Title: **Multiple memories can be simultaneously reactivated during sleep as effectively as a single memory**

Abbreviated title: **Parallel memory reactivation during sleep**

Eitan Schechtman[1]*, James W. Antony[2], Anna Lampe[1], Brianna J. Wilson[1], Kenneth A. Norman[2] & Ken A. Paller[1]

1 – Department of Psychology, Northwestern University, Evanston, IL 60208, USA.

2 – Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA

* Corresponding author: eitan.schechtman@northwestern.edu, Department of Psychology, Northwestern University, Evanston, IL 60208, USA

Number of pages: 40

Number of figures (excluding supplementary figures): 5

Number of words (Main text): 4,355

Number of words (Online methods): 2,596

Conflict of interest: The authors declare no competing financial interests.

Abstract

Memory consolidation during sleep involves reactivation of memory traces. Targeting specific memories by presenting learning-related cues during sleep selectively enhances memory, but the mechanism behind this benefit is not fully understood. To better characterize the consolidation process in humans, we tested whether multiple memories can be reactivated in parallel using a spatial-memory task. After learning the locations of images belonging to semantically related sets of one, two, or six items, half of the sets were reactivated during a nap. Results showed a selective benefit in location recall for cued versus non-cued items regardless of set size, implying that reactivation may occur in a simultaneous, promiscuous manner. Intriguingly, sleep spindles and delta power modulations were sensitive to set-size and reflected the extent of previous learning. Taken together, our results refute the notion that resource availability strictly reduces the capacity of simultaneous sleep-reactivation and bring forward alternative testable models for sleep-related consolidation.

Acknowledgments: This work was supported by NSF grants BCS-1461088 and BCS-1533511. E.S. was funded by the Human Frontier Science Program and the Zuckerman STEM Leadership Program.

Author contribution: All authors contributed to the design of this study and helped revise the manuscript. E.S and B.W collected the data. E.S conducted the analyses and wrote the manuscript.

More than a century after researchers started to explore the beneficial effects of sleep on memory[1], the mechanism by which this benefit is achieved in still debated[2]. The leading hypothesis, termed active systems consolidation[3], postulates that memories are stored in the hippocampus and then reactivated during sleep, subsequently shaping neocortical memory traces. Reactivation of memories during sleep was first observed in rodents[4, 5]. Sequential learning-related spiking activity was shown to "replay" during sleep, and this phenomenon has since been connected to the off-line consolidation process[6]. In Humans, recent studies using multivariate pattern classification analysis in fMRI have also shown evidence for reactivation of cortical and hippocampal memory-related patterns during sleep and awake rest[7, 8, 9, 10].

An important milestone in the study of sleep-related reactivation has been the development of targeted memory reactivation (TMR), a paradigm designed to reactivate specific memories by unobtrusively presenting learning-related cues during sleep[11]. TMR has been shown to improve various forms of learning, including spatial[12, 13], skill[14, 15] and vocabulary[16] learning. A recent TMR-fMRI study[17] demonstrated that cuing reactivated category-level learning-related cortical activity, further establishing the link between TMR-related memory reactivation and spontaneous reactivation during sleep.

TMR spatial-learning paradigms have predominantly used olfactory and auditory cues to reactivate memories. Whereas olfactory designs have associated multiple learned items to a single odor (e.g., 15 items[13, 18, 19]), auditory designs have commonly used sounds that were associated with a single item (and more recently, with two items[20, 21]). Both techniques have consistently shown benefits for cued items, but the question of whether effect sizes rely on the number of items reactivated has never been directly addressed. The relationship between the number of items cued and the cuing benefit may shed light on fundamental attributes of sleep reactivation. Importantly, it may help address the question of whether memories may be reactivated simultaneously or whether limited reactivation resources strictly govern sleep-related consolidation and limit it to some number of items at any given time.

To examine these questions, we designed a study contrasting the effects of TMR for single items and larger groups of items. In a spatial-memory task with auditory TMR, we required participants to learn the locations of 90 images on a 2D circular grid (Figure 1a). Items were strategically grouped into distinct sets, and each set included either six items (e.g., six distinguishable images of different cats), two items, or a single item. Each set was associated with a single unique sound (e.g., a single meow linked to all six cats). The locations of unique items within a set were learned to criterion in separate blocks to minimize competition. Participants next were allowed a 90-minute nap opportunity. During non-REM sleep (NREM), half of the sounds for each set-size were unobtrusively presented (Figure 1b). The benefits of TMR were compared within-participant between cued and non-cued items.

Using this design, we aimed to tease apart three hypotheses regarding the neurocognitive mechanisms by which TMR achieves its benefit during sleep (Figure 1c):

1. If reactivating multiple items involves the separate reactivation of each item, in a manner similar to path-specific replay, these activations may each incur some cost, or may rely on common limited resources. The Limited Capacity Hypothesis (LCH; Figure 1c, top) assumes a finite capacity for simultaneous reactivation during sleep. By this view, if multiple items are reactivated together, the benefit of reactivation is split among all items associated with the presented cue. The LCH thus predicts that items in smaller sets will be reactivated to a larger extent and benefit more than items in larger sets. A helpful analogy to consider is a pipeline in which the volume of liquid each pipe receives depends on the number and diameter of downstream pipes.
2. Alternatively, reactivation may be achieved in a parallel, independent manner. The Parallel Reactivation Hypothesis (PRH; Figure 1c, center) suggests that there is no “cost” to reactivating multiple items at the same time as opposed to a single item. In our design, this hypothesis predicts similar benefits to all items in a set, regardless of set-size.

3. Another possibility is that sleep-related reactivation is limited to a single item at any given time. The Sampling Hypothesis (SH; Figure 1c, bottom) suggests that following each cue presentation during sleep, exactly one item of the set is reactivated and benefits. This hypothesis is supported by a recent study that found that only one item of a pair but not the other benefits from cuing[20]. The reactivated item may be selected randomly with each presentation of a cue or there could be some systematic bias towards reactivating a specific item (e.g., weakly activated memories might more likely be reactivated[8, 22]). An extension of this hypothesis postulates that a small subset (i.e., more than one – but not all items) is reactivated and benefit.

Although these hypotheses are not mutually exclusive, our design was ideally suited to contrast these hypotheses in a controlled setting. The results challenge all hypotheses except for the PRH, suggesting that reactivation may be simultaneous and, in a sense, cost-free. We use these findings to question the assumption that sleep-related consolidation is necessarily based on reactivation of individual memories and not of a wider, more generalized context or semantic structure.

Results

The benefit of sleep reactivation was uninfluenced by set size

Spatial-memory for learned items was evaluated before and after sleep. In previous spatial-memory TMR paradigms with single distinct images, each associated with a unique sound, spatial accuracy was measured as the pixel distance between correct and recalled positions[12, 21]. Our task included a categorical structure, whereby several distinguishable images (e.g., different cats) were associated with the same sound (e.g., meow). Consequently, we could ascertain the influence of set-size on the benefit of TMR. However, another source of error that may vary with set-size is that participants may mistakenly recall the location of an object other than the one they attempted to recall. We therefore

designed the study in a manner that allowed us to dissociate these so-called swap errors from placement errors that more directly reflect degree of spatial accuracy in recall (see Methods).

