Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds

F. Sayako Earle and Emily B. Myers[a)]
Department of Speech, Language, and Hearing Sciences, University of Connecticut, 850 Bolton Road, Unit 1085, Storrs, Connecticut 06269
Frances.Earle@uconn.edu, Emily.Myers@uconn.edu

Abstract: This investigation explored the generalization of phonetic learning across talkers following training on a nonnative (Hindi dental and retroflex) contrast. Participants were trained in two groups, either in the morning or in the evening. Discrimination and identification performance was assessed in the trained talker and an untrained talker three times over 24 h following training. Results suggest that overnight consolidation promotes generalization across talkers in identification, but not necessarily discrimination, of nonnative speech sounds.

[AC]
Date Received: August 21, 2014 **Date Accepted:** November 13, 2014

1. Introduction

Acquisition of nonnative speech sounds posits a learning challenge for adults. One difficulty concerns the generalization of training beyond the learning context, such that the distinctive features of the nonnative sounds can be applied to an unfamiliar talker's voice or to sounds occurring in a different vowel environment. Evidence suggests that generalization is not ubiquitous in perceptual training. For example, Lively, Logan, and Pisoni (Lively *et al.*, 1993) found that Japanese listeners trained to identify /r/ and /l/ in one talker's voice did not improve in identifying /r/ and /l/ produced by an unfamiliar talker, whereas listeners trained on multiple talkers did. Such studies have established that careful selection of varied phonetic tokens facilitates the generalization of perceptual training over a closed stimulus set in adults (Bradlow *et al.,* 1996; Jamieson and Morosan, 1989; Logan *et al.,* 1991). Pisoni (1992) suggests that variation provides insight into invariant features. However, as the training tokens employed in these studies were varied in vowel context, it is yet undetermined (to our knowledge) if training on a closed set of tokens *can* generalize to the recognition of target sounds produced by a different talker or occurring next to a different vowel.

Limiting the variability in input is crucial to identifying factors beyond the training environment that contribute to the generalization of phonetic information. Specifically, our approach highlights the memory encoding processes by which the invariant cues are thought to be abstracted from the acoustic instances experienced during training. Recent literature suggests that sleep precipitates these qualitative changes to memory (Diekelmann and Born, 2010; Rasch and Born, 2013). The *Complementary Systems Account of Learning* (McClelland *et al.,* 1995) predicts that a period of off-line systems consolidation of the episodic trace leads to the abstraction and integration of new information with preexisting knowledge (Davis and Gaskell, 2009; Tamminen *et al.*, 2013). The purpose of this study therefore is to identify the role of sleep in the generalization of training across talkers and across vowel contexts, following perceptual training of a nonnative contrast with a limited set of training tokens.

[a)]Also at: Department of Psychology, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020 and Haskins Laboratories, Yale, New Haven, CT 06511.

2. Methods

Thirty-eight (16 male, 22 female) students between 18 and 24 yr of age were recruited from University of Connecticut (UCONN) and received course credit for participation. Participants reported being monolingual speakers of American English with normal hearing and vision and provided informed consent according to the UCONN Institutional Review Board guidelines.

All experiment sessions were administered using E-PRIME 2.0 (Psychology Software Tools, Pittsburgh, PA). Two novel visual objects[1] were used to pair with the auditory tokens during training. Five unique tokens each of the syllables /ɖɛ/, /d̪ɛ/, /ɖa/, /d̪a/, naturally spoken by two male native speakers of Hindi, were digitally recorded onto a Macintosh laptop at Haskins Laboratories (New Haven, CT). Syllables were rescaled to mean amplitude of 70 dB sound pressure level (SPL) and cut to the onset of the burst using PRAAT software (Boersma and Weenink, 2011). Auditory tokens were presented through Hi-Fi digital sound monitor headphones (SONY MDR-7506) at an average listening level of 75 dB SPL (range: 44–80 dB SPL).

