Brain & Language 167 (2017) 83–93

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Sleep-based memory processing facilitates grammatical generalization: Evidence from targeted memory reactivation

Laura J. Batterink *, Ken A. Paller

Northwestern University, Department of Psychology, 2029 Sheridan Road, Evanston, IL 60208-2710, USA

ARTICLE INFO

Article history:
Received 26 March 2015
Revised 20 August 2015
Accepted 1 September 2015
Available online 9 October 2015

Keywords:
Language acquisition
Learning
Syntax
Generalization
Abstraction
Sleep
Memory consolidation
Targeted memory reactivation

ABSTRACT

Generalization—the ability to abstract regularities from specific examples and apply them to novel instances—is an essential component of language acquisition. Generalization not only depends on exposure to input during wake, but may also improve offline during sleep. Here we examined whether targeted memory reactivation during sleep can influence grammatical generalization. Participants gradually acquired the grammatical rules of an artificial language through an interactive learning procedure. Then, phrases from the language (experimental group) or stimuli from an unrelated task (control group) were covertly presented during an afternoon nap. Compared to control participants, participants re-exposed to the language during sleep showed larger gains in grammatical generalization. Sleep cues produced a bias, not necessarily a pure gain, suggesting that the capacity for memory replay during sleep is limited. We conclude that grammatical generalization was biased by auditory cueing during sleep, and by extension, that sleep likely influences grammatical generalization in general.

1. Introduction

The ability to *generalize* is a key aspect of many basic types of learning, such as motor learning and perceptual learning (e.g., Fenn, Nusbaum, & Margoliash, 2003; Shadmehr & Moussavi, 2000). Generalization involves abstracting regularities from specific examples and applying these regularities to new instances or situations. In contrast to rote learning or to episodic encoding, generalization allows learners to respond adaptively to stimuli that fall outside the original conditions of training. Generalization therefore represents a powerful learning mechanism whereby the learner can transfer acquired knowledge to never-before-experienced stimuli and situations.

Generalization also plays a central role in language acquisition. A hallmark feature of language is that it allows a virtually infinite set of meaningful and grammatically correct utterances to be produced (Hauser, Chomsky, & Fitch, 2002; Pinker & Jackendoff, 2005). Because language is open-ended, language users must be able to generalize common linguistic principles to new combinations of words, rather than relying upon memory of meanings of individual phrases and sentences that they have already heard. This ability to generalize depends upon rules or regularities that are found in virtually every linguistic subsystem, including phonology, morphology, semantics, and syntax. During language acquisition, these overarching linguistic rules or patterns are abstracted over multiple learning episodes, and then applied in order to comprehend and produce novel phrases and sentences. For example, learners of English exposed to a sufficient number of plural nouns will eventually learn that the morpheme *–s* is typically used to denote plurality, and can then apply this rule to novel words. The "Wug Test" is a well-known demonstration of this phenomenon (Berko, 1958). Research using this test has shown that young children are able to correctly produce the plural form of a made-up pseudoword (*wug*), providing evidence that they have extracted generalizable rules from linguistic input, rather than simply memorizing words that they have heard (Menn & Ratner, 2000).

Processes contributing to the generalization of rules from input operate not only during online learning, but during sleep as well. Sleep has been shown to facilitate generalization processes involved in a number of different aspects of language, including speech perception (Fenn et al., 2003), grammar learning (Batterink, Oudiette, Reber, & Paller, 2014; Gómez, Bootzin, & Nadel, 2006; Nieuwenhuis, Folia, Forkstam, Jensen, & Petersson, 2013), and speech production (Gaskell et al., 2014). These experimental results have often implicated generalization above and beyond any improvement in rote or exemplar-based learning. In an artificial grammar learning task, sleep leads to improvement in classification driven specifically by an enhancement of rule

* Corresponding author.
E-mail address: lbatterink@northwestern.edu (L.J. Batterink).

http://dx.doi.org/10.1016/j.bandl.2015.09.003
0093-934X/

abstraction, and not by the strengthening of memory for "chunks," the bigrams and trigrams that make up parts of the presented sequences (Nieuwenhuis et al., 2013). Similarly, infants who were exposed to an artificial language consisting of nonadjacent dependencies and then napped showed greater rule abstraction, whereas infants who remained awake showed improved veridical memory for specific nonadjacent word pairs (Gómez et al., 2006). Sleep also leads to generalization of phonetic constraints in speech production, an effect that is specifically associated with slow-wave sleep (Gaskell et al., 2014). These findings dovetail with numerous results from nonlinguistic tasks demonstrating the importance of sleep for the extraction of overarching rules or patterns (e.g., Djonlagic et al., 2009; Durrant, Cairney, & Lewis, 2013; Durrant, Taylor, Cairney, & Lewis, 2011; Ellenbogen, Hu, Payne, Titone, & Walker, 2007; Wagner, Gais, Haider, Verleger, & Born, 2004). Generalization may be promoted by sleep through simultaneous reactivation of individual memories that share common elements, leading to strengthening of the shared connections (Lewis & Durrant, 2011).

In the present study, we tested whether effects of sleep on rule generalization could be manipulated or enhanced by experimentally inducing reactivations of linguistic patterns during sleep. A series of recent studies has shown that presenting memory cues associated with a prior learning episode during non rapid-eye-movement (NREM) sleep benefits consolidation of both declarative and procedural memories (e.g., Antony, Gobel, OðHare, Reber, & Paller, 2012; Bendor & Wilson, 2012; Diekelmann, Büchel, Born, & Rasch, 2011; Fuentemilla et al., 2013; Rasch, Buchel, Gais, & Born, 2007; Rihm, Diekelmann, Born, & Rasch, 2014; Rudoy, Voss, Westerbrg, & Paller, 2009; Schreiner & Rasch, 2014). For example, re-exposure of an odor during slow-wave sleep that had been previously presented as context during an object-location learning task improved later memory for object locations (Rasch et al., 2007). Individual memories for object-location associations can also be selectively strengthened, when auditory cues associated with individual objects are presented again during sleep (Creery, Oudiette, Antony, & Paller, 2015; Rudoy et al., 2009). Procedural memories also benefit from cueing; presenting a previously learned melody during sleep results in improved performance on a melody production task for the cued relative to the non-cued melody (Antony et al., 2012; Cousins, El-Deredy, Parkes, Hennies, & Lewis, 2014; Schonauer, Geisler, & Gais, 2013). Collectively, these cueing procedures are referred to as *targeted memory reactivation* (TMR; Oudiette & Paller, 2013). Although TMR has been shown to have clear benefits in terms of strengthening associative memories, whether it also results in qualitative changes to memory with improvements in rule abstraction and generalization is unknown.

The goal of the present study was to examine whether TMR influences rule abstraction and generalization in a language-learning context. Participants gradually acquired the grammatical rules of an artificial language through an interactive, trial-and-error-based learning procedure. They also completed a second learning task involving passive exposure to a tone sequence following a probabilistic pattern, which has been previously shown to be sensitive to sleep (Durrant et al., 2011, 2013). By including two learning tasks we hoped to control for nonspecific effects of cueing on consolidation. Each participant was randomly assigned to one of two cueing conditions, involving either the presentation of auditory recordings of the artificial language (grammar-cued condition) or segments of the tone sequence (tone-cued condition). After initial learning, participants took a 90-min nap, during which auditory cues from the selected task were covertly presented during slow-wave sleep. Upon awakening, participants were tested on both learning tasks.

Our central hypothesis was that participants in the grammar-cued condition would show enhanced acquisition of the grammatical rules relative to participants in the tone-cued condition. In addition, we also examined potential mechanisms whereby TMR may influence grammar learning. As laid out by theoretical frameworks of artificial grammar learning (AGL), classification performance on the AGL task can be driven by abstract, rule-based knowledge and by knowledge of chunks (e.g., Knowlton & Squire, 1994, 1996; Lieberman, Chang, Chiao, Bookheimer, & Knowlton, 2004; Meulemans & Van der Linden, 1997). Adopting this reasoning, we examined whether "chunk strength"—the degree of superficial similarity between training items and test items (Knowlton & Squire, 1996)—interacts with cueing improvements. Given previous evidence that sleep specifically benefits the abstraction of grammar rules without enhancing the effect of chunk knowledge (Nieuwenhuis et al., 2013), we hypothesized that TMR would primarily enhance rule knowledge. Finally, we tested whether oscillatory and spindle activity during sleep predicts cueing-related gains in grammar acquisition by examining correlations between sleep physiology and behavioral improvements on the grammar task.