Focusing first only on sets with two or six items, we considered the effects of set-size and cuing during sleep on the change in swap errors. The average number of swap errors for six-item sets was 2.31 before sleep and 2.34 after sleep (change = +0.03). The average number of swap errors for two-item sets was 0.35 before sleep and 0.36 after sleep (change = +0.01). Using a repeated-measures ANOVA we found no effect of cuing ($F(1,30)=0.43$, $p=0.52$) and no interaction between cuing and set-size ($F(1,30)=0.24$, $p=0.63$).

Previous spatial TMR studies have found effects of cuing on accuracy[12, 21, 22]. For all subsequent analyses, we therefore shifted our focus to accuracy errors and included only items that were not considered as swapped (see Methods). We first analyzed pre-sleep test results to confirm that they did not differ between cued and non-cued sets. Using a repeated-measures ANOVA we found that pre-sleep error rates were not different between cued and non-cued sets ($F(2,30)=0.74$, $p=0.4$), nor was there an interaction between cuing status and set-size ($F(2,30)=0.14$. $p=0.87$). Error rates differed between set-sizes ($F(1,30)=7.52$, $p<0.01$), with greater errors for sets of six and two items relative to single items ($p<0.01$ for one vs. six; $p<0.01$ for one vs. two; $p=0.7$ for two vs. six; Tukey's HSD, Figure 2).

The three aforementioned hypotheses make different predictions regarding how set-size would influence cuing. To test these effects, we calculated, per participant, the average change (in percentage) in spatial error over the period of sleep for different set-sizes (six, two, or one) and cuing status (cued or non-cued). As shown in Figure 2, cuing influenced recall accuracy ($F(1,30)=13.6$, $p<0.001$), producing enhanced recall for cued compared to non-cued locations. However, accuracy was not modulated by set-size ($F(2,30)=0.07$, $p=0.93$) or the interaction between set-size and cuing status ($F(2,30)=0.21$, $p=0.81$). Using the Bayesian Information Criterion to estimate the Bayes factor[23, 24, 25] indicated strong

evidence that there was no interaction effect (BF_{10}≈25). On average, spatial-memory increased by 1.72% for cued sets and decreased by 10.77% for non-cued sets.

The aforementioned differences in pre-sleep accuracy among sets of different sizes (i.e., error rates were significantly smaller for one-item sets) could potentially have influenced our results. To deal with this possible bias, we conducted two complementary analyses (Supplementary Figure 1). First, we regressed out pre-sleep results to eliminate set-dependent effects of regression to the mean[20, 26] (see Methods). This analysis produced similar results, with a significant effect for cuing ($F(1,30)=17.89$, $p<0.001$) and a nonsignificant interaction between set-size and cuing status ($F(2,30)=0.11$, $p=0.9$). For the second analysis, we chose 500 subsampled versions of our dataset that eliminated significant pre-sleep differences (see Methods). The results were consistent with the results obtained for the complete dataset, with 68% of the subsampled datasets showing a significant ($p<0.05$) cuing effect (mean(p)=0.056, median(p)=0.025), compared with 0.8% of data sets showing a significant ($p<0.05$) cuing by set-size interaction (mean(p)=0.59, median(p)=0.61).

Because training involved a binary criterion short of perfection, in that placements within 100 pixels from the target were deemed correct, participants may have been shaped to ignore small errors in accuracy. We therefore repeated the analysis including a lenient measure of accuracy – the number of items considered correct using the training criterion. Results were qualitatively similar, with an effect of cuing ($F(1,30)=5.55$, $p<0.03$), no effect of set-size ($F(2,30)=1.42$, $p=0.25$), and no interaction ($F(2,30)=1$, $p=0.37$). On average, the number of correct trials increased by 0.04 for cued sets and decreased by 0.04 for non-cued sets.

These results indicate that set-size within the range sampled did not influence the benefit of sleep for the individual items within a set. The notion that there is a finite capacity for reactivation during sleep (the LCH), favoring the smaller set-size, is therefore not supported by the results.

Within-set relationships between items did not support the Sampling Hypothesis

After ruling out the LCH, we shifted our focus to test the SH. The predictions made by this hypothesis depend on whether an item (or a small subset of items) is randomly selected and reactivated with each cue presentation during sleep or whether some bias governs which item will be selected and reactivated over multiple presentations. To examine the latter, we calculated the intra-class correlation (ICC) for sets consisting of six and two items[27, 28]. The ICC, commonly used in reliability analyses, provides a measure of total agreement among measurements within a set. If a small subset of items within a set benefited from cuing whereas the rest were on par with items of non-cued sets, we expected higher agreement for non-cued items. To obtain a measure of significance, we compared the difference between the ICC for the cued and non-cued sets within each set-size to the differences obtained after randomly shuffling the data (see Methods). The ICC values for six-item sets did not differ from those obtained by random shuffling for cued and non-cued data (p>0.47 for both), but for two-item sets the ICC values were higher than those obtained by shuffling for both cued and non-cued sets (p<0.01 for both). The critical question for this test, however, was whether the differences between ICCs for cued and non-cued item was higher than chance. Regardless of set-sizes, the ICC was not different between cued and non-cued sets (p=0.45 for six items; p=0.81 for two items; Figure 3a).

The ICC is essentially a measure of variability. However, large variability does not necessarily stem from a subset of deviant measurements (i.e., it could be due to a wider distribution). To complement the ICC analysis, we considered a primitive measure of deviance within set. For each set, we calculated the Z-scores of each item and used the maximum absolute value as a representative statistic to consider whether the set includes an outlier-like data point. These scores were not different between cued and non-cued sets of six items (t(119)=-1.1, p=0.27; Figure 3b). Taken together, the ICC and Z-scores analyses did not support the biased form of the SH.

Alternatively, SH could posit that a small subset of random items would be reactivated with each cue presentation during sleep. This hypothesis is inconsistent with the lack of an effect of set-size on cuing benefit, as described above. If items were randomly sampled with each cue presentation, the probability for an item within a larger set to benefit from the cue would be smaller than the probability for an item within a smaller set. Like the LCH, the random version of the SH would therefore also predict a graded, set-size dependent benefit of cuing on memory. This prediction was not supported by our data.

Another prediction, based on this hypothesis, is that memory benefit would correlate with the number of times a cue was presented during sleep. Previous studies with single items have often failed to show correlations between the number of repetitions of a cue during sleep and the benefit of TMR[29, 30]. However, if random sampling occurred, the number of repetitions in sets with more than a single item would be a proxy of the probability of an item to be reactivated (e.g., the probability for an item in a six-item set being reactivated at least once would be 0.17, 0.31 and 0.42 for one, two or three repetitions, respectively). We tested this prediction across participants and found no significant correlations (r=0.29, p=0.29 for single-item sets; r=0.01, p=0.95 for sets of two item; r=0.09, p=0.63 for sets of six items; Figure 3c). Taken together, these data contradict the predictions made by the random-selection version of the SH. Altogether, our results do not support the SH, making PRH the only one of our three hypotheses that is not challenged by our findings.

