech sounds into long-term memory (Earle and Myers, 2015a; Earle *et al.*, 2017; see Earle and Myers, 2014, for review), recover learned speech information that had been interfered with through throughout the day (Fenn *et al.*, 2003), and generalize their knowledge of new speech categories to a new talker (Earle and Myers, 2015b; Fenn *et al.*, 2013). However, there may be some limitations to the benefits of sleep for non-native speech sound learning. Specifically, Earle and Myers (2015a) trained two groups (one in the evening and another in the morning) on the voiced dental and retroflex stop consonants found in Hindi (a difficult contrast for English speakers; Best *et al.*, 2001). Participants returned for testing approximately 12 and 24 h later. Participants who learned in the evening improved after sleep, but those who learned in the morning did not. They argued that this was a result of interference from the native language before sleep: Morning-trained participants likely heard and produced a great deal of native-language speech sounds before sleeping, and hearing native-language speech tokens (especially those that are perceptually similar to the non-native sounds being learned) may have destabilized memory traces of the non-native speech sounds and attenuated behavioral improvements that result from sleep consolidation processes. Qin and Zhang (2019) found similar results on a non-native tone learning task. They found that, at least for the trained identification task, evening-trained participants improved marginally after sleep, but those trained in the morning did not. Both of these studies suggest that memory traces of non-native speech sounds may decay or be interfered with, but that sleep may help restore or protect memory traces from interference (similar to findings from Fenn *et al.*, 2003). In addition, these studies suggest that proximity to sleep is important for learning.

1.1 Interference in memory consolidation

Poor maintenance of phonetic information could result from weaker encoding of memory traces of the sounds or interference during a critical stabilization period after learning (see Fuhrmeister, 2019, for review). For instance, theories of consolidation by Müller and Pilzecker (1900) and Walker (2005) both suggest that newly formed memory traces need to undergo a stabilization phase. In Walker's (2005) model, stabilization relies on the passage of time, but task improvement

[a]Author to whom correspondence should be addressed.

results from a subsequent phase of consolidation that relies on sleep. Related to this idea, research from many domains suggests that consolidation in part depends on how strongly information is encoded during learning (e.g., Ebbinghaus, 1885; Hauptmann *et al.*, 2005). Shibata *et al.* (2017) had participants perform two similar visual learning tasks consecutively. This often causes the learner to forget what they learned on the first task when tested after a delay, due to interference from the second task (e.g., Brashers-Krug *et al.*, 1996). In Shibata *et al.* (2017), one group of participants overlearned the first task (continued to practice the task even after it was mastered) and one group did not. The group that overlearned showed no interference effects from the second task when tested the next day, whereas the group who practiced the task for the typical amount of time did experience interference (i.e., they showed weaker learning of the first task). These theories and empirical findings suggest that newly formed memory traces can be disrupted by what happens after learning, but interference may be mitigated if information is encoded strongly to begin with, or "overlearned." Testing whether theories of consolidation can be applied to non-native speech sound learning is both of theoretical and practical importance: Many second-language courses take place in the morning hours when learners will immediately leave the classroom and hear and speak their native language. Therefore, it is important to identify ways, such as overlearning, to mitigate interference effects. From a theoretical standpoint, testing predictions from the memory consolidation literature can shed light on whether some of the challenges of non-native speech sound learning stem from failures of consolidation.

1.2 Current study

The current study draws from the consolidation literature and tests the hypothesis that participants who learn a difficult non-native phonetic contrast in the morning hours [as in Earle and Myers (2015a), where these participants did not improve after sleep] will indeed show overnight improvement if they overlearn the phonetic contrast (as in Shibata *et al.*, 2017). We trained two groups of participants on the Hindi dental/retroflex contrast in the morning, with reassessments approximately 12 and 24 h later. We based the number of training trials on pilot testing which suggested that performance tended to asymptote at about 300 trials. To ensure that both overlearning and non-overlearning training groups had an equal opportunity to learn, we set the non-overlearning group at 300 trials [similar to training amounts in Earle and Myers (2015a) and Fuhrmeister and Myers (2020)], and following methods in Shibata *et al.* (2017), added 20 min of additional training for the overlearning group (800 trials total). Based on Earle and Myers (2015a), we predicted that the non-overlearning group would not show improvement after the overnight interval. If the overlearning group does improve after sleep, this would support the idea that strong initial encoding of memory traces is important for sleep-mediated consolidation processes and that it might be sufficient to mitigate interference caused by exposure to native-language speech sounds. If neither group shows overnight improvement, this would suggest that the time of day of initial learning may be more important for consolidation-based improvements rather than how strongly the information is learned.