2. Methods

2.1. Participants

We recruited 44 participants from the university community (30 female; mean age = 22.4 years) for this study. Participants were randomly assigned to one of two sleep-cueing conditions (grammar-cued condition versus tone-cued condition). Of the 44 participants, 35 were successfully cued, 17 in the tone condition and 18 in the grammar condition. The 9 remaining participants were not successfully cued, either because insufficient slow-wave sleep (SWS) prevented cueing from being attempted ($n = 3$), or because cueing attempts resulted in arousals ($n = 6$).

2.2. Artificial language task

2.2.1. Stimuli

The artificial language was composed of 20 monosyllabic nonsense words (e.g., *pilk*). Sixteen of these words were taken from previous artificial language studies (Saffran, 2001, 2002). Each of the nonsense words was assigned to one of six categories (denoted here by A–F), with each category containing 2, 3, or 4 different words (Table 1). An artificial grammar was created using five rules:

Rule 1: A → B → C
Rule 2: A → D → B → C
Rule 3: A → B → C → E
Rule 4: A → D → B → C → E
Rule 5: A → B → F → C

This artificial grammar was designed to be characteristic of natural languages and contained both optional elements and predictive dependencies between word categories. For example, D is an

Table 1
Word categories from the artificial language.

Category				
A	biff	hep	mib	rud
B	cav	lum	neb	sig
C	dupp	jux	loke	vot
D	klor	pell		
E	pilk	tiz	gak	
F	tood	kice	zic	

optional element, but if present, it must follow an A. The inclusion of these optional elements adds an additional level of complexity to the artificial grammar and precludes the rote memorization of pairs of words from being an effective learning strategy. For example, learning all possible word combinations contained in the word pair A → B would facilitate the acquisition of rules 1, 3, and 5, but would hinder acquisition of rules 2 and 4. In addition, because each category was represented by 2–4 words, rote memorization of entire phrases would not on its own allow learners to generalize the grammar to novel phrases. Because participants were not provided with explicit representations of the grammatical rules, the process of learning the underlying grammar was designed to mimic rule abstraction processes that would normally occur during exposure to an unfamiliar language.

For the learning phase, we created a set of 56 phrases according to the five underlying grammatical rules. For example, a valid phrase following rule 1 would be composed of a word from category A, followed by a word from category B, followed by a word from category C (e.g., biff → cav → dupp). Each rule was represented either 8 times (rules 1, 2, and 3) or 16 times (rules 4 and 5). Rules 1–3 were embedded within rules 4 and 5, and thus rules 4 and 5 were represented more frequently. A trained male native-English speaker produced auditory recordings of these phrases, speaking at a rate of approximately 2 words/s and using natural intonation. Auditory recordings of individual words were also created by splicing individual word tokens from the longer phrase recordings. These auditory recordings were embedded in the learning paradigm, as described in detail below.

For the test phase, we created a set of 54 phrases that were not used in the learning phase. The test phrases were composed of the same 20 nonsense words and followed the same five grammatical rules as the original set of 56 phrases. All 5 rules were represented in the test phase. However, in order to capture a range of different chunk strengths (as described in greater detail below), some rules were represented more frequently than others in the test phrase pool (range: 6–18 total phrases per rule).

Auditory cueing stimuli (presented during the nap period) consisted of the same 56 auditory recordings that were presented during the original learning phase. Cueing stimuli consisted of entire phrases; single elements were not presented during sleep.

2.2.2. Learning and testing procedures

Participants were not given any explicit information about the structure of the artificial grammar, but gradually acquired the grammatical rules over time by completing an interactive learning paradigm. Before beginning the task, they were instructed that their goal was to learn to create phrases following an artificial grammar, and that they would accomplish this learning through trial and error, with feedback given for each response. On each trial, words from a phrase in the artificial language were presented visually, in a vertical arrangement and in scrambled order (Fig. 1). Participants indicated the order for each phrase by clicking on each word one at a time using a computer mouse. When an incorrect response was made, an auditory error signal was presented and the visual representation of the word remained in its current location. When a correct response was made, an auditory recording of the word was played, and the visual representation of the word moved to the correct location at the bottom of the screen. Once all words of the phrase were successfully ordered, the entire phrase was presented auditorily before the next trial began. The learning set consisted of 56 valid phrases. Participants performed this task until they successfully met a predetermined two-tier accuracy criterion, as follows. A computer algorithm calculated the number of errors made over the last 10 phrases (on every trial beginning with the 10th trial). If fewer than the full set of 56 phrases had been presented, the task continued until 5 or fewer errors had been made over the last 10 phrases. Thirty-two participants (73%) reached the learning criterion after 56 phrases or less. For the other participants, phrases were recycled and the task continued until 10 or fewer errors had been made over the last 10 phrases, representing an easier level of accuracy to attain. Overall, participants required an average of 48.3 phrases (SD = 25.9) to reach the accuracy criterion, making an average of 73.0 incorrect responses (SD = 62.6). Individual learning rates were highly variable, with the number of total training phrases required to reach criterion ranging from 15 to 143. We verified after testing that all participants had been exposed multiple times to all 5 rules, taking into account that rules 1–3 were embedded within rules 4–5.

The test was similar to the initial learning task, except that no feedback was provided. The test consisted of 54 novel phrases. Again, the visual representation of each word was shown in a vertical arrangement and in scrambled order. Participants were

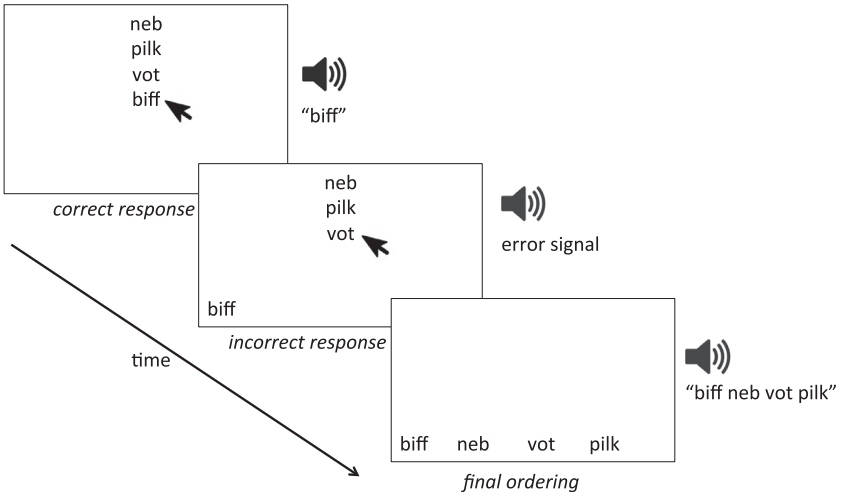

Fig. 1. Schematic diagram of artificial language learning task. Participants learned to correctly order phrases according to abstract hidden grammatical rules through an interactive, trial-and-error based learning procedure. Participants selected individual words one at a time and were given auditory and visual feedback on whether their choice was correct or not. For a correct choice, the word moved to the bottom of the screen and the spoken word was presented out loud. Auditory recordings of the phrases were embedded in the learning paradigm and presented later during sleep. Examples of both a correct and incorrect response are shown.

instructed to reorder the words according to the rules that they had acquired during the learning phase. After completing each phrase, they provided a confidence rating on a 3-point scale, indicating how confident they were that they had ordered the words correctly. They were instructed that the lowest level of confidence indicated a guess response. Confidence data was not collected for the first four participants.

Half of the participants (n = 22) completed two test blocks: one immediately after the initial learning phase (Pre-Nap test), and a second after the nap period (Post-Nap test). For these participants, the Post-Nap test was composed of novel phrases but was otherwise identical to the Pre-Nap test. Given that interpretations might be complicated due to any incorrect responses made during the Pre-Nap test, such that participants might consolidate those incorrect grammatical representations rather than correct ones, we ran the other participants (n = 22) with a Post-Nap test but no Pre-Nap test. The effect of the Pre-Nap test was evaluated using this between-subjects factor (Pre-Nap test: present, absent), and in all analyses this effect was nonsignificant, so our main analyses collapsed between these two groups, focusing on performance during the Post-Nap test.