Delta and sigma activity after cue presentation during sleep was modulated by set-size

Throughout the experiment, we collected scalp electroencephalographic data. We used these data for sleep staging and also to analyze responses to cue presentation during sleep, focusing on two frequency bands shown to be modulated by cues in previous studies[26, 31]. The sigma range (11-16 Hz) encompasses sleep spindles, which are ramped waves lasting between 500-3000 ms. Spindles have been linked to

memory consolidation (see [32] for review). The delta band (0.5-4 Hz) includes both slow waves (the defining characteristic of N3, the deepest stage of NREM sleep) and K-complexes (waveforms consisting of a fast negative peak and a slower positive component that commonly appear individually during stage N2 and N3). Examining the time-frequency response following cue presentation, we identified two modulated clusters corresponding to these two frequency bands, similar to the ones identified in other studies[31].

Considering these two bands, we examined the power modulation for each set-size (Figure 4a). In addition to playing the sounds that had been associated with one, two, or six items before sleep, we presented a single novel sound as well and included data from that condition in these analyses. Pooling across participants and individual sounds, results showed that power for both frequency bands was modulated by the number of items previously associated with the sound (delta $F(3,6887)=4.42$, $p<0.01$; sigma $F(3,6887)=4.84$, $p<0.01$; Figure 4b). In the delta range, the power modulation for all set-sizes of previously presented sounds was higher than the power modulation for novel sounds ($p<0.01$ for six-item sounds; $p<0.02$ for two-item sounds; $p<0.06$ for one-item sounds, Tukey's HSD). All other differences were nonsignificant ($p>0.16$). In the sigma range, the pattern was different. The power increase for six-item sounds was significantly higher relative to two- and one-item sounds ($p<0.01$ for each, Tukey's HSD) and marginally higher relative to novel sounds ($p<0.07$, Tukey's HSD). All other differences were nonsignificant ($p>0.95$).

This measure of power in the sigma range may be affected by two factors: spindle probability and spindle power. Therefore, the sigma results could represent two scenarios: (a) more spindles occurred in response to six-item sounds relative to other sounds; or (b) a similar number of spindles occurred, but their amplitudes were lower for smaller set-sizes. To disentangle these two scenarios, we used an automated algorithm to detect spindles in single trials and calculated the probability of a spindle following cue presentation as a function of set-size (Figure 4c). Considering the time frame identified for

the previously described cluster, we found that spindle probability depended on set-size ($F(3,6887)=2.79$, $p<0.04$). The probability of a spindle following six-item sounds was higher than that for one-item sounds ($p<0.04$, Tukey's HSD). All other differences were nonsignificant ($p>0.14$).

These analyses were complemented by correlation analyses. Omitting the data for the novel sound, we found modest positive linear correlations between the number of items per set and power modulation in the delta range ($r=0.03$, $p<0.01$), power modulation in the sigma range ($r=0.04$, $p<0.001$), and spindle probability ($r=0.03$, $p<0.01$). Taken together, the results demonstrate that spindle probability was modulated by set-size.

Finally, we also tested whether delta power, spindle power, or spindle probability were correlated with performance in sets of different sizes. None of these correlations proved significant ($p>0.04$, uncorrected; Supplementary Table 1).

Discussion

We sought to investigate neurocognitive mechanisms of consolidation during sleep using targeted memory reactivation of sets of items of different sizes. Results showed that TMR produced a clear benefit in the form of better recall for cued items than for non-cued items. Moreover, this memory benefit was not influenced by set-size, ruling out the possibility that the benefit of reactivation in our study can be conceptualized as a "limited resource" that is divided among individual memories (i.e., the LRH). Alternatively, memory reactivation may benefit a small subset of items (e.g., one of the items) that are sampled either randomly or in a biased manner (e.g., weaker items more likely to be reactivated[8, 22]), as described in the SH. Our data do not support this hypothesis. The random SH predicts bigger benefits for smaller sets and a correlation between the number of times a cue was repeated during sleep and the average benefit for the associated set. Neither of these predictions are supported by our data. The

biased SH predicts higher within-set agreement between benefit scores for non-cued sets, yet our data showed no significant differences between ICC scores and outlier scores for cued and non-cued sets.

Taken together, our results lend support to the PRH, suggesting that reactivation occurs in a parallel manner and that multiple memories can be reactivated independently, either simultaneously or in very rapid succession. This reasoning implies that reactivation does not appear to be a limited resource, opening up the possibility that even larger numbers of memories could conceivably be reactivated simultaneously.

Whereas our findings constitute a step forward in constructing a neurocognitive model for memory reactivation, they also raise new and uninvestigated questions. In the context of TMR, cue presentation first initiates primary sensory processing (i.e., processing that depends on the stimulus properties and not the scope of the memories associated with it). What happens next is open for exploration. One possibility is that the stimulus representation is linked with multiple, independent, highly-specific memory traces (e.g., the spatial position of a single item out of a set of multiple items), which are then reactivated in a parallel (or almost parallel) manner (Figure 5a). The major characteristic of this model is parallel offline reactivation at the single-item level. These reactivations must be independent of each other, to avoid interference and contamination between different memory traces.

A potentially important property of our paradigm is that stimuli and sounds within sets had a close semantic relationship. It is possible that generalized representations of a set – and not only the representations of the individual items – may be of importance for sleep's effect. As an alternative to the item-reactivation model, it could therefore be suggested that the semantic theme of a set forms a context that is reactivated, benefitting all embedded items. This model is reminiscent of computational models of the effects of context on episodic memory, such as the Context Maintenance and Retrieval (CMR) model[33, 34]. Briefly, this model predicts that the memory search will be determined by

associations between items and the context in which they are embedded. The context includes both a semantic clustering component and a temporal clustering component (i.e., items learned in temporal proximity will share the same context). Whereas the interplay between context and memory has been extensively studied, focusing both on its role in recall[35, 36, 37] and with regard to the role of the hippocampus in binding item memories to context[38, 39], the role of context in sleep reactivation has not been systematically explored. If contexts are directly reactivated (whether spontaneously or using TMR), they may then reactivate individual embedded items (Figure 5b). Note that mathematical models of memory posit competition between items linked to a context[40]; however, these models have focused on data from wake, not sleep, and it is possible that memory systems operate differently during sleep. For example, low levels of acetylcholine during slow-wave sleep may put the hippocampus into a strong retrieval mode that makes it possible to retrieve more traces at once[41, 42] (but see[43]).

The idea that TMR cues activate contexts (as opposed to directly activating items) may help to explain data on the time course of TMR. Bendor and Wilson[44] presented learning-related auditory cues during sleep and found a bias toward reactivation of cued memories, but intriguingly, this bias was sustained for multiple seconds after cue offset. At that point in time, any replay that was directly elicited upon initiation of the auditory cue should have ended. One of the characteristics of context, at least during wake, is its slow temporal drift[45], and therefore these data may be better fit by a model in which the TMR cue activates a context representation that persists over time and cues individual items.