Participants were trained in a self-paced, forced-choice identification task (ID) of the dental and retroflex sounds in a modified version of a previous experiment (Earle and Myers, 2013). To address the effect of sleep, participants were assigned to two groups, those trained in the morning (morning group) or the evening (evening group) and then returned twice over 24 h for reassessment such that the overnight between-session interval occurred during sessions two and three for the morning group and between sessions one and two for the evening group [means and standard deviations of session times: Morning group, 8:27 a.m. (31 min), 6.05 p.m. (49 min), 8:22 a.m. (30 min); evening group, 7:01 p.m. (38 min), 8:30 a.m. (30 min), 6:36 p.m. (48 min)]. Participants completed a sleep questionnaire for the 24-h experiment period at the end of the experiment.

2.1 Training

Each participant was trained on a set of 10 tokens produced by a single talker in a single vowel context (5 tokens/target consonant) with the training talker and vowel counterbalanced across participants. Instructions indicated that participants would hear "words" that corresponded to two visual objects (the two Fribbles). Training consisted of 300 trials (150 each beginning with dental and retroflex, 30/token): Participants heard a /CV/ token with a target sound (dental or retroflex) at the onset and were prompted to choose the corresponding picture. Feedback was given after every trial.

2.2 Assessments

We tracked task performance on both ID and AX discrimination. In the ID posttests, participants completed 100 trials of the training task without feedback: The first 50 trials (5 trials/token) in the trained talker's voice, followed by 50 trials on the untrained talker, both in the trained vowel context only.[2] ID posttests were completed at three time points: Immediately after training (ID posttest 1) and at sessions 2 and 3 (ID posttest 2, ID posttest 3).

AX discrimination included 160 trials with 40 trials (20 "same" and 20 "different") in which the stimuli were of the same talker and the same vowel as training (TTTV), 40 trials with the trained talker and an untrained vowel (TTUV), 40 trials with an untrained talker and the trained vowel (UTTV), and 40 trials with the different talker and an untrained vowel (UTUV). The speaker and vowel context remained consistent within each trial. During each trial, participants were presented with two acoustic tokens (e.g., "/ɖɛ/… /ɖɛ/") separated by a 500 ms ISI and prompted to indicate whether the sounds at the onset belonged to the same or a different speech sound category. No two acoustically identical tokens were used in a single trial, such that listeners were required to make discrimination judgments on the basis of membership to the variant category. Participants completed the AX discrimination at four time points:

Before training (pretest) and immediately after the ID posttests at session 1 (posttest 1), at session 2 (posttest 2), and at session 3 (posttest 3).

3. Analyses and results

Of the 38 who participated, 11 participants were excluded: 7 for non-compliance with the experimental task, 2 for non-completion, 1 due to experimenter error, and 1 who reported sleeping during the day during the 24-h experiment period. Percent accuracy on the discrimination and identification tasks were converted to d′ scores (MacMillan and Creelman, 2004). In addition, to ensure that our data reflects participants who achieved even minimal success in training, we included data only from participants who obtained a d′ score higher than 0 (equivalent to >50% accuracy) on session 1 ID posttest on the trained talker (i.e., the trained task). We excluded data from five additional participants based on this criterion (two from morning, three from evening). Data from the remaining 22 participants (12 male, 10 female; 11/group) are included in the following analyses.

3.1 ID performance

Our results suggest that both groups improved their identification of tokens produced by an unfamiliar talker immediately following sleep but not before. To assess changes in ID performance over time, we conducted a $2 \times 3 \times 2$ mixed models analysis of variance (ANOVA) with group as the fixed factor and time (3 levels) and talker (2 levels) as within-subjects factors (see Table 1). There was a significant three-way interaction among time, talker, and group and two-way interactions between time and group and talker and group. In addition, we observed a trend toward a main effect of time. To examine the factors driving the three-way interaction, we ran additional 2×3 repeated measures ANOVAs for each talker separately (trained and untrained) with group as the fixed factor and time (3 levels) as the within-subjects factor (see Table 2). For the trained talker, there were no significant effects or interactions. For the untrained talker, there was a significant interaction between time and group, a significant time main effect, but no main effect of group. Thus it appears that the three-way interaction is driven by differences over time by group in the untrained talker.