2.3. Probabilistic tone task

2.3.1. Stimuli

The stimuli in this task were designed to replicate those used by Durrant et al. (2011). As in that prior study, five pure sinusoidal tones were created, using frequencies taken from the Bohlen–Pierce scale (261.63 Hz, 300.53 Hz, 345.22 Hz, 396.55 Hz, 455.52 Hz). Each tone had a duration of 200 ms and was Gaussian-modulated to avoid edge effects. The learning sequence consisted of a total of 1818 tones separated by 20-ms gaps. The order of the tones followed a probabilistic pattern, determined by a transition matrix in which the identity of the next tone is predicted by the previous two tones, forming a second-order Markov chain. The probability of a second-order transition was 90%. Therefore, a given pair of tones was followed by a particular third tone 90% of the time, and 10% of the time by one of the four other tones (p = 2.5% for each low-probability tone). Zero- and first-order transitions were fully balanced, ensuring that they could not provide additional structural information.

A total of 168 test sequences were created, each consisting of 18 tones. Half of the test sequences followed a random order, with an equal probability for each tone at every position in the sequence. The other half followed the same probabilistic pattern as the learning sequence. Structured sequences were generated by randomly sampling the transition matrix. The number of likely second-order transitions in the test sequences, as defined above, ranged between 8 and 16 across the 18 tones (mean = 12.6).

Auditory cueing stimuli (presented during the nap period) consisted of streams that followed the same probabilistic pattern as the original learning sequence, with the exception that the likely-transition probability of the transition matrix was increased to 96%. The goal of increasing the likely-transition probability was to reduce probabilistic error and emphasize the underlying stimulus structure, thereby facilitating the extraction of this structure by the sleeping brain. Eliminating probabilistic error entirely (i.e., setting the likely transition probability to 100%) was not feasible, as this resulted in a shortened sequence loop in which only a subset of possible transitions were presented. Each cued sequence consisted of 35 tones.

2.3.2. Learning and testing procedures

The learning task involved presentation of the auditory learning sequence for approximately 7 min. Participants were instructed to listen carefully to the tones and that their memory for the tones would be subsequently tested. After exposure to the stream, participants completed a forced-choice recognition test composed of 84 two-alternative forced-choice trials. Each trial consisted of two short sequences, one structured and one random. On each trial, participants indicated which sequence sounded more familiar and provided a confidence rating on a 3-point scale, with the lowest level of confidence indicating a guess. Responses were made without time pressure. After the nap period, participants completed a second recognition test composed of novel sequences but otherwise identical to the first test. Test order was counterbalanced across participants. Because the current paper focuses on grammar learning, results from the probabilistic tone task are not reported.

2.4. Procedure

The experimental session began between 11:00 AM and 3:00 PM with electrode application for ERP analysis and standard sleep EEG recording. After electrode application, participants completed the two experimental learning tasks (the artificial language task and the probabilistic tone task, run in counterbalanced order across participants). After the second learning task, a short story was presented for approximately 3 min. The goal of presenting this story was to reduce order effects and any recency benefits that might have been associated with performing the second task immediately prior to the nap period.

Participants then reclined in a quiet, darkened room to sleep. Low-intensity white noise at ~40 dB was present for the duration of the sleep period to dampen the influence of possible noise from outside the room and to embed the cues. Sound presentation began once indications of slow-wave sleep were observed. Stimulation was paused if signs of arousal were observed, and was restarted only if a stable pattern of slow-wave sleep re-emerged. For participants in the grammar-cued condition, the 56 phrases were cycled through twice, for a maximum of 112 phrases total. Phrases were separated by 12–14 s. The mean number of phrases presented during sleep was 92 (*minimum* = 51, *SD* = 23.5). Duration of the phrases varied from 2 to 3.5 s, depending on the length of the phrase. For participants in the tone-cued condition, a maximum of 60 streams was presented during the nap period, with each stream separated by 8–10 s. The mean number of tone streams presented was 51 (*minimum* = 36, *SD* = 7.8). Duration of the tone sequences was approximately 8 s. Thus, maximum auditory stimulation time was approximately 8 min in the tone condition and approximately 5.5 min in the grammar condition. Cueing parameters for each task were designed to equate auditory stimulation between conditions, considering both overall auditory stimulation time and number of auditory reactivation events, insofar as that was possible given the differences between the two tasks. Sound presentation was designed to be covert. Participants were not told that auditory cues would be played during their naps, and—as described in more detail in Results—the vast majority of participants did not report hearing any sounds other than white noise during the nap period.

The nap period ended after 90 min, but was extended if the participant was still in slow-wave sleep. After awakening, participants were given a 10-min break before completing Post-Nap tests for both the artificial language task and the probabilistic tone task, run in the same order as the Pre-Nap learning tasks. At the end of the experiment, participants were asked to describe any patterns or rules that they had noticed in the language. They were also asked whether they had heard any sounds presented during their nap. To assess subjective sleepiness levels, sleepiness ratings were collected from a subset of participants (n = 31) using the Stanford Sleepiness Scale at two time points: once at the beginning of the session, and again after the nap period prior to the Post-Nap tests.

2.5. EEG recording and analysis

EEG was recorded from 21 tin electrodes mounted in an elastic cap, along with two electrooculogram (EOG) channels and one chin electromyogram (EMG) channel, using a 250-Hz sampling rate. EEG was recorded throughout both learning blocks and over the nap period.

For sleep analyses, data from EEG and EOG channels were filtered with a bandpass from 0.5 to 30 Hz, and EMG data were filtered from 10 to 62 Hz. Sleep staging was conducted offline using standard criteria recommended by the American Academy of Sleep Medicine, with 30-s EEG epochs scored as corresponding to wake, Stage 1, 2, or 3, or REM. SWS was defined as Stage 3, NREM sleep was defined as Stages 1–3, and sleep was defined as any epoch not staged as wake. EEG spectral analyses were conducted following artifact removal based on visual inspection. Time–frequency decompositions were computed using fast Fourier transform with a Hamming window over 5-s epochs. As an overall measure of sleep quality, we computed delta power (1–4 Hz) across epochs of NREM sleep at electrode Fz, where delta power is typically maximal (Grigg-Damberger, 2012).

Spindles are bursts of EEG oscillations at 12–15 Hz lasting 0.5–2.0 s that are possibly related to consolidation (Nir et al., 2011; Nishida & Walker, 2007). Spindles were quantified using a MATLAB/EEGLAB algorithm (Ferrarelli et al., 2007; Mander, Santhanam, Saletin, & Walker, 2011). The algorithm identifies amplitude fluctuations exceeding threshold values, with the lower and upper values set at two and eight times the average amplitude, respectively. The peak amplitude for each spindle is defined as the local maximum above the threshold, with the beginning and end of the spindle defined as points immediately preceding or following this peak, when the amplitude of the time series dropped below the cut-off threshold. Prior to spindle analysis, the EEG signal was filtered at 11–16 Hz and then subjected to artifact rejection by visual inspection. The algorithm-determined spindles were restricted only to those events falling within this frequency range, subsequently classified as either fast (13.5–16 Hz) or slow (11–13.5 Hz), derived from a median split of the frequency range (11–16 Hz), similar to previously reported fast and slow spindle analyses (Knoblauch, Martens, Wirz-Justice, & Cajochen, 2003; Milner, Fogel, & Cote, 2006; cf. Mander et al., 2011). This analysis yielded two main spindle measures, corresponding to fast and slow spindle density. Fast spindle density was computed at electrode Pz and slow spindle density was computed at electrode Fz, as fast spindles predominate over centro-parietal areas whereas slow spindles predominant over frontal areas (e.g., De Gennaro & Ferrara, 2003; Schabus et al., 2007; Urakami, 2008).

2.6. Behavioral data analysis

Performance during the artificial language learning task was quantified for each participant by calculating (a) the total number of phrases required before the participant reached the accuracy criterion and (b) the total number of incorrect responses made during the learning task. Performance during the Post-Nap test was quantified for each participant by calculating the percentage of phrases that were ordered correctly (out of a total of 54 phrases).