To reveal neural correlates of reactivation of multiple items over sleep, we analyzed modulations in sigma and delta frequency bands following cue presentation during NREM sleep. Both sigma power, consisting mostly of sleep spindle activity, and delta power, consisting of slow waves and K-complexes, increased in the seconds following sound onset. Crucially, these boosts in activity were modulated by the size of the set associated with the presented sound, with the highest increase in both bands

measured after sounds associated with six items. This pattern was also apparent when considering spindle probability over the same timeframe.

Spindles have long been associated with memory consolidation during sleep and have been hypothesized to embed previously learned information[32]. Recently, spindles have been shown to coincide with time windows in which decoding of stimulus properties is possible[31], further solidifying their significance for memory consolidation. Previous studies have contrasted spindles following previously learned and novel stimuli[46], yet our results seem to be the first to manipulate memory load and consider its effect on spindle power and probability, showing that the extent of previous learning affects these measures.

Unlike spindles, whose role in memory has been thoroughly explored, memory-related aspects of stimulus-evoked delta activity have attracted little attention. Slow-wave activity during sleep has been shown to be higher after learning both globally[47] and locally for previously engaged brain regions[48], but these effects reflect a coarse, task-related increase that is not linked to any specific memory. K-complexes, which together with slow waves make up most of the modulations in the delta range, are usually explored in the context of sensory processing and have been hypothesized to protect against arousal[49]. However, this gating is not purely sensory and depends at least to some degree on cognitive analysis, as evidenced by larger K-complexes to cognitively salient stimuli such as one's own name[50, 51]. Recently, Andrillon and colleagues[52] found that learning increases triggered delta power, but our results seem to be the first showing that delta power is correlated with the scope of information linked with a stimulus. This finding may suggest that delta power (and possibly K-complexes) are sensitive to information at higher levels than previously thought.

Despite the apparent discrepancy in our experiment between the behavioral finding of set-size-independent memory benefits and the physiological finding of set-size-dependent increases in delta and

sigma power, we propose that the results can be reconciled as follows. Given the finding that the average benefit per item remains the same regardless of set-size, the cumulative benefit across items within a set would be much larger for larger sets. Delta and sigma power may represent the cumulative benefit for the set associated with the presented cue. Alternatively, it may represent the process by which memories are retrieved and modified—in which case delta and sigma power may reflect the extent of previous knowledge made available for consolidation. Either way, our physiological and behavioral findings can be reconciled and are not at odds with each other.

Finally, some limitations must be acknowledged. First, our results are restricted to sets of relatively few items. It may be, for example, that sets larger than six items would have shown a graded benefit. It is also possible that our findings were influenced by our choice of stimulus categories (i.e., similar and highly associated items with the same congruent sound, that may have been chunked together to constitute a single, unitized memory), or other design features, and will not generalize to other situations. For example, Antony and colleagues[20], using a slightly different spatial-TMR procedure, found that benefits for items within a cued-pair were anti-correlated, supporting the SH (but see contrasting results in [21]). Differences across designs do not allow direct comparisons with our results, so explanations for these discrepancies remain outstanding.

The growing appreciation of the importance of sleep for memory consolidation has raised unexplored questions regarding the relevant neurocognitive mechanisms. Our results suggest that multiple memory reactivations can occur simultaneously and independently during sleep, benefiting several memories in parallel. Reactivation capacity during sleep may therefore be larger than would be assumed with serial reactivation, as in typical working-memory operations. Although our data is consistent with models in which single memory items are reactivated directly, an alternative, previously unexplored model in which generalized contexts are reactivated seems equally likely and worthy of further exploration. Either

way, the notion that reactivation benefits memories in a parallel and promiscuous suggests that the capacity for sleep-related consolidation is far beyond that which was previously imagined.

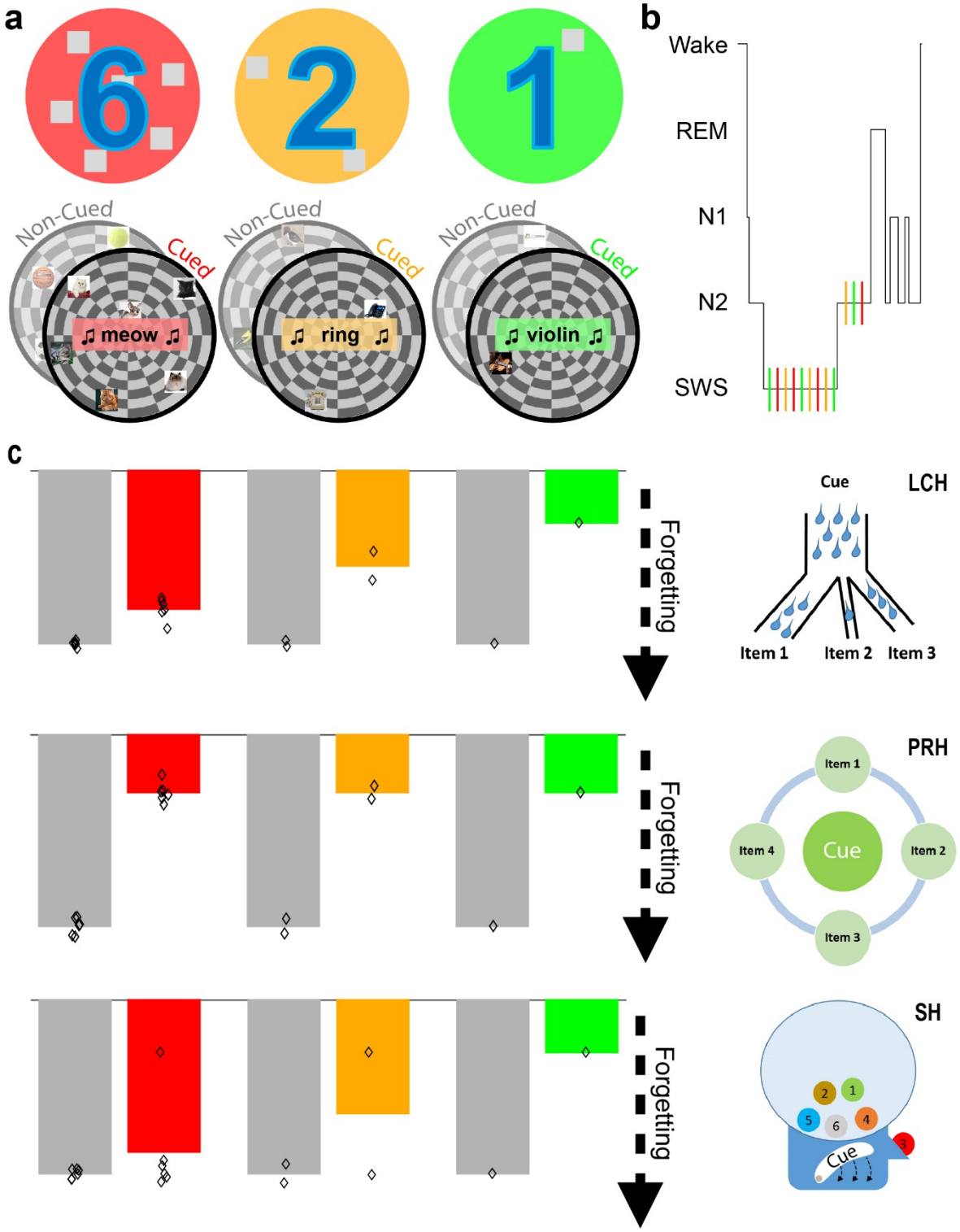


Figure 1: Design and predictions. (a) In a spatial-memory task, participants learned the specific locations of images that appeared on a circular grid. Images belonged to sets that included six images, two