Inspection of the results (Fig. 1), suggests that groups differ in (a) the time at which improvements on the untrained voice are observed and (b) the magnitude of this improvement. To further examine the reported time by group interaction within the untrained talker, we conducted two additional mixed models ANOVAs with group as the fixed factor and time (2 levels) as the within-subjects factor. The first of these compared performance between groups at sessions 1 and 2, where sleep has occurred for the evening but not the morning group. There was an interaction between time and group ($F_{2,20} = 16.884$, $p = 0.018$, $\eta^2 = 0.251$), a main effect of time ($F_{1,20} = 15.737$, $p = 0.021$, $\eta^2 = 0.238$), and a main effect of group ($F_{1,20} = 11.767$, $p = 0.037$, $\eta^2 = 0.200$). We examined this interaction by conducting paired samples *t*-tests to compare performances at

Table 1. ($2 \times 3 \times 2$) Mixed models analysis of variance on ID performance.

Main effects	df	MS	*F*	*p*	η^2
Time	2,40	5.113	5.113	0.057	0.134
Talker	2,40	1.546	0.754	0.396	0.036
Group	1,20	4.967	1.302	0.267	0.061
Interactions					
Talker × group	1,20	9.823	4.790	0.041[a]	0.193
Time × group	2,40	7.009	4.232	0.022[a]	0.175
Time × talker	2,40	3.172	2.023	0.146	0.092
Time × talker × group	2,40	1.051	3.860	0.029[a]	0.162

[a]Significant at 0.05 level.

Table 2. (2 × 3) Mixed models analysis of variance, by trained and untrained talker.

	Trained talker					Untrained talker				
Main effects	df	MS	F	p	η^2	df	MS	F	p	η^2
Time	2,40	0.018	0.012	0.915	0.001	2,40	8.018	3.735	0.033[a]	0.157
Group	1,20	4.967	1.302	0.269	0.061	1,20	4.856	2.409	0.136	0.107
Interactions										
Time × group	2,40	1.183	0.779	0.388	0.037	2,40	12.432	5.792	0.006[a]	0.225

[a]Significant at 0.05 level.

posttest 1 and posttest 2 by group. We applied the Bonferroni method of correction to control for family-wise error rate (FWER) at 0.05 in calculating confidence intervals (CI). Session 2 performance was higher then session 1 for the evening group [$t_{10} = -2.698$, $p = 0.022$, CI: (−4.813, −0.057)]; but there was no difference for the morning group [$t_{10} = 0.133$, $p = 0.897$, CI: (−0.812, 0.897)].[3] The second follow-up analysis compared maintenance of performance over the 24 h interval. To this end, we performed a 2 × 2 ANOVA comparing session 1 and session 3 posttests across groups. There was a significant main effect of time ($F_{1,20} = 5.125$, $p = 0.035$, $\eta^2 = 0.204$), with greater sensitivity to the contrast in the untrained talker's voice at session 3 than session 1, but no main effect of group ($F_{1,20} = 1.215$, $p = 0.283$, $\eta^2 = 0.057$) nor an interaction between time and group ($F_{1,20} = 0.130$, $p = 0.722$, $\eta^2 = 0.006$).