Before investigating whether cueing significantly influenced performance on the Post-Nap test, we examined whether performance during the learning task significantly predicted performance during the Post-Nap test across all participants ($n = 44$). The goal of this analysis was to quantify and reduce possible effects of inter-subject baseline variability in learning (as shown to be critical in the prior TMR study reported by Creery et al., 2015). A regression model was conducted with Post-Nap performance as the dependent variable and three predictor variables: (a) total number of learning phrases, (b) total number of incorrect responses during learning and (c) Pre-Nap test (present, absent). Pre-Nap test was not found to be a significant predictor and was subsequently dropped as a factor [$F(1,40) = 0.14$, $p = 0.71$]. The final model significantly predicted Post-Nap performance across all participants [$F(1,41) = 8.16$, $p = 0.001$]. Better Post-Nap performance was associated with fewer phrases to reach criterion and fewer errors, indicating that participants who more quickly learned the underlying grammatical rules and made fewer errors during the learning task performed better on the subsequent test. The results of this analysis confirm that variability in Pre-Nap learning performance systematically contributed to variability in Post-Nap test performance. Based on this analysis, we subtracted each participant's predicted Post-Nap performance (as estimated by the final regression model) from his or her actual observed Post-Nap performance. This difference, which we term the *Accuracy Residual*, reflects whether participants over- or under-performed compared to what would be predicted based upon their initial learning performance.

The Accuracy Residual was used to examine our main experimental question—whether cueing condition significantly influenced Post-Nap performance. A regression model was initially conducted with the Accuracy Residual as the dependent variable, and with cue condition (tone-cued, grammar-cued), Pre-Nap test (present, absent), and Task Order (grammar task first, tone task first) as predictors. Pre-Nap test [$F(1,31) = 0.70$, $p = 0.41$] and Task Order [$F(1,31) = 0.087$, $p = 0.77$] were not significant predictors and were subsequently dropped from the model. We hypothesized that participants in the grammar-cued condition should show significantly higher Accuracy Residual values than participants in the tone-cued condition, indicating that grammar-cued participants performed better on the Post-Nap test than would be expected based upon initial learning. Note that our Accuracy Residual approach is essentially equivalent to conducting an ANOVA with baseline learning performance variables (total number of phrases, number of incorrect responses) entered as covariates (Analysis of Covariance, or ANCOVA), but allows us to include data from our entire pool of participants ($N = 44$) in our estimate of the effect of baseline learning on Post-Nap performance, rather than only cued participants ($n = 35$). Thus, this approach provides more statistical power compared to an ANCOVA with cued participants ($n = 35$).

In addition, we tested whether cueing especially benefitted performance on either high-chunk or low-chunk test items in order to examine the underlying mechanisms contributing to potential performance improvements. Chunk strength was calculated according to previous procedures (e.g., Knowlton & Squire, 1996; Lieberman et al., 2004). Chunks were defined as the bigrams and trigrams that appeared in each test item. First, the overall frequency of each bigram and trigram across all the training items was calculated, yielding an "associative strength" value for each chunk. For example, if the bigram "biff cav" appeared in two different training phrases (e.g., *biff cav dupp* and *biff cav tood jux*), this bigram would have an associative strength of 2. Next, the chunk strength of each test item was calculated by averaging the associative strength for each chunk that occurred in the test item. For example, the test item "biff lum vot" contains the chunks "biff lum", "lum vot", and "biff lum vot." If these chunks had associative strengths of 2, 1 and 0 respectively, the chunk strength of this test item would be computed as $(2 + 1 + 0)/3 = 1$. Because the number and identity of phrases presented as part of the training set were not constant across participants (due to individual variability in reaching the learning criterion), these calculations were performed at an individual level for each participant. For each participant, the top 50% of test phrases with the highest chunk strength were designated as high-chunk-strength items, while the 50% of test phrases with the lowest chunk strength were designated as low-chunk-strength

items. In order to examine whether learning is influenced by chunk strength (cf. Knowlton & Squire, 1996), we intentionally selected test phrases with a range of chunk strengths (range = 0–3.4 when all 56 learning phrases were presented once during training).

Accuracy Residuals were calculated separately for both low-chunk and high-chunk items, using the same procedure described previously for all items. A repeated-measures ANOVA was then conducted with the Accuracy $Residual_{high\ chunk}$ and Accuracy $Residual_{low\ chunk}$ as the dependent variables, chunk strength (high, low) as a within-subjects factor, and Pre-Nap test (present, absent), Task Order (grammar task first, tone task first), and cue condition (grammar-cued, tone-cued) as between-subjects factors. Again, Pre-Nap test [Pre-Nap test × Chunk Strength: $F(1,27) = 1.40$, $p = 0.25$] and Task Order [Task Order × Chunk Strength: $F(1,27) = 0.42$, $p = 0.52$] were not significant predictors and were subsequently dropped from the model. If TMR specifically enhances abstraction of grammar rules, no effect of chunk strength would be expected, indicating that all items benefitted approximately equally from cueing. In contrast, if TMR specifically enhances knowledge of parts of the presented phrases (i.e., "chunk" knowledge), we would expect a significant effect of chunk strength, with high-chunk-strength items benefitting more from cueing than low-chunk-strength items.

To examine whether explicit or implicit knowledge contributed to performance on the task, we quantified the effect of confidence on accuracy by conducting a repeated-measures ANOVA with confidence (high, medium, guess) as a within-subjects factor.

The Accuracy Residual was also used to examine possible correlations between grammatical performance and sleep physiology. The following sleep physiology measures were included: overall percentage of time asleep, percentage of time in each sleep stage (Stage 1, Stage 2, SWS, REM), delta power (during epochs of SWS, NREM, and across the entire nap), fast spindle density, and slow spindle density. These analyses were designed to examine whether a larger gain in grammatical knowledge from pre- to post-nap (as indicated by a larger Accuracy Residual) was associated with certain aspects of sleep previously linked to memory consolidation.

3. Results

3.1. Overall behavioral performance

Across all participants, 47.6% ($SD = 18.9\%$) of phrases on the Post-Nap test were ordered correctly. Given that only 4.9% of phrases would be ordered correctly by chance, this represents a highly significant learning effect [$t(43) = 15.0$, $p < 0.0001$]. Overall—that is, independently of cueing condition—performance on the grammar test did not change over the sleep interval but was stable from Pre- to Post-Nap. Within the subset of participants who completed both the Pre-Nap and Post-Nap tests and who were cued successfully ($n = 17$, both cueing conditions), 44.8% of phrases were ordered correctly at Pre-Nap and 46.0% at Post-Nap [Pre- to Post-Nap difference: $F(1,16) = 0.56$, $p = 0.46$]. Individual performance in the Pre-Nap and Post-Nap tests was highly correlated ($r = 0.93$, $p < 0.001$).

3.2. Effects of cueing on behavioral performance

Consistent with our main hypothesis, participants in the grammar-cued condition showed significantly higher Accuracy Residual values than participants in the tone-cued condition [$F(1,33) = 4.83$, $p = 0.035$]. This result reflects the finding that the Post-Nap test performance of participants in the grammar-cued condition was higher than expected based upon their Pre-Nap learning scores, whereas performance on the Post-Nap test of participants in the tone-cued condition was lower than expected (Fig. 2). By controlling for individual differences in Pre-Nap learning, the Accuracy Residual represents a more powerful test of cueing benefits relative to merely comparing Post-Nap test performance between the two groups.

Learning performance was not significantly different between the two groups, as indicated by both the total number of phrases to reach criterion [grammar-cued group: 51.2, $SD = 33.1$; tone-cued group: 45.8, $SD = 23.3$; $t(33) = 0.56$, $p = 0.58$] as well as the total number of incorrect responses made during learning [grammar-cued group: 85.1, $SD = 82.7$; tone-cued group: 64.9, $SD = 50.2$; $t(33) = 0.86$, $p = 0.39$]. Predicted Post-Nap performance [$t(33) = 1.20$, $p = 0.24$] and actual Post-Nap performance [$t(33) = 1.07$, $p = 0.29$] were both not significantly different between the two groups, but there was a significant interaction between cue condition and the difference between predicted/actual performance, as reflected by the Accuracy Residual computation above [$F(1,33) = 4.83$, $p = 0.035$]. The learning parameters (trials to criterion and errors made) yielded the prediction that Post-Nap performance would be less accurate for the grammar-cued group than the tone-cued group, cueing effects aside; the cueing effect is best evaluated in this context by considering the Accuracy Residual effect, as shown in Fig. 3.