images, or one image (e.g., the cat set consisted of six cats whereas the phone set consisted of two phones). Each set was associated with a set-specific sound (e.g., “meow”, “ring”). Some sets were cued during subsequent sleep (colored sets) and others were not (gray sets). (b) Sounds were presented during stages N2 and N3 in an afternoon nap. (c) Simulated data are shown for three hypotheses regarding how TMR achieves its benefit. Diamonds represent average benefit over sleep for each item within a set (i.e., lower values represent more forgetting) and bars represent average benefit for the all items in the set. Colors signify set-size, corresponding with colors in (a). Predictions from the three hypotheses are figuratively shown in illustrations at right. The Limited Capacity Hypothesis (LCH; top) predicts that larger set-sizes would benefit less from TMR. The Parallel Reactivation Hypothesis (PRH; center) predicts that the benefit per item would not depend on group size. The Sampling Hypothesis (SH; bottom) predicts that a small subset of items (e.g., one) would benefit from cue-presentation regardless of set-size.

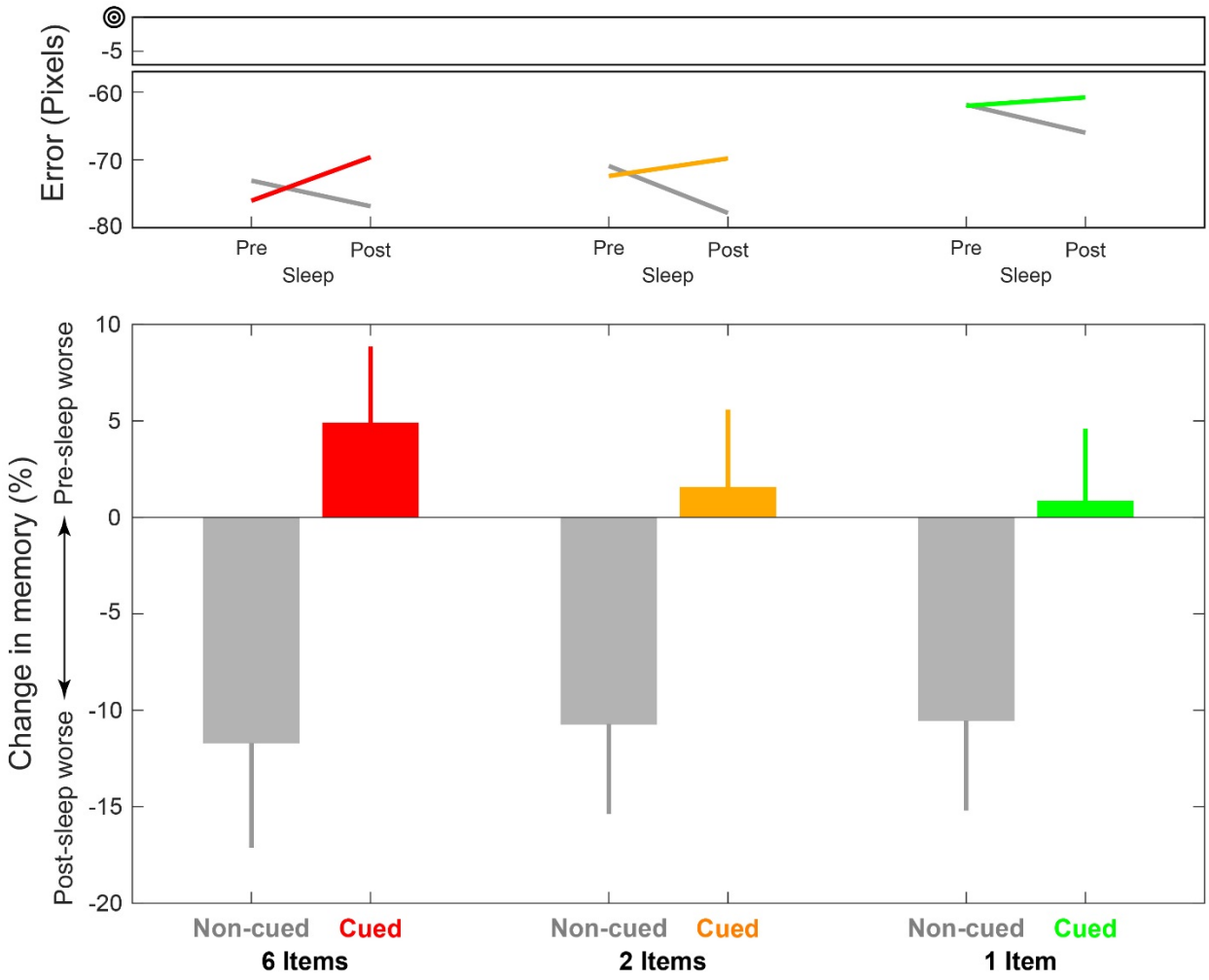


Figure 2: Cuing status, but not set-size, affected sleep-related benefits. The upper panel shows mean error rates before and after sleep as a function of set-size (red – six-item sets; yellow – two-item sets; green – one-item sets) and cuing status (colored – cued; gray – non-cued). Zero error signifies exact recall of the correct location. The lower panel shows the change in memory over sleep (i.e., sleep-related benefit) in percentages. Negative values represent higher errors after sleep. There was a significant effect of cuing status ($p<0.001$) but no cuing-set-size interaction. Error bars signify standard errors of the means (between-subjects).