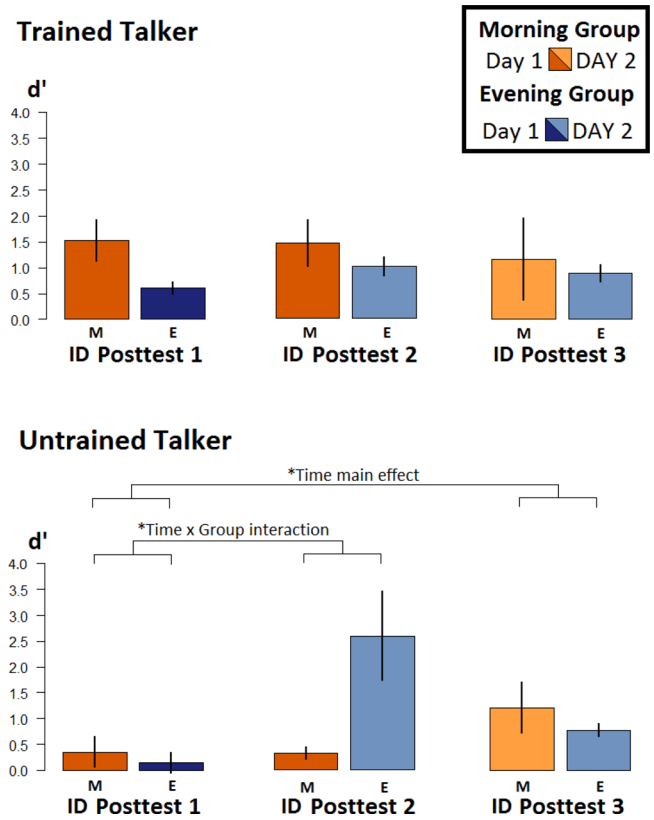

Fig. 1. (Color online) Identification performance over 24 h by group. Profile of performance changes across time underlying the significant interaction among time, group, and talker in the 2 × 3 × 2 mixed models ANOVA (see Tables 1 and 2). Error bars denote standard errors of the mean. *, significance at 0.05 level.

Table 3. Means of discrimination performance by condition over 24 h expressed in d′.[a]

	TTTV	UTTV	TTUV	UTUV	Mean
Morning group					
Pretest	0.02(0.43)	0.25(0.40)	0.11(0.46)	0.29(0.46)	0.17(0.22)
Posttest1	0.36(0.42)	0.75(1.64)	0.11(0.27)	0.66(0.61)	0.46(0.41)
Posttest2	0.08(0.45)	0.41(0.58)	0.19(0.99)	0.55(0.40)	0.31(0.26)
Posttest3	0.38(0.55)	0.11(0.52)	0.28(0.41)	0.49(0.52)	0.31(0.30)
Evening group					
Pretest	0.21(0.28)	0.07(0.54)	0.13(0.54)	0.40(0.29)	0.20(0.21)
Posttest1	0.26(0.41)	0.20(0.57)	0.32(0.51)	0.33(0.45)	0.28(0.19)
Posttest2	0.26(0.38)	0.32(0.29)	0.35(0.32)	0.28(0.73)	0.30(0.30)
Posttest3	0.25(0.59)	0.06(0.44)	0.09(0.47)	0.32(0.68)	0.18(0.28)

[a]Means and standard deviations in the four trial conditions.

To determine whether there was any evidence of generalization immediately after training, we performed one-sample *t*-tests on posttest 1 performances by group. Neither group differed from 0 [morning: $t_{10} = 1.210$, $p = 0.254$, CI: (−0.415, 1.228); evening: $t_{10} = 0.750$, $p = 0.471$, CI: (−0.376, 0.672); Bonferroni correction applied]. Cumulatively, this suggests that no generalization to the untrained talker is evident immediately after training but that significant generalization in the evening group emerges at session 2 and for both groups at session 3.

3.2 Discrimination performance

We conducted a $2 \times 4 \times 2 \times 2$ repeated measures ANOVA on the discrimination scores with group as the fixed factor and time (4 levels), talker (2 levels), and vowel (2 levels) as the within-subjects measures. There was a main effect of time ($F_{1,20} = 3.455$, $p = 0.040$, $\eta^2 = 0.379$), but no other main effects or interactions. We further explored the time main effect by collapsing across all four conditions across Groups at each time point, and conducting paired T-tests between pretest and posttest 1, posttests 1 and 2, and posttests 2 and 3. We found that posttest 1 trended higher than pretest after correction for FWER [$t_{21} = -2.205$, $p = 0.039$, CI: (−0.041, 034)] but that differences between posttests 1 and 2 and 2 and 3 were not statistically significant [$t_{21} = 0.785$, $p = 0.441$, CI: (−0.138, 0.268); $t_{21} = 0.734$, $p = 0.471$, CI: (−0.106, 0.222), respectively]. The means and standard deviations of discrimination performance for the four conditions are displayed in Table 3.

4. Discussion

Generalization of phonetic learning to a new talker's voice requires that the learner abstract acoustic-phonetic features from the details that disambiguate the contrast in the trained talker's voice. While we found evidence for the effects of sleep on generalization across talkers for ID performance, we observed no such effects on discrimination.