Tests of simple effects indicated that in the tone-cued group, Post-Nap test performance was significantly lower than predicted based on Pre-Nap learning performance [$t(16) = -2.14$, $p = 0.048$]. In the grammar-cued group, Post-Nap test performance was higher than predicted, but not significantly so [$t(17) = 1.25$, $p = 0.23$]. It would be interesting to know whether the tone cues impaired generalization or whether the grammar cues facilitated generalization over the course of the nap period (or whether both mechanisms occurred). However, with the current design this issue cannot be conclusively evaluated, as we did not collect data to evaluate how performance would change merely with the passage of time without sleep, or with sleep but no cueing.

In a separate analysis, we examined the Accuracy Residual in the subset of participants who were not successfully cued ($n = 9$), but with the caution that this group was not randomly assigned to a condition of sleep without cues (rather, these participants either awoke during cueing or failed to achieve SWS and so were not subjected to cues). The Accuracy Residual in this "no-cues" group was close to zero, indicating that Post-Nap performance was very similar to the level expected based on Pre-Nap learning, and fell between values for the grammar-cued and tone-cued

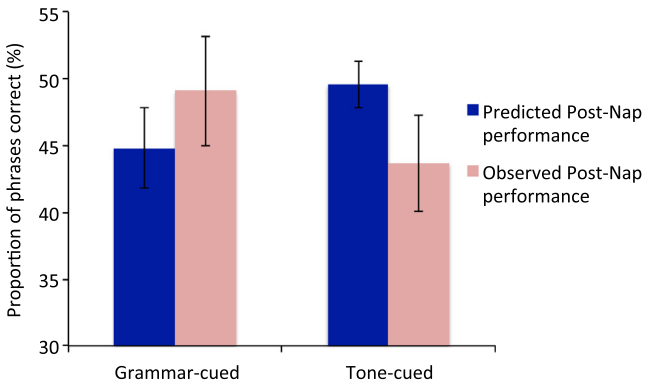

Fig. 2. Data contributing to the Accuracy Residual. The proportion of phrases predicted to be correct based on Pre-Nap learning performance is shown in dark blue for each group. Observed Post-Nap performance is shown in light red. A significant interaction was found, indicating that participants in the grammar-cued condition showed significantly higher Accuracy Residual values than did participants in the tone-cued condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

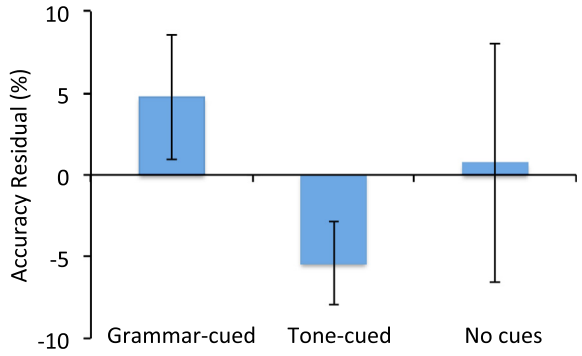

Fig. 3. The Accuracy Residual in the grammar-cued group ($n = 18$) and the tone-cued group ($n = 17$), as well as participants who were not successfully cued due to low sleep quality, the no-cues group ($n = 9$). The Accuracy Residual was computed as the difference between each participant's predicted Post-Nap performance, as estimated based upon performance during the Pre-Nap learning task, and his or her observed Post-Nap performance. A larger Accuracy Residual indicates that the participant performed better at the Post-Nap test than would be expected based upon Pre-Nap learning.

groups (Fig. 3). There was no significant difference in Accuracy Residual between participants in the "no-cues" group and either the grammar-cued group [$t(41) = 0.63$, $p = 0.53$] or tone-cued group [$t(41) = 0.95$, $p = 0.35$]. Although this analysis provided some indication of the Accuracy Residual in participants who received no or very few cues during sleep, participants in this group also differed from the other groups in systematic ways, as their sleep quality was generally poorer.

We also analyzed cueing effects in the subset of participants who were tested on the grammar both pre- and post-nap ($n = 9$ in the grammar-cued condition; $n = 8$ in the tone-cued condition) by directly comparing performance on the Pre- and Post-Nap tests. No significant effect of cueing was found in this subset of participants [Cue Condition × Pre-Nap/Post-Nap: $F(1,15) = 0.26$, $p = 0.62$]. If cueing effects are generally somewhat small in magnitude, they may not be detectable with such a small sample.

3.3. Effects of chunk strength on behavioral performance

Next, we examined the effect of chunk strength on ordering performance. Across all participants, there was no significant effect of chunk strength [$F(1,43) = 0.23$, $p = 0.64$]. High-chunk-strength items were ordered correctly at a rate of 48.2% ($SD = 21.3\%$), compared to 47.1% ($SD = 18.9\%$) for low-chunk-strength items. As a more sensitive test of the effect of chunk strength on performance, we computed the correlation between chunk strength and accuracy at the individual level. The average of these correlations across participants was not significantly greater than 0, again indicating that there was no significant effect of chunk strength on ordering performance [mean correlation coefficient (r) = −0.0093, $SD = 0.173$; $t(42) = -0.35$, $p = 0.73$]. In addition, we examined whether ordering performance was significantly greater than chance for items that had a chunk strength of 0. Because these items are composed of bigrams and trigrams that had never been presented in training, the ability to correctly construct these phrases cannot be even partially supported by knowledge of chunks without generalization. Only participants who were tested on 5 or more zero-chunk-strength items were included in this analysis ($n = 10$). Performance for these zero-chunk-strength items was significantly greater than the chance level of 4.9% accuracy [mean proportion correct = 53.4%, $SD = 22.2\%$, $t(9) = 6.90$, $p < 0.001$], and was similar to overall performance for all items. Within this subset of participants, there was no significant difference between performance for high chunk and zero-chunk-strength items [mean proportion correct for high chunk items: 48.5%, $SD = 18.6\%$; $F(1,9) = 0.56$, $p = 0.47$].

We also examined whether there was a specific benefit of cueing for high-chunk or low-chunk items. No significant effect of chunk strength was found on cueing-related gains [$F(1,33) = 0.003$, $p = 0.96$]. In other words, the benefit associated with being exposed to grammar cues during sleep extended to both high-chunk and low-chunk items at approximately equal levels.

3.4. Effects of subjective confidence on behavioral performance

Among the subset of participants who completed confidence judgments ($n = 40$), a significant effect of confidence on accuracy was found [$F(2,62) = 109.1$, $p < 0.001$; linear effect of confidence: $F(1,31) = 154.4$, $p < 0.001$]. Not surprisingly, participants were most accurate when they expressed high levels of confidence [mean = 67.5%, $SD = 22.8\%$], showed a medium level of accuracy when they expressed moderate levels of confidence [39.4%, $SD = 18.0\%$], and were least accurate when they reported to be guessing [*mean* = 17.4%, $SD = 20.0\%$]. Relatively few guess responses were made. Among those participants who made at least one guess response, average proportion of guesses was 13.0% ($SD = 16.6\%$); 7 participants made no guess responses at all. Accuracy for guess responses was significantly above chance [$t(32) = 3.64$, $p = 0.001$].

No significant effect of confidence was found on cueing-related improvements [$F(1,26) = 0.22$, $p = 0.64$]. In other words, accuracy for guesses, medium-confident, and high-confident responses all benefitted from TMR.

3.5. Sleep physiology

Sleep physiology measures for grammar-cued and tone-cued participants are shown in Table 2. No significant differences in sleep physiology between the two groups were found on any measure (overall sleep duration, duration on Stage 1, Stage 2, SWS, and REM, delta power during NREM sleep, and slow and fast spindle density during NREM sleep; all p values >0.19).

3.6. Self-reported measures

Participants who were successfully cued and who provided both Pre- and Post-Nap ratings on the Stanford Sleepiness Scale reported feeling marginally more alert after the nap period compared to before (mean Pre-Nap rating = 3.2, $SD = 1.1$; mean Post-Nap rating = 2.7, $SD = 1.1$; $F(1,23) = 4.05$, $p = 0.056$; lower values indicate higher levels of alertness). The change in sleepiness levels from pre-nap to post-nap did not differ significantly between the two groups (grammar-cued group: 0.46 decrease; tone-cued group: 0.63 decrease; $t(22) = 0.32$, $p = 0.75$). This result indicates that although testing was carried out relatively soon after the nap period ended, sleep inertia effects are not a major concern.

After the nap period, participants were asked to estimate how long they slept, how long it took them to fall asleep, the number of awakenings they had, and the length of time they were awake after falling asleep. They also gave ratings on 1–5 scale for the following measures: how well they slept, how refreshed they felt after waking, whether they slept soundly or restlessly, whether they slept throughout the allocated nap time, how easy it was for them to wake up, and how easy it was for them to fall asleep. None of these self-report sleep measures differed significantly between the two groups (all p values >0.12).