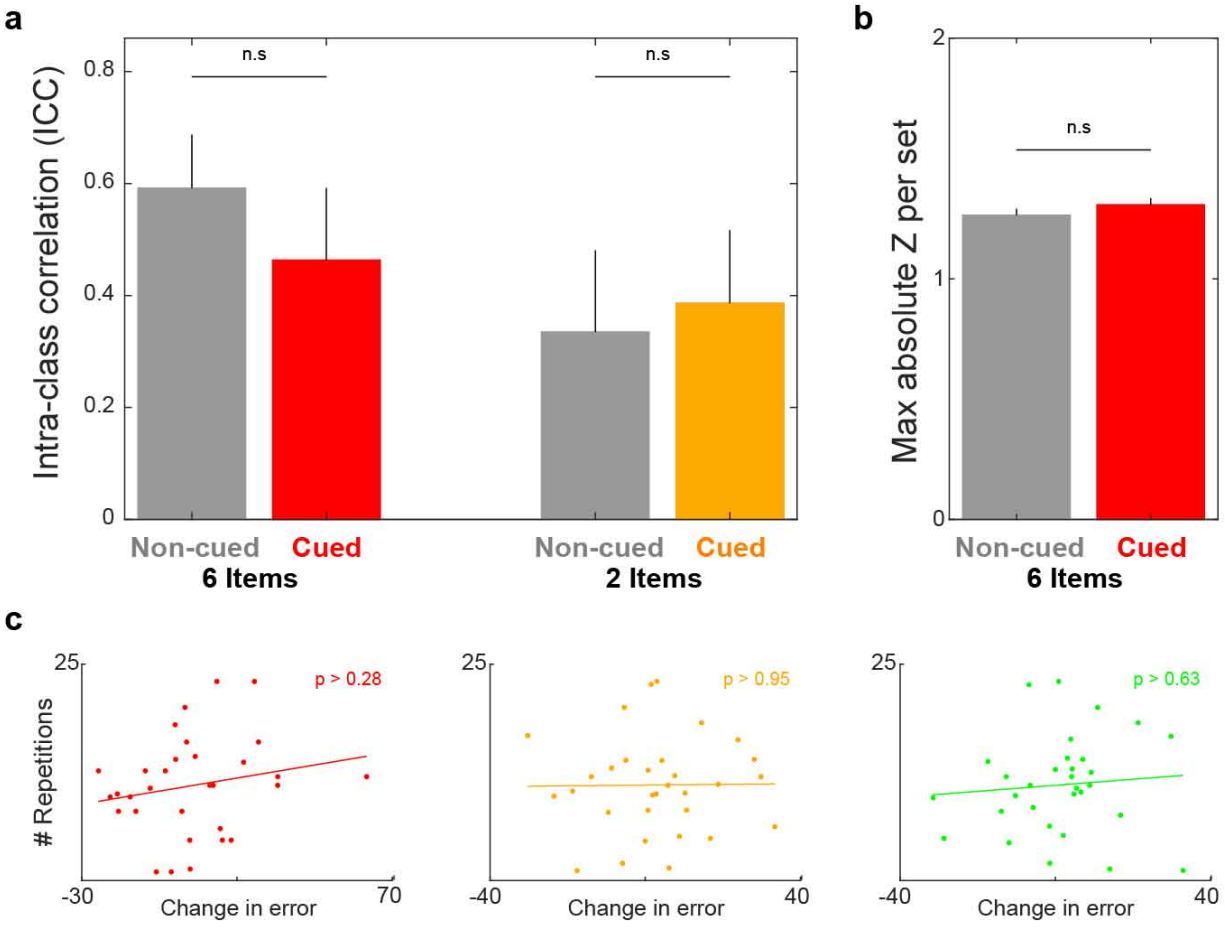


Figure 3: The Sampling Hypothesis is not supported by the behavioral data. (a) If a small subset of items were to benefit from cuing, we would expect significantly higher inter-class correlations between benefits of individual items for non-cued sets relative to cued sets, but no such effect was observed. (b) Another prediction is that the cued sets would be more prone to outlier-like results (e.g., a single item's benefit would be much higher than the others'). Comparisons between the maximal absolute *Z*-score within set for cued and non-cued sets of six items failed to show a significant difference on this measure. (c) If a small subset of items randomly benefited from each presentation, we would expect that more repetitions per cue would results in higher benefits to memory, but these correlations were not significant for any set-size (red – six-item sets; yellow – two-item sets; green – one-item sets). Error bars signify standard errors of the means. n.s – nonsignificant comparisons.

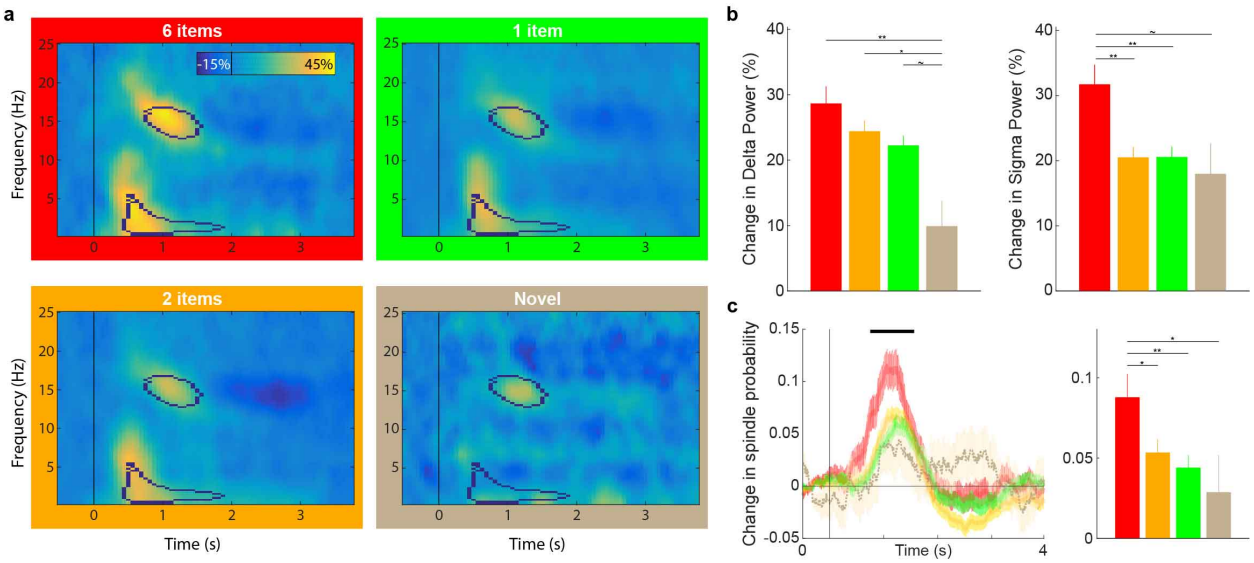


Figure 4: Sleep spindles and delta power following cue presentation was influenced by set-size. (a) Spectrograms displaying the time-frequency responses to cue presentations during sleep for different set-sizes. Time zero represents cue onset. Black outlines mark the identified delta cluster (encompassing K-complexes and slow waves) and sigma cluster (encompassing sleep spindles). (b) Changes in delta and sigma power as a function of set-size. (c) Time-locked modulation of spindle probability following cue onset for different set-sizes (left) and average differences in spindle probability modulation over the timeframe of the sigma cluster as a function of set-size (right). Error bars and shaded areas signify standard errors of the mean. Data from electrode location Cz. ** - p<0.01; * - p<0.05; ~ - p<0.07.