2. Method

2.1 Participants

Sixty-three participants (17 male, 46 female, ages 18–34) were recruited from the University of Connecticut community and the Psychology Department participant pool. Participants were monolingual native speakers of North American English with no history of speech, hearing, or language disorders. Five participants were eliminated for not completing all sessions, two because of missing data files, two for previous exposure to speech sounds in Hindi, and two who reported that they were not native speakers of English. Data from the remaining 52 participants (overlearning = 29, non-overlearning = 23) are included in the analyses reported. Participants gave informed consent according to the University of Connecticut Institutional Review Board and were paid $10 per hour or received course credit for participation.

2.2 Stimuli

Experimental stimuli were presented using OpenSesame experiment presentation software (Mathôt *et al.*, 2012). Auditory stimuli were presented using over-ear headphones (SONY MDR-7506, New York), and participants could adjust the volume to a comfortable level. Auditory stimuli were five acoustically unique recordings each of /ɖug/ and /d̪ug/ recorded by a female native speaker of Hindi and scaled to mean amplitude of 65 dB in Praat (Boersma and Weenink, 2018). Visual stimuli were two novel objects (stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/). Participants responded via key press.

https://doi.org/10.1121/10.0000943

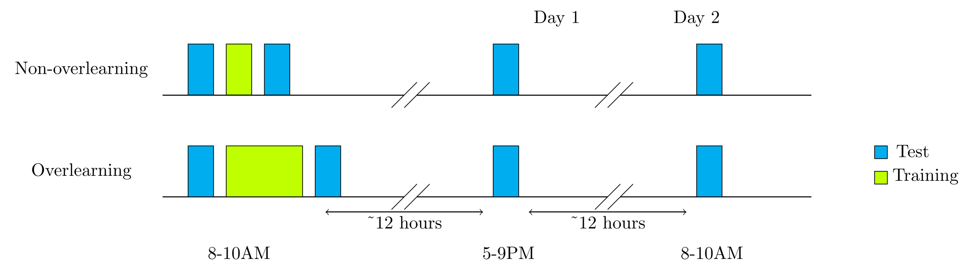

Fig. 1. (Color online) Schematic of experiment procedure for overlearning and non-overlearning groups. Procedure was identical except for the number of training trials groups completed.

2.3 Procedure

Participants visited the lab three times (see Fig. 1). We trained all participants in the morning hours to test whether overlearning would facilitate overnight improvement (or would mitigate native-language interference that participants experienced throughout the day). The first visit took place between 8:00 and 10:00 a.m., the second on the same day between 5:00 and 9:00 p.m., and the third the following morning between 8:00 and 10:00 a.m. The first session included an AX discrimination test (same/different judgment) to measure baseline discrimination of the contrast, identification training, an identification assessment, and a post-training AX discrimination assessment. The second and third sessions consisted of reassessments of identification and AX discrimination.

2.4 Training and assessments

AX discrimination. Participants heard two tokens (e.g., /ɖug/ and /d̪ug/) in a row with a 1 s interstimulus interval (64 trials). Half of the trials contained tokens from the same category, and half contained trials from different categories. Same-category trials always contained two acoustically distinct recordings.

Identification training and test. For the identification tasks, participants heard one auditory token and were asked to choose one of two novel visual objects. Before training, participants saw one object at a time on the screen and heard five examples of the corresponding auditory stimulus to learn the stimulus pairings. The non-overlearning group completed 300 identification training trials (about 10–15 min of practice), and the overlearning group completed 800 trials (approximately 20 min of extra training). Participants received visual feedback on each trial during training (e.g., "Correct!" or "Incorrect"). Identification assessments consisted of 50 trials without feedback. Identification and discrimination assessments were used in tandem, as these tasks may differ in overnight consolidation patterns (e.g., Earle and Myers, 2015a; Qin and Zhang, 2019; see Earle and Myers, 2014, for a discussion).