The profile of changes over time in identification performance (see Fig. 1) leads to the following interpretations. First, for ID performance in the trained talker, the lack of a main effect of time suggests that systems consolidation during sleep has little effect on identifying the tokens on which participants were trained. Given that abstraction away from trained material is not necessary in this case, this finding is not surprising. In the untrained talker, d′ scores are near chance immediately following training, suggesting that training does not immediately generalize across talkers. The morning and evening groups then diverged with both showing improvements in the untrained talker only following the overnight interval. In the evening group, performance appears to decline at session 3, but remains comparable to the morning group (as indicated by the 2×2 ANOVA on time points 1 and 3). This decline may be due to interference following reactivation of the memory trace after overnight consolidation (see

Dudai, 2004 for review). Previous data from our lab suggest that exposure to English (alveolar) /d/ subsequent to training interferes with the consolidation of trained dental and retroflex sounds (Earle and Myers, 2014). Similarly, reactivation of the memory trace during our tests may make listeners vulnerable to interference from subsequently heard talkers and/or English language input. This kind of exposure is more likely to occur during the daytime interval (between sessions 2 and 3 for the evening group; between sessions 1and 3 for the morning group) than at night (between sessions 1 and 2 for the evening group; between sessions 2 and 3 for the morning group). This interference-related decline subsequent to reactivation is documented in the learning and memory literature outside the linguistic domain (Stickgold and Walker, 2007; Talamini *et al.,* 2008; Walker *et al.,* 2003). Overall, the evidence suggests that significant generalization across talkers only appears to emerge following a period of sleep and not before. We note, however, that as our ID task only assessed performance in a single vowel context, whether or not sleep facilitates generalization across vowel contexts remains unanswered.

Discrimination performance differed markedly from performance on the ID task. The paired comparisons suggest that performance improves slightly immediately after training but that subsequent performance is stable for 24 h. Notably discrimination performance remained close to chance levels (see Table 3). Expected gains in discrimination may have been attenuated by our task design. Specifically, our discrimination blocks contained quite variable tokens with three-quarters of the trials containing untrained tokens. This variability may have made it difficult for listeners to generate consistent criteria for performing the discrimination task. Based on the current data alone, we are unable to make conclusive claims about perceptual ability as measured by our discrimination task.

The pattern shown in the ID data closely resembles other work showing abstraction or generalization as a function of sleep (e.g., Davis and Gaskell, 2009; Durrant *et al*., 2013; Fenn *et al*., 2003; Tamminen *et al.,* 2010; Tamminen *et al.,* 2012). For example, in Tamminen *et al.* (2012), reaction times to a speeded shadowing task decreased for the trained material immediately following morphological training, but generalization of performance to unfamiliar tokens was observed only when retested 2 days later. This suggests that consolidation during the between-session interval facilitated the improvement in task performance for tokens presented in an unfamiliar context. In this study and the present data, observed changes in performance are similarly indicative of qualitative changes in the sound-level information applied to the task. In particular, the contributing information no longer appears to be specific to the acoustic-phonetic context (our data) or morphophonemic context (Tamminen *et al.*, 2012) encountered during training. There are at least two explanations for this qualitative change. First, overnight consolidation may induce a qualitative change in the memory trace itself during information transfer, transforming it from a specific, acoustic/sensory-based trace to a more abstracted representation. Alternatively, generalization may instead reflect differences in the way information is retrieved from memory with a shift from episodic (training-specific) retrieval on day 1 to the use of features that are stripped of task-irrelevant details (abstracted) on day 2. Under either account, these findings support the interpretation that sleep promotes the abstraction of novel phonetic information, such that recall of category labels invoke nonspecific information that can be applied to identify tokens produced by an unfamiliar talker. This view is consistent with the *Complementary Systems Account of Learning* (McClelland *et al.,* 1995), and reports of sleep-mediated abstraction in other domains. In conclusion, while previous literature has focused on the role of input to the talker generalization of phonetic learning, this process appears to be facilitated also by the memory encoding processes during the sleep that follows training.

Acknowledgments

This work was supported by NIH NIDCD Grant Nos. R03 DC009495 and R01 DC013064 to E.B.M. and NIH NICHD Grant No. P01 HD001994 (Rueckl, PI). The content is the responsibility of the authors and does not necessarily represent official views of the NIH, NIDCD, or NICHD.