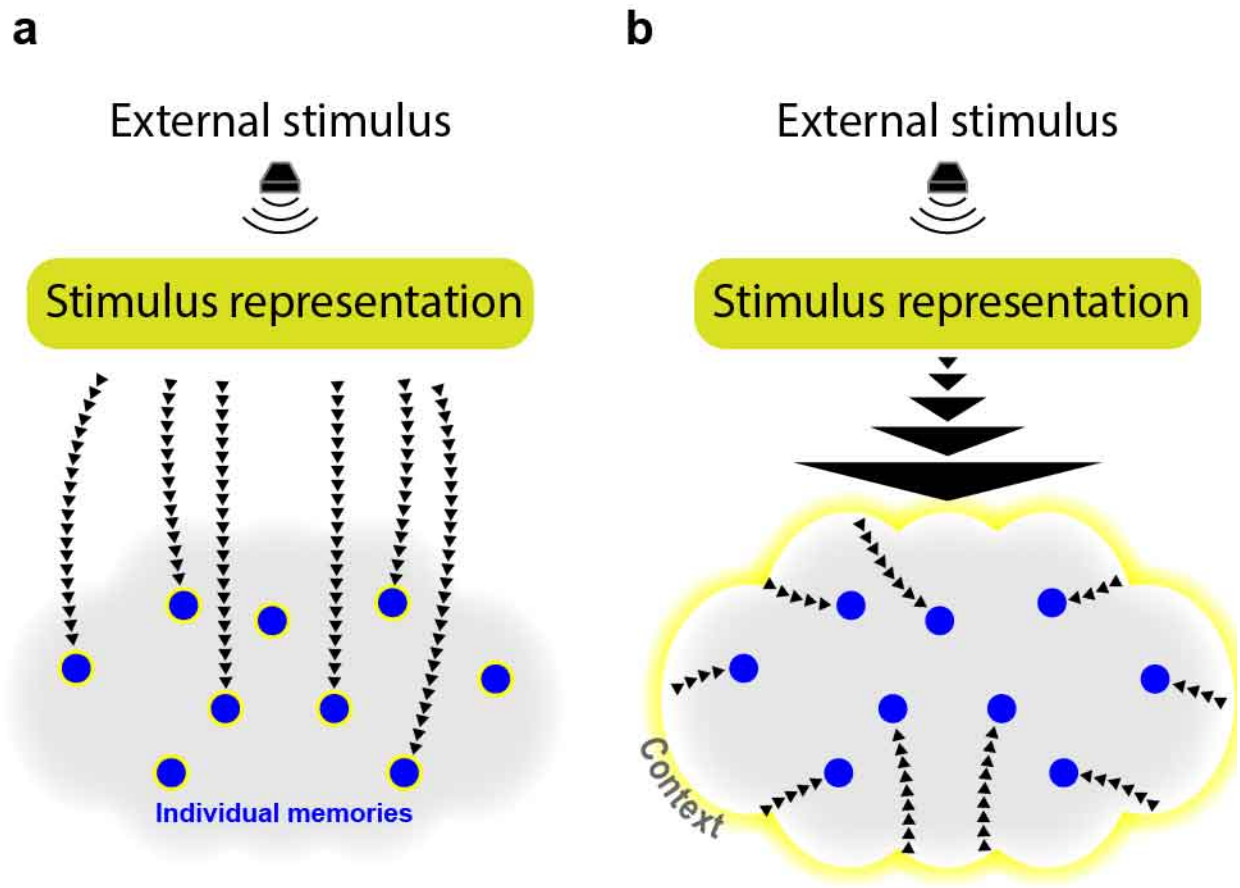


Figure 5: Models explaining the behavioral results, supporting the Parallel Reactivation Hypothesis. (a) In one model, individual memories are directly and simultaneously reactivated following stimulus presentation, regardless of the context in which they are embedded (e.g., locations of different cat images activated following a "meow" sound). Note that not all items in the context are reactivated (e.g., not all cat-related memories will be reactivated). (b) In contrast, in another model the stimulus reactivates a generalized context and its reactivation subsequently benefits individual memories embedded in the context.

Table 1: Themes used for sets of images

Theme	Used as
Balls	Two- or six-item sets
Fire	Two- or six-item sets
Birds	Two- or six-item sets
Documents and Literature	Two- or six-item sets
Cameras	Two- or six-item sets
Automobiles	Two- or six-item sets
Cats	Two- or six-item sets
Timepieces	Two- or six-item sets
Clothes with zippers	Two- or six-item sets
Dogs	Two- or six-item sets
Doors	Two- or six-item sets
Drinks	Two- or six-item sets
Food	Two- or six-item sets
Frogs	Two- or six-item sets
Heart	Two- or six-item sets
Kettles	Two- or six-item sets
Toilets	Two- or six-item sets
Trains	Two- or six-item sets
Phones	Two- or six-item sets
Musical Keyboard	Two- or six-item sets
Airplane	Two- or six-item sets
Cough	Two- or six-item sets
Flowers	Two- or six-item sets
Shoes	Two- or six-item sets
Pen	One-item sets
Trumpet	One-item sets
Violin	One-item sets
Monkeys	One-item sets
Kiss	One-item sets
Pen	One-item sets
Cow	One-item sets
Pig	One-item sets
Record	One-item sets
Money	One-item sets
Laugh	One-item sets
Toothbrush	One-item sets
Robot	One-item sets
Owl	One-item sets
Gong	One-item sets
Lobby bell	One-item sets
Drop	One-item sets
Boiling water	One-item sets
Slinky	Practice sets
Balloon	Practice sets
Computer keyboard	Practice sets

Methods

Participants

Participants had no known history of neurological or sleep disorders and claimed to be able to nap in the afternoon. They were asked to go to bed later than usual the night before the study, wake up earlier in the morning, and avoid any caffeine on the day of the study. Our sample included 40 participants, but 9 of these were excluded from the study because they were not exposed to each of the to-be-cued stimuli during sleep. The final sample included 31 participants (9 males, 21 females, and 1 non-binary person) between the age of 18 and 30 years (mean ± SD = 20.81 ± 2.96). The Northwestern University Institutional Review Board approved the procedure.

Materials

Visual stimuli were presented on a screen (1920 x 1080 pixels, P2418HT, Dell Inc., TX). Sounds were delivered using speakers (AX-210, Dell Inc., TX). Stimulus presentation was controlled by Neurobs Presentation (v17.2). Scripts and stimuli are available upon request.

Visual stimuli consisted of images of objects, parts of objects, or people, each shown as 125 x 125 pixels (34.6 x 34.6 mm). A total of 45 sets of pictures that shared the same theme were used (e.g., images of different cats; different parts of an airplane; Table 1). Each set consisted of main images for the spatial-memory task and lure images only needed for the item recognition task, as described below. Twenty-four of the sets included 6 main images and 12 lure images (“multi-item sets”). Twenty-one of the sets included one main image and two lure images (“single-item sets”). Each set was matched with a single, distinguishable, congruent sound with a maximal duration of 600 ms (e.g., a meow; a take-off sound).

The 24 multi-item sets were randomly assigned, per participant, to 18 two-item sets and 6 six-item sets. The rationale behind this random assignment was to avoid any systematic differences between six- and two-item sets that were independent of set-size. For each of the two-item sets, two main images and

four lure images were randomly chosen to be part of the set. Three of the single-item sets were practice sets that were used for initial training and also acted as fillers in pre- and post-sleep tests. A single, 500-ms sound not associated with any item was only presented during sleep ("novel sound").

Together, the 18 one-item sets, 18 two-item sets, and 6 six-item sets consisted of 90 different images. For the spatial task, each of these images was paired with a single location on a circular grid with a radius of 540 pixels (149.4 mm). The position for the center of each image was randomly selected to obey the following rules: 1) it was at least 50 pixels from the center of the grid; 2) it was at least 50 pixels from the external border of the grid; 3) it was at least 41 pixels from the location of any of the other 89 items; and 4) it was at least 400 pixels from any other item in its own set. The rational for this last rule was to allow us to disentangle swapping errors and accuracy errors (see below). However, because this rule is applied differently to sets of different size, changing the distribution of potential locations in a set-dependent manner, we randomly generated sets of six locations for each set, regardless of its real size. We then assigned all six locations to items in six-item sets and randomly chose two locations or one location for the two- and one-item sets, respectively.