2.5 Analysis approach

For analyses of discrimination, d′ scores were calculated for each participant to minimize response bias (MacMillan and Creelman, 2005). The d′ scores served as the dependent variable in a linear mixed effects model. Identification data were analyzed using mixed effects logistic regression models.[1] Analyses were performed in R (R Development Core Team, 2008), and mixed effects models were fit with the lme4 package (Bates *et al.*, 2015). For linear mixed effects models, *p*-values were estimated using the Satterthwaite approximation with the afex package (Singmann *et al.*, 2019). For analyses of identification data, a backwards stepping procedure was used to determine the random effects structure that best fit the data (Matuschek *et al.*, 2017). Raw data and analysis scripts can be found at https://osf.io/hm24w/.[2]

3. Results

3.1 Discrimination

First, we analyzed data from the discrimination task to test for learning immediately after training, and maintenance or improvement over the two post-tests. The dependent variable was the d′ score, and fixed effects included factors for time (pretest, immediate post-test, evening post-test, and next-morning post-test) and group (overlearning and non-overlearning). To test for group differences, the factor group was deviation coded (overlearning = −0.5, non-overlearning = 0.5). To test for group differences over time, we employed backwards difference coding using the cont.sdif() function in the MASS package to test the following contrasts: immediate post-test–pretest (improvement after training), evening post-test–immediate post-test (maintenance over the day), and next-morning post-test–evening post-test (overnight improvement). Random intercepts for

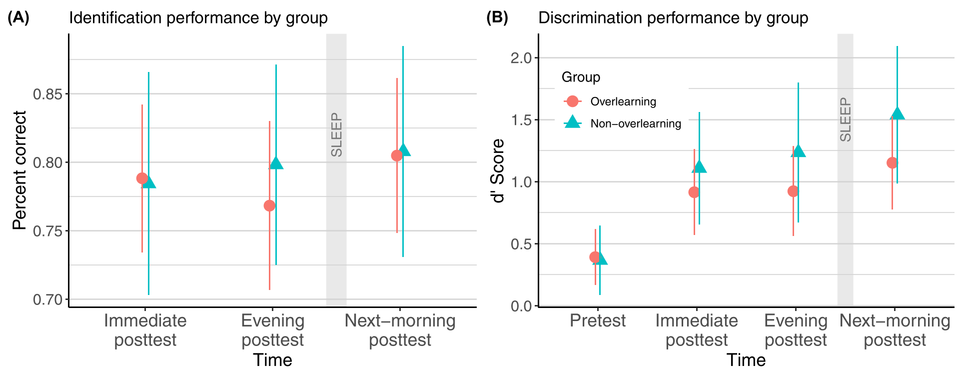

Fig. 2. (Color online) Identification and discrimination performance by group at each time point. Participants improved after training on the discrimination task, maintained training-induced gains on identification and discrimination measures, and improved after a period of sleep on both discrimination and identification. No group differences or interactions were found. Error bars indicate 95% confidence intervals.

participants were included. The model revealed a significant difference between the pretest and immediate post-test, $\beta = 0.63$, SE $= 0.11$, $t = 5.60$, $p < 0.001$, which means that, overall, participants improved as a result of training. No difference was seen between the immediate post-test and the evening post-test, suggesting that participants maintained training-induced gains over the course of the day. A significant difference was found between the evening post-test and next-morning post-test, $\beta = 0.27$, SE $= 0.11$, $t = 2.35$, $p = 0.02$, which indicates that performance improved overnight. No group differences or interactions were found, suggesting that the extra training did not result in superior learning, maintenance, or consolidation of the sounds (see Fig. 2).

3.2 Identification

Next, we turn to an analysis of the identification task, which was the task that was explicitly trained in this study. The dependent variable of this model was the log odds of selecting a correct response on each trial, and fixed effects included time (immediate post-test, evening post-test, and next-morning post-test) and group (overlearning and non-overlearning, coded as in discrimination analysis). Time was backwards difference coded to compare contrasts of evening post-test–immediate post-test (maintenance over the day) and next-morning post-test–evening post-test (overnight improvement). The random effects structure in the final model included by-subject random intercepts. The model revealed a significant difference only between the evening post-test and the next-morning post-test, $\beta = 0.16$, SE $= 0.07$, $z = 2.13$, $p = 0.03$, which suggests that performance improved overnight. We observed no effect of group and no interactions, suggesting that the groups did not differ in their identification performance at any time point or in their degree of overnight improvement (see Fig. 2).