Reference and links

[1]Fribbles stimulus images are available from Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University at http://www.tarrlab.org/.

[2]Note that it was not possible for participants to perform the identification task in the untrained vowel context. Because participants are trained to treat each CV syllable as a "word" matching a novel object (e.g., /ɖɛ/ is the name of one Fribble, /ɖɛ/ is the name of the other Fribble), participants had no match to any object with untrained "words" (e.g., /ɖa/ and /ɖa/).

[3]Between-group comparisons at sessions 1 and 2 furthermore confirmed that performance between groups were not significantly different until time point 2 (Bonferroni correction applied). Values obtained for independent samples t-tests at each time point are as follows. Session 1: ($t_{20} = 0.579$, $p = 0.569$, CI: [−0.850, 760]), session 2: ($t_{20} = -2.605$, $p = 0.017$, CI: [−1.109, −3.816]).

Boersma, P., and Weenink, D. (**2011**). Praat: Doing phonetics by computer [Computer program]. Version 5.3. 03. 2011.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. I. (**1996**). "Three converging tests of improvement in speech production after perceptual identification training on a non-native phonetic contrast," J. Acoust. Soc. Am. **100**(4), 2725.

Davis, M. H., and Gaskell, M. G. (**2009**). "A complementary systems account of word learning: Neural and behavioural evidence," Philos. Trans. R. Soc. B. Biol. Sci. **364**(1536), 3773–3800.

Diekelmann, S., and Born, J. (**2010**). "The memory function of sleep," Nat. Rev. Neurosci. **11**(2), 114–126.

Dudai, Y. (**2004**). "The neurobiology of consolidations, or, how stable is the engram?," Annu. Rev. Psychol. **55**, 51–86.

Durrant, S. J., Cairney, S. A., and Lewis, P. A. (**2013**). "Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum," Cereb. Cortex **23**(10), 2467–2478.

Earle, S., and Myers, E. (**2013**). "The effect of sleep on learned sensitivity to a non-native phonetic contrast," J. Acoust. Soc. Am. **134**(5), 4107.

Earle, S., and Myers, E. B. (**2014**). "Native language interference on the overnight consolidation of a learned nonnative contrast," J. Acoust. Soc. Am. **135**(4), 2352.

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

Jamieson, D. G., and Morosan, D. E. (**1989**). "Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques," Can. J. Psychol. Rev. Can. de Psychol. **43**(1), 88–96.

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.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**(2), 874–886.

Macmillan, N. A., and Creelman, C. D. (**2004**). *Detection Theory: A User's Guide* (Psychology Press, New York).

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (**1995**). "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," Psychol. Rev. **102**(3), 419–457.

Pisoni, D. B. (**1992**). "Some comments on talker normalization in speech perception," in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha Publishing, Tokyo), pp. 143–151.

Rasch, B., and Born, J. (**2013**). "About sleep's role in memory," Physiol. Rev. **93**(2), 681–766.

Stickgold, R., and Walker, M. P. (**2007**). "Sleep-dependent memory consolidation and reconsolidation," Sleep Med. **8**(4), 331–343.

Talamini, L. M., Nieuwenhuis, I. L., Takashima, A., and Jensen, O. (**2008**). "Sleep directly following learning benefits consolidation of spatial associative memory," Learn. Mem. **15**(4), 233–237.

Tamminen, J., Davis, M. H., Merkx, M., and Rastle, K. (**2012**). "The role of memory consolidation in generalization of new linguistic information," Cognition **125**(1), 107–112.

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., and Gaskell, M. G. (**2010**). "Sleep spindle activity is associated with the integration of new memories and existing knowledge," J. Neurosci. **30**(43), 14356–14360.

Tamminen, J., Ralph, M. A. L., and Lewis, P. A. (**2013**). "The role of sleep spindles and slow-wave activity in integrating new information in semantic memory," J. Neurosci. **33**(39), 15376–15381.

Walker, M. P., Brakefield, T., Hobson, J. A., and Stickgold, R. (**2003**). "Dissociable stages of human memory consolidation and reconsolidation," Nature **425**(6958), 616–620.