Table 2
Sleep physiology measures by group.

	Time awake	Overall time asleep	Time in S1	Time in S2	Time in SWS	Time in REM	Delta power NREM (Fz) (μV^2)	Slow spindle density NREM (Fz) (#/min)	Fast spindle density NREM (Pz) #/min)
Grammar-cued	17.9 (13.2)	73.8 (13.5)	5.5 (3.3)	32.1 (7.6)	29.3 (13.4)	7.0 (8.1)	147.5 (94.6)	3.4 (0.85)	2.8 (0.89)
Tone-cued	19.3 (11.7)	73.3 (8.61)	5.5 (4.8)	35.6 (14.2)	23.3 (12.9)	8.9 (8.0)	114.5 (57.9)	3.6 (0.68)	2.7 (1.2)
p	0.75	0.90	0.97	0.37	0.19	0.47	0.22	0.40	0.84

Values are min ± *SD* unless stated otherwise. *p* values are uncorrected for multiple comparisons.

3.7. Sleep physiology correlations

Across all participants who were presented with cues during sleep (grammar-cued and tone-cued conditions combined; $n = 35$), a marginal correlation was found between the Accuracy Residual and overall time asleep ($r = 0.32$, $p = 0.063$). This result indicates that participants who slept more tended to show better Post-Nap performance than would be expected based on their Pre-Nap learning performance, providing a suggestion that behavior on this task may be sleep-sensitive. No other sleep physiology measure (duration of Stage 1, Stage 2, SWS, and REM sleep, delta power during NREM sleep, slow spindle density, and fast spindle density) correlated significantly with the Accuracy Residual (all *p* values >0.13).

3.8. Questionnaire data

When participants were asked about rules or patterns in the artificial language, they generally described only a small level of the overall statistical structure. For example, participants generally described noticing that specific words came first or last within the phrase (e.g., "*biff* and *hep* came first"), or noticing that there was a hierarchy between individual words within the phrase. Of the 44 participants, only 5 mentioned learning particular word pairs or longer word combinations (e.g., "*cav* and *vot* went together," or "I noticed 3 specific phrases that started every sentence").

When questioned whether they had heard any sounds during their nap, none of the participants in the grammar-cued condition reported hearing any cue-related sounds. Three participants in the tone-cued condition reported hearing 1–2 sequences of tones at some point during their nap, often with low confidence, likely reflecting brief arousals that occurred during the time of stimulus presentation. However, given that 50–60 sequences were actually presented, even these three participants remained unaware of the vast majority of stimuli presented during sleep.

4. Discussion

4.1. Cueing during sleep influences grammar learning

Our findings provide evidence that auditory cueing during sleep can influence grammatical rule learning and generalization. By extension, these results suggest that spontaneous memory processing during sleep may generally be useful for learning and generalizing grammatical rules. Participants who were presented with auditory recordings of the artificial language while napping showed significantly greater gains on grammatical rule generalization compared to those who were presented with auditory cues associated with an unrelated control task. This effect could not be attributed to gross differences in sleep physiology, as both participant groups showed very similar sleep architecture based on standard EEG signals. This result extends previous demonstrations that sleep facilitates rule generalization (Djonlagic et al., 2009; Durrant et al., 2011, 2013; Ellenbogen et al., 2007; Fenn et al., 2003; Gaskell et al., 2014; Gómez et al., 2006; Nieuwenhuis et al., 2013; Wagner et al., 2004), and further demonstrates that rule-abstraction mechanisms can be influenced by sensory stimulation during sleep.

As cited above, the major evidence implicating sleep in generalization comes from comparisons between sleep and wake during the retention interval. In studies of this sort, greater interference during the wake condition could influence the results and contribute to an apparent benefit for sleep, thus limiting the strength of the conclusions that can be drawn. The present TMR design avoids this potential confound with interference, as well as with differential levels of alertness due to intervening sleep, providing stronger evidence that sleep plays a role in generalization and abstraction.

Sleep-related enhancements in rule generalization have been attributed to replay of separate but overlapping item memories (Lewis & Durrant, 2011; Stickgold & Walker, 2013). This overlapping reactivation strengthens the interconnections between multiple memories, creating a schema in which shared memory components are most strongly represented (Lewis & Durrant, 2011). This schema formation enables the extraction of relationships or rules governing the item set, as well as generalization to novel items (Stickgold & Walker, 2013). TMR procedures may promote rule generalization by artificially triggering the simultaneous reactivation of multiple item memories at levels above those that occur spontaneously during normal sleep. Presenting the artificial language during sleep may have acted as a cue, eliciting overlapping reactivation of individual phrases presented during the prior learning episode. This reactivation could then lead to the extraction and generalization of underlying grammatical rules.

It is also worth considering the alternative possibility that TMR effects on generalization were merely a consequence of increased exposure to the artificial language. Phrases presented during sleep might be processed and encoded into memory, perhaps providing additional opportunities for consolidation and generalization without necessarily reactivating previously learned material. However, there is no evidence that any episodic encoding occurred during sleep. Moreover, some doubts about this second possibility come from the principle that new learning does not occur for information presented only during sleep (e.g., Emmons & Simon, 1956; Simon & Emmons, 1956), although classical conditioning appears to be an exception (Arzi et al., 2012), so other types of implicit learning may constitute analogous exceptions that deserve further investigation.

In any event, our results demonstrate that presenting task-specific cues during sleep leads to a relative improvement in expected grammatical performance compared to presenting unrelated cues. However, with our current design we cannot conclusively determine whether the difference between cueing conditions was driven by an active enhancement in grammatical consolidation from the grammar cues, interference in grammatical consolidation from the unrelated tone cues, or a combination of both mechanisms. Although a definitive answer to this question awaits further study, we tentatively hypothesize that both mechanisms were at play. In tone-cued participants, some spontaneous reactivations of the grammar during SWS may have been

prevented by the presentation of the tone stimuli, consistent with the general idea that there is a limited capacity for memory reactivation during sleep (Oudiette & Paller, 2013).

This speculation is supported by at least three prior TMR studies. First, Cousins et al. (2014) trained participants on two different sequences of button presses. In an experimental group, participants were cued on one of the two sequences during sleep, whereas in a separate control group participants did not receive any cues during sleep. Both procedural and explicit performance measures for the control group fell midway between levels observed for the cued and uncued sequences in the experimental group, suggesting that the cues biased memory processing in favor of the cued sequence at the expense of the uncued sequence. A similar pattern of results was observed by Antony et al. (2012), who trained participants on two different melodies, including both an experimental group who was cued on one melody during sleep and a control group who received no cues during sleep. Again, although differences were not significant, the control group showed a level of improvement in performance that fell halfway between the cued and uncued conditions in the experimental group. Bendor and Wilson (2012) trained rats to associate two auditory signals with two different running directions (left versus right). Presenting one of the auditory cues during SWS caused hippocampal place cells representing the cued spatial direction to fire more frequently than place cells representing the uncued spatial direction. Notably, the total number of replay events elicited by control cues not associated with a particular running direction was similar to that elicited by task-related cues, indicating that task-related cues only biased the content of replay events, rather than increasing their overall frequency.

These findings fit with the pattern of results found in the present study, in which post-nap grammatical knowledge was numerically better than predicted in the grammar-cued group and was significantly lower than predicted in the tone-cued group. In addition, for the subset of participants who were not successfully cued ($n = 9$), the difference between predicted and actual performance on the post-nap test was close to zero, and fell between values observed in the grammar-cued and tone-cued groups (Fig. 3). It is important to note that this group was designated post hoc and does not represent an ideal "no-cues" control group, as these participants had systematically poorer sleep quality than successfully cued participants. Nonetheless, this comparison does provide some indication that the presentation of the tone cues may have been worse for grammatical consolidation and generalization than no auditory stimulation at all. Assuming a limited number of reactivation events during a finite period of SWS, the tone cues may have biased reactivation events toward the tone task at the expense of the grammar task, disrupting grammatical consolidation and generalization that would normally occur due to spontaneous reactivations. Taken together, these results suggest that TMR cues produce a consolidation bias, rather than a pure gain.