3.3 Bayesian analyses of discrimination and identification

Null effects in frequentist analyses are difficult to interpret; one cannot validly conclude that a lack of a significant difference between conditions is evidence *for* the absence of a difference. However, Bayesian approaches do allow one to provide evidence for the lack of an effect using Bayes factor analyses (Kass and Raftery, 1995; Kruschke, 2015). Bayes factors quantify the relative evidence for the null hypothesis (e.g., no difference between overlearning and non-overlearning) compared to the alternative hypothesis (e.g., a difference between groups). To take advantage of this, we reanalyzed the discrimination and identification data using Bayesian mixed effects models. The results were fully compatible with the results of the frequentist analysis reported here. The Bayes factor analyses revealed no support for group differences and no evidence for an interaction with time. See supplementary material for a detailed description of the Bayesian analyses.[2]

4. Discussion

The goal of the current study was to test whether overlearning a difficult non-native speech contrast would help participants capitalize on sleep-dependent consolidation processes and improve after sleep as compared to a group that did not overlearn. We found that, overall, participants improved after training on the discrimination task, maintained these gains throughout the day, and, in contrast to prior work from our lab and others, also improved on both discrimination and identification of the contrast after sleeping. Surprisingly, overlearning the non-native contrast with an additional 20 min of training did not confer any benefits above and beyond the shorter amount of training. A possible explanation is that although our chosen way of manipulating overlearning followed prior literature (Shibata *et al.*, 2017), there are many potential ways to test whether overlearning supports consolidation. For example, future research could calibrate the amount of training to the individual (i.e., have each participant practice until some threshold is reached, and then overlearn). It is still of interest that, in contrast to Earle and Myers (2015a) and Qin and Zhang (2019), we saw an increase in performance following the overnight interval when participants were trained in the morning, as indicated by the effects of time in both the discrimination and identification analyses. We discuss potential reasons for this below.

4.1 Destabilization of memory traces as a result of phonological variability

One key difference between the current study and Earle and Myers (2015a) is that participants in Earle and Myers (2015a) were additionally tested on the sounds they learned in the context of an untrained vowel, and in the current study, participants only heard the sounds presented in one vowel context. Similarly, Qin and Zhang (2019) tested generalization of tone learning to a non-native talker. This suggests that exposure to variability, either in the form of multiple phonological contexts or multiple talkers, may destabilize learning to the point that task improvement as a result of sleep-dependent consolidation does not occur (as in Fuhrmeister and Myers, 2017). The current study and others have shown overnight improvement when phonological variability in test and training is limited (Earle *et al.*, 2017; Fuhrmeister and Myers, 2017), whereas findings of improvement after sleep in studies that included phonological variability in training or test are inconsistent (Earle and Myers, 2015a; Fuhrmeister and Myers, 2017, 2020). This idea may be counterintuitive because several studies have found that exposure to variability in non-native phonetic training is beneficial for learning and generalization (e.g., Lively *et al.*, 1993). However, in many of these studies, participants engaged in many training sessions over several weeks. In this case, variability is likely helpful, but in situations where large doses of training are not possible, it is of interest to understand how stability of learning (which seems to be influenced by exposure to variability) interacts with memory consolidation.

The current findings add to our understanding of sleep-dependent memory consolidation and interference in memory consolidation of non-native speech sounds in two important ways. First, they bolster arguments for sleep made by Earle and Myers (2014, 2015a) and Earle *et al.* (2017). Based on the findings of Earle and Myers (2015a) alone, it was hard to make the argument that sleep *per se* resulted in task improvement after a period of sleep because this was only seen in participants who trained during the evening hours. The current study suggests that overnight improvement is possible when participants learn in the morning hours and that limited phonological variability during testing or training may be a more reliable predictor of overnight improvement than native-language exposure, as was found in Earle and Myers (2015a). Second, the current findings add to a growing body of evidence in the domain of non-native speech sound learning that suggests that stability or strength of learning is important for consolidation-based improvements after sleep and that there may be multiple ways to achieve stability in learning. For example, the current study, combined with evidence from Fuhrmeister and Myers (2017) suggests that limiting phonological variability in training and testing can stabilize learning, and findings from Earle and Myers (2015a) and Qin and Zhang (2019) suggest that sleeping soon after information is learned can have a stabilizing effect. These findings are consistent with domain-general models of memory consolidation that posit that strong initial learning (Ebbinghaus, 1885) or stabilization of memory traces is necessary for task improvement during sleep (Müller and Pilzecker, 1900; Walker, 2005). An open question is whether overlearning could help mitigate interference from variability that is induced by training or testing on additional vowel contexts or talkers, and we suggest that future research address this question.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grant Nos. DGE-1747486 and NSF IGERT DGE-1144399 to the University of Connecticut, and NSF BCS 1554510 to E.B.M. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References and links

[1]In tests of identification with no feedback, participants sometimes confuse the category labels, even when they can distinguish the sounds. This results in a very low accuracy score, which is unlikely to occur by chance. For any participants that scored lower than chance as indicated by a binomial test, we recoded the trial-level data (0 was recoded as 1, and 1 was recoded as 0; 1 = correct, 0 = incorrect). Only one participant in the overlearning group at the next-morning post-test was affected.

[2]See supplementary material at https://doi.org/10.1121/10.0000943 for tables with coefficients from each model as well as a detailed description of the Bayesian analyses.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**(1), 1–48.

Best, C. T., McRoberts, G. W., and Goodell, E. (**2001**). "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," J. Acoust. Soc. Am. **109**(2), 775–794.

Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer" [Computer Software], Version 6.0.43.

Brashers-Krug, T., Shadmehr, R., and Bizzi, E. (**1996**). "Consolidation in human motor memory," Nature **382**(6588), 252–255.

Earle, F. S., Landi, N., and Myers, E. B. (**2017**). "Sleep duration predicts behavioral and neural differences in adult speech sound learning," Neurosci. Lett. **636**, 77–82.

Earle, F. S., and Myers, E. B. (**2014**). "Building phonetic categories: An argument for the role of sleep," Front. Psychol. **5**, 1192.

Earle, F. S., and Myers, E. B. (**2015a**). "Sleep and native language interference affect non-native speech sound learning," J. Exp. Psychol. **41**(6), 1680–1695.

Earle, F. S., and Myers, E. B. (**2015b**). "Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds," J. Acoust. Soc. Am. **137**(1), EL91–EL97.

Ebbinghaus, H. (**1885**). *Memory: A Contribution to Experimental Psychology* (Verlag von Duncker und Humblot, Leipzig, Germany).

Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (**2013**). "Sleep restores loss of generalized but not rote learning of synthetic speech," Cognition **128**(3), 280–286.

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (**2003**). "Consolidation during sleep of perceptual learning of spoken language," Nature **425**(6958), 614–616.

Fuhrmeister, P. (**2019**). "Interference in memory consolidation of non-native speech sounds," in *Speech Production and Perception: Learning and Memory*, edited by S. Fuchs, J. Cleland, and A. Rochet-Capellan (Peter Lang, Berlin, Germany), pp. 207–243.

Fuhrmeister, P., and Myers, E. B. (**2017**). "Non-native phonetic learning is destabilized by exposure to phonological variability before and after training," J. Acoust. Soc. Am. **142**(5), EL448–EL454.

Fuhrmeister, P., and Myers, E. B. (**2020**). "Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on non-native phonetic learning," Attn., Percept., Psychophys. (published online).

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

Kass, R. E., and Raftery, A. E. (**1995**). "Bayes factors," J. Am. Stat. Assoc. **90**(430), 773–795.

Kruschke, J. K. (**2015**). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, 2nd ed. (Academic Press, New York).

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. **94**(3), 1242–1255.

Macmillan, N. A., and Creelman, C. D. (**2005**). Detection Theory: A User's Guide.

Mathôt, S., Schreij, D., and Theeuwes, J. (**2012**). "OpenSesame: An open-source, graphical experiment builder for the social sciences" [OpenSesame 3.1.5] https://osdoc.cogsci.nl/3.1.5/ (Last viewed 19 March 2020).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (**2017**). "Balancing Type I error and power in linear mixed models," J. Memory Lang. **94**, 305–315.

Müller, G. E., and Pilzecker, A. (**1900**). *Experimentelle Beiträge zur Lehre vom Gedächtnis* (J. A. Barth, Leipzig), Vol. 1.

Qin, Z., and Zhang, C. (**2019**). "The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts," PLoS One **14**(12), e0221498.

R Development Core Team (**2008**). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (Last viewed 19 March 2020).

Shibata, K., Sasaki, Y., Bang, J. W., Walsh, E. G., Machizawa, M. G., Tamaki, M., and Watanabe, T. (**2017**). "Overlearning hyper-stabilizes a skill by rapidly making neurochemical processing inhibitory-dominant," Nature Neurosci. **20**(3), 470–475.

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (**2019**). *afex: Analysis of Factorial Experiments*, R package version 0.23-0. https://CRAN.R-project.org/package=afex (Last viewed 19 March 2020).

Walker, M. P. (**2005**). "A refined model of sleep and the time course of memory formation," Behav. Brain Sci. **28**(1), 51–64.