4.2. Rule abstraction and not chunk knowledge primarily contributes to grammar learning in this novel task

In this experiment, we designed and implemented a novel, interactive learning task, in which learning of grammatical rules proceeds entirely on the basis of trial-and-error, in order to examine the effect of cueing during sleep on grammatical rule extraction. Interestingly and perhaps somewhat surprisingly, we found that performance on this task was primarily driven by abstract rule knowledge and not by knowledge of chunks (bigrams and trigrams). If explicit memories for chunks presented during training contributed to task performance, phrases containing chunks that had been more frequently presented during training should have shown better accuracy rates than phrases containing less familiar chunks. In contrast, ordering performance for high chunk and low chunk strength items was virtually identical (48% compared to 47%), and across all participants there was no relationship between chunk strength and accuracy (mean $r = 0$). In addition, performance for items that were composed entirely of bigrams and trigrams not previously presented during training ("zero-chunk-strength" items) was both significantly above chance and not different from performance for high-chunk-strength items. In sum, the surface similarity between test items and training items did not influence ordering performance, suggesting that performance was supported by abstract rule knowledge of the underlying grammar, rather than by memory for specific bigrams and trigrams presented during training.

This result is notable because it contrasts with findings from the artificial grammar learning literature, in which it has been demonstrated that both grammaticality and chunk strength can influence classification judgments (e.g., Chang & Knowlton, 2004; Forkstam, Hagoort, Fernandez, Ingvar, & Petersson, 2006; Knowlton & Squire, 1996; Lieberman et al., 2004; Meulemans & Van der Linden, 1997; Perruchet & Pacteau, 1990). This discrepancy suggests that there may be interesting differences between the experimental task used in the current experiment and the traditional AGL task. In the AGL task, the artificial language is composed of strings of letters generated according to a finite state grammar. Therefore, specific letter strings may often be repeated within a typical training set, leading to a strong memory for individual chunks in healthy learners, which can then be used as a basis for classification judgments. In contrast, the artificial language used in our study makes use of word categories, in which each category contains two to four nonsense words. Because there is no one-to-one correspondence between each word category and word exemplar, the training phrases consisted of a large number of different word combinations. This variability may prevent the formation of strong memory representations of specific bigrams and trigrams, leading participants to rely instead upon abstract grammatical rule representations to support task performance. The idea that learners may preferentially rely upon abstract rule knowledge when more concrete exemplar specific knowledge is not available (or available at only weak levels) runs parallel to findings from amnesic patients on the AGL task. It has been shown that patients with amnesia exhibit intact classification performance on the AGL task (Knowlton & Squire, 1994, 1996), indicating that learners can rely upon abstract rule knowledge to support performance when explicit memory for bigrams and trigrams is impoverished.

4.3. TMR effects on grammar learning are mediated through rule-based knowledge

Theoretically, task-specific TMR cues could improve grammar learning through at least two possible routes. One possibility is that TMR could strengthen the associations between individual words that commonly co-occur. For example, a particular word pair (e.g., *biff lum*) presented as part of a cued phrase during sleep could trigger the reactivation of all phrases containing that word pair from the initial training set. This would strengthen the association between "biff" and "lum," leading to a stronger representation of the "biff lum" chunk. Because frequent chunks would be reactivated more often, this would lead to a selective enhancement of TMR for high chunk items.

An alternative possibility is that presenting phrases from the artificial language during sleep improves grammar learning in a more general way, for example by serving as a cue that triggers reactivation of the prior learning context (e.g., Diekelmann et al., 2011; Rasch et al., 2007) or of an entire category of learned items (e.g., Oudiette, Antony, Creery, & Paller, 2013). Generalization during sleep is thought to occur when individual memories that share

common elements are simultaneously reactivated, promoting or strengthening the shared connections, while the idiosyncratic aspects of each memory are gradually eroded (Lewis & Durrant, 2011). In theory, this process should occur regardless of whether the individual memories are reactivated spontaneously or through external stimulation (i.e., through presentation of TMR cues). Thus, general reactivation of the learning context through presentation of individual phrases could trigger the reactivation of multiple recently learned phrases in memory, regardless of the exact content of each phrase. This would then strengthen the common features shared between the reactivated phrases, over time leading to the abstraction of the grammatical information governing the collective set of training phrases. For example, if several phrases or phrase segments containing "A" and "D" elements were reactivated, this may lead to the extraction of the relationship between "D" and "A" words (i.e., "D" words, when present, follow "A" words).

Our results more strongly support the second possibility. Cueing during sleep influenced ordering performance for both high-chunk and low-chunk phrases, with no significant effect of chunk strength on TMR-related effects. This result suggests that TMR impacts grammar learning by promoting rule abstraction, rather than by strengthening individual memories for highly frequent chunks. This idea converges with results from a recent AGL study, which found that sleep-related improvements in classification performance were driven specifically by an enhancement of rule abstraction; the effect of chunk frequency was unaltered by sleep (Nieuwenhuis et al., 2013). This conclusion is also consistent with evidence that TMR can often reactivate multiple related memories rather than just single item memories in isolation. For example, studies using contextual TMR cues—in which a background odor is presented first during an object-location learning task and then again during a subsequent period of SWS—report an improvement in overall task performance relative to control conditions (Diekelmann et al., 2011; Rasch et al., 2007). These results suggest that presentation of the odor cue during SWS reactivated a set of multiple learned spatial associations. In another study, participants were trained on associations between a set of individual items and spatial locations, with each individual item paired with an associated auditory cue (e.g., cat–meow; Oudiette et al., 2013). Items were designated as either high or low value, and participants were instructed to plan learning to maximize their score. Critically, when half of the low value cues were presented during sleep, the entire set of low-value associations benefitted, with no specific benefit for the cued low-value items relative to the uncued ones. This finding suggests that sleep cues for the low-value objects triggered reactivation of the whole set of low-value objects, reinforcing memory for the entire domain. Thus, cueing can benefit the entire set of items within a discrete category, consistent with our finding that TMR benefits extended to all phrases within the grammar, regardless of chunk strength. By reactivating multiple newly acquired memories simultaneously, TMR may then promote generalization and rule abstraction for memories that share an underlying structure.

4.4. The contribution of explicit and implicit knowledge to task performance

According to one widely accepted framework, knowledge is implicit when structural knowledge—knowledge of the structure of training items, which may consist of rules, particular items, or fragments knowledge—is unconscious (Dienes & Scott, 2005). Unconscious structural knowledge can be inferred from unconscious judgment knowledge, defined as the ability to know whether a particular test item has the same structure as a training item (Dienes & Scott, 2005). Judgment knowledge is implicit when participants lack meta-knowledge of what they have learned, either because they believe they are guessing when in fact they are above chance (*the guessing criterion*), or because their confidence is unrelated to their accuracy (*the zero-correlation criterion*; Dienes & Berry, 1997). According to these criteria, we found evidence that task performance was supported by both explicit and implicit knowledge. Overall accuracy was strongly related to confidence, with more accurate responses associated with higher degrees of confidence. In other words, participants could recognize whether a phrase that they had produced was grammatical or not, indicating that they had conscious judgment knowledge. As evidenced by responses in the written questionnaire, most participants were also capable of verbally describing a small number of grammatical rules (e.g., "*biff* and *rud* usually came first"), providing evidence of at least some conscious structural knowledge. However, implicit knowledge also partially contributed to task performance, at least on a subset of trials. When participants claimed to be guessing, accuracy rates were still far above chance, indicating that some level of grammatical knowledge was present even in the absence of conscious judgment knowledge. In sum, the present learning task generates both explicit and implicit knowledge of the underlying grammar, but it is likely that explicit knowledge makes a larger contribution to test performance than does implicit knowledge.

4.5. Experimental design limitations

One potential issue is that participants showed a large amount of individual variability in initial learning of the grammatical rules. By implementing an accuracy criterion, we hoped to roughly equate end-of-learning performance across participants. However, this approach necessarily created variability in overall training durations, with some participants requiring a far greater number of phrases to reach the accuracy criterion than others. Ideally both end-of-training accuracy and duration of training would have been less variable across participants, which may have led to greater power to detect sleep-related effects on behavioral changes (e.g., correlations with specific sleep stages and sleep physiology). Because of this variability, perhaps TMR effects were small in magnitude and detectable only with the full sample and the Accuracy Metric approach. TMR effects were not significant within the small subgroup of participants who received both the Pre-Nap and Post-Nap tests.

4.6. Conclusions

Results from this study demonstrated that TMR procedures, previously shown to improve declarative memory and skill learning (e.g., Antony et al., 2012; Diekelmann et al., 2011; Fuentemilla et al., 2013; Rasch et al., 2007; Rihm et al., 2014; Rudoy et al., 2009; Schreiner & Rasch, 2014), can also be used to bias grammatical rule acquisition. These findings suggest that TMR functions not only to strengthen item memories or specific associations between cues and motor responses, but may also potentially influence abstract or qualitative changes in memory. Although effects were modest (approximately 10%), with further development this procedure may eventually be applied to improve grammar learning with natural languages. This technique could be especially beneficial for second-language learners, for whom the acquisition of new grammatical rules typically poses a particular challenge (e.g., Johnson & Newport, 1989; Weber-Fox & Neville, 1996). Most importantly, these TMR results show that sleep provides benefits for generalization learning. Most previous support for this idea has come from wake-versus-sleep comparisons (e.g., Djonlagic et al., 2009; Durrant et al., 2011; Fenn et al., 2003; Gaskell et al., 2014; Gómez et al., 2006; Nieuwenhuis et al., 2013; Wagner et al., 2004), or from

sleep-physiology and behavioral correlations (Batterink et al., 2014; Djonlagic et al., 2009; Durrant et al., 2011). By directly manipulating processing during sleep, our study provides novel evidence for this idea (e.g., avoiding limitations due to greater interference during wake in studies with wake-versus-sleep comparisons). The finding that generalization performance is sensitive to the content of sleep cues implies that mechanisms contributing to generalization learning are at play spontaneously during natural sleep.

Acknowledgments

This work was funded through NIH grants T32 NS047987 and F32 HD 078223, and NSF grant BCS-1461088.

References

Antony, J. W., Gobel, E. W., OðHare, J. K., Reber, P. J., & Paller, K. A. (2012). Cued memory reactivation during sleep influences skill learning. *Nature Neuroscience, 15*, 1114–1116.

Arzi, A., Shedlesky, L., Ben-Shaul, M., Nasser, K., Oksenberg, A., Hairston, I. S., & Sobel, N. (2012). Humans can learn new information during sleep. *Nature Neuroscience, 15*, 1460–1464.

Batterink, L., Oudiette, D., Reber, P. J., & Paller, K. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia, 65*, 169–179.

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature Neuroscience, 15*, 1439–1444.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.

Chang, G. Y., & Knowlton, B. J. (2004). Visual feature learning in artificial grammar classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 714–722.

Cousins, J. N., El-Deredy, W., Parkes, L. M., Hennies, N., & Lewis, P. A. (2014). Cued memory reactivation during slow-wave sleep promotes explicit knowledge of a motor sequence. *The Journal of Neuroscience, 34*, 15870–15876.

Creery, J. D., Oudiette, D., Antony, J. W., & Paller, K. A. (2015). Targeted memory reactivation during sleep depends on prior learning. *Sleep, 38*, 755–763.

De Gennaro, L., & Ferrara, M. (2003). Sleep spindles: An overview. *Sleep Medicine Reviews, 7*, 423–440.

Diekelmann, S., Büchel, C., Born, J., & Rasch, B. (2011). Labile or stable: Opposing consequences for memory when reactivated during waking and sleep. *Nature Neuroscience, 14*, 381–386.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review, 4*, 3–23.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*, 338–351.

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning and Memory, 16*, 751–755.

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

Durrant, S., Taylor, C., Cairney, S., & Lewis, P. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*, 1322–1331.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 7723–7728.

Emmons, W. H., & Simon, C. W. (1956). The non-recall of material presented during sleep. *American Journal of Psychology, 69*, 76–81.

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

Ferrarelli, F., Huber, R., Peterson, M. J., Massimini, M., Murphy, M., Riedner, B. A., ... Tononi, G. (2007). Reduced sleep spindle activity in schizophrenia. *American Journal of Psychiatry, 164*, 483–492.

Forkstam, C., Hagoort, P., Fernandez, G., Ingvar, M., & Petersson, K. M. (2006). Neural correlates of artificial syntactic structure classification. *NeuroImage, 32*, 956–967.

Fuentemilla, L., Miro, J., Ripolles, P., Vila-Ballo, A., Juncadella, M., Castaner, S., ... Rodriguez-Fornells, A. (2013). Hippocampus-dependent strengthening of targeted memories via reactivation during sleep in humans. *Current Biology, 23*, 1769–1775.

Gaskell, M. G., Warker, J., Lindsay, S., Forst, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep underpins the plasticity of language production. *Psychological Science, 25*, 1457–1465.

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science, 17*, 670–674.

Grigg-Damberger, M. M. (2012). The AASM scoring manual four years later. *Journal of Clinical Sleep Medicine, 8*, 323–332.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569–1579.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60–99.

Knoblauch, V., Martens, W. L., Wirz-Justice, A., & Cajochen, C. (2003). Human sleep spindle characteristics after sleep deprivation. *Clinical Neurophysiology, 114*, 2258–2267.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 79–91.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 169–181.

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences, 15*, 343–351.

Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., & Knowlton, B. J. (2004). An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. *Journal of Cognitive Neuroscience, 16*, 427–438.

Mander, B. A., Santhanam, S., Saletin, J. M., & Walker, M. P. (2011). Wake deterioration and sleep restoration of human learning. *Current Biology, 21*, R183–R184.

Menn, L., & Ratner, N. B. (2000). Methods for studying language production. In L. Menn & N. B. Ratner (Eds.), *In the beginning was the wug: Forty years of language-elicitation studies* (pp. 1–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meulemans, T., & Van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1007–1028.

Milner, C. E., Fogel, S. M., & Cote, K. A. (2006). Habitual napping moderates motor performance improvements following a short daytime nap. *Biological Psychology, 73*, 141–156.

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep promotes the extraction of grammatical rules. *PLoS ONE, 8*, e65046.

Nir, Y., Staba, R. J., Andrillon, T., Vyazovskiy, V. V., Cirelli, C., Fried, I., & Tononi, G. (2011). Regional slow waves and spindles in human sleep. *Neuron, 70*, 153–169.

Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PLoS ONE, 4*, e341.

Oudiette, D., Antony, J. W., Creery, J. D., & Paller, K. A. (2013). The role of memory reactivation during wakefulness and sleep in determining which memories endure. *The Journal of Neuroscience, 33*, 6672–6678.

Oudiette, D., & Paller, K. A. (2013). Upgrading the sleeping brain with targeted memory reactivation. *Trends in Cognitive Sciences, 17*, 142–149.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119*, 264–275.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition, 95*, 201–236.

Rasch, B., Buchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science, 315*, 1426–1429.

Rihm, J. S., Diekelmann, S., Born, J., & Rasch, B. (2014). Reactivating memories during sleep by odors: Odor specificity and associated changes in sleep oscillations. *Journal of Cognitive Neuroscience, 26*, 1806–1818.

Rudoy, J. D., Voss, J. L., Westerbrg, C. E., & Paller, K. A. (2009). Strengthening individual memories by reactivating them during sleep. *Science, 326*, 1079.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*, 493–515.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47*, 172–196.

Schabus, M., Dang-Vu, T. T., Albouy, G., Balteau, E., Boly, M., Carrier, J., ... Maquet, P. (2007). Hemodynamic cerebral correlates of sleep spindles during human non-rapid eye movement sleep. *Proceedings of the National Academy of Sciences, 104*, 13164–13169.

Schonauer, M., Geisler, T., & Gais, S. (2013). Strengthening procedural memories by reactivation in sleep. *Journal of Cognitive Neuroscience, 26*, 143–153.

Schreiner, T., & Rasch, B. (2014). Boosting vocabulary learning by verbal cueing during sleep. *Cerebral Cortex*, pii: bhu139. [Epub ahead of print].

Shadmehr, R., & Moussavi, Z. M. K. (2000). Spatial generalization from learning dynamics of reaching movements. *The Journal of Neuroscience, 20*, 7807–7815.

Simon, C. W., & Emmons, W. H. (1956). EEG, consciousness and sleep. *Science, 124*, 1066–1069.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience, 16*, 139–145.

Urakami, Y. (2008). Relationships between sleep spindles and activities of cerebral cortex as determined by simultaneous EEG and MEG recording. *Journal of Clinical Neurophysiology, 25*, 13–24.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature, 427*, 352–355.

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8*, 231–256.