The 90 items were split into six learning blocks of 15 items, so that items of the same set were never learned in the same block. Items belonging to sets of two items were always learned in sequential blocks (i.e., blocks 1-2, 2-3, 3-4, 4-5, or 5-6).

Procedure

After consenting to the study, the participant was fitted with an electroencephalography (EEG) cap. Next, a test was administered to measure pre-sleep response time (RT) in order to later evaluate sleep inertia (as described below). This RT task consisted of a red square that shifted between left and right positions at 10 Hz and finally stopped at one of the two locations. The participant was required to click the correct mouse button (i.e., left/right) before the square began flickering again. The task ended when

the participant responded correctly for eight out of the last ten trials. Initially, the square paused for 450 ms, but if the participant failed to reach the criteria within 30 trials, this duration was extended by 50 ms and the task restarted, iterating until the criterion was reached.

The RT task was followed by instructions and practice trials for the spatial-memory task. This task consisted of six blocks, each including 15 items. Each block started with exposure trials, in which the grid appeared with a sound. One second later, a corresponding image appeared in its location. The image disappeared 3 seconds later, coinciding with the offset of another presentation of the sound. Two seconds later, the grid disappeared, followed by a 1-second white-screen inter-trial interval (ITI), progressing through all 15 items.

Next, the positioning trials began. Each of these trials started with the simultaneous presentation of the circular grid, the item, and the related sound. The item was positioned in a random location at least 100 pixels from the true location. The participant used the mouse to drag and drop the item in its correct position. Trials were self-paced and every time an item was picked up or dropped the associated sound was presented. The participant signaled their choice by clicking the right button, which triggered a 3-second feedback screen with the image at its correct location. Some movement from the initial position was always required, and placements were considered correct if <100 pixels (27.7 mm) from the correct location. For incorrect placements, a red arrow linked the participant's choice with the correct location. After feedback, a 1-second white-screen ITI commenced, followed by the next trial.

Positioning trials ended when the participant responded correctly for each item twice in a row. When this criterion was reached for an item, it was no longer presented. Items were presented pseudo-randomly, with the only limitation being that items were not displayed twice in a row (unless only a single item remained).

After all six blocks were completed, the pre-sleep test began. It consisted of positioning trials for all 90 items (plus the three practice items, which were presented first). These positioning trials were identical to the positioning trials during training, except that no sounds were presented and no feedback was provided. The order of items in this stage was pseudo-random, with any two items belonging to the same set separated by at least two items from a different set.

Based on the results from this stage, half of the sets were chosen to be cued during sleep. Across all six-item sets, those chosen to be cued were balanced with the remaining sets by first minimizing differences between the numbers of incorrectly placed items. If those were equal between sets, the number of swaps and the absolute error in pixels were also balanced. The same procedure was also used for two-item sets and one-item sets.

Next, the futon chair on which the participant was seated was converted to a bed. During the ensuing 90-minute nap opportunity, white noise was presented over a set of speakers. When the participant entered slow-wave sleep (stage N3) a sequence of randomly ordered 22 sounds was repeatedly and unobtrusively presented: three six-item sounds, nine two-item sounds, nine one-item sounds, and one novel sound. The inter-cue interval was randomly chosen per presentation to be either 4.5, 5, or 5.5 seconds. The participant's EEG, electrooculography (EOG), and electromyography (EMG) were continuously monitored and sounds were discontinued when the participant showed signs of arousal. If cueing was not completed after 45 minutes and the participant was not in stage N3, cues were also presented during stage N2. Out of 31 participants, 5 reported upon awakening that they heard sounds during sleep. At the very end of the task, they were presented with all the sounds and asked to specify for each whether they heard it during sleep or not. Importantly, all five participants were all at chance level.

After waking up, the bed was converted back into a chair. Testing commenced at least 5 minutes after sleep offset. The participant was first required to reach criterion in the RT task. The time-window for response was based on their own responses in the pre-sleep RT task. Only one participant had trouble reaching that criteria, and was given an opportunity to freshen up before trying again, subsequently succeeding to reach criterion.

The participant then started a post-nap spatial-memory test, which was identical to the pre-sleep test. This testing was followed by an item-recognition task, in which the participant heard the sound for each set along with a set images and had to indicate for each image whether it was previously presented as part of the spatial-learning task or not. For each old image, two lures of the same semantic category were presented. Each image was presented twice. Data from this task were not used for this manuscript.

For debriefing, participants were asked whether they heard any noises during the nap and what they thought about the purpose of the study. They were then paid and dismissed.

Electroencephalography and polysomnography

EEG was recorded using Ag/AgCl active electrodes (Biosemi ActiveTwo, Amsterdam). In addition to the 64 scalp electrodes, contacts were placed on the mastoids, next to the eyes, and on the chin. All recordings were made at 512 Hz. Noisy channels were replaced with interpolated data from neighboring electrodes. Sleep scoring was based on the guidelines published by the American Academy of Sleep Medicine [53]. Scoring was approximated online while the participant was sleeping, and completed in full offline using the EEGLAB [54] and sleepSMG (http://sleepsmg.sourceforge.net) packages for Matlab (MathWorks Inc, Natick, MA). Offline scoring was done by two independent raters, both of whom were not privy to when sounds were presented. Any discrepancies were subsequently reconciled by one of

the two raters. For all analyses, only cues that were presented during NREM sleep (N2 and N3) were considered. Of these, an average of 84.8% (across participants) were presented during N3.

Sleep spindles were automatically detected using custom Matlab scripts. Data from electrode Cz during artifact-free periods of NREM sleep were used. The EEG signal was filtered between 11–16 Hz, and then the root mean square (RMS) was calculated at every time point with a moving window of 200 ms. Spindles were detected when the RMS crossed and remained above a threshold of 1.5 standard deviations of the signal for 500-3000 ms.

Analyses

Analyses consisted of ANOVAs, repeated-measures ANOVAs, paired and non-paired t-tests, Tukey post-hoc highly significant differences (HSD), correlation analyses, and interclass correlation (ICC) analyses, as described in the main text. For the spatial-memory task, our design allowed us to disentangle swapping errors and accuracy errors. For item placements in the pre- and post-nap tests, we found the closest correct position for any item within the same set. If that location was not the correct location of the placed item, the item was considered as swapped. Only non-swapped items were considered for the accuracy analyses. The rationale behind detecting and removing swap errors from accuracy estimates is that they inflate errors in a manner that does not purely reflect spatial-memory.

A caveat of our method of detecting swap errors is that pure guesses may also be considered as such error, if they happen to be closer to another image location than to the correct one. We were aware of this limitation, but preferred to overestimate rather than under-estimate swap errors to avoid their contamination of accuracy errors. However, this does raise the concern that our definition of swapping errors creates an unfair advantage in error rate for bigger sets. For example, a large error due to guessing (e.g., 800 pixels) would likely have been close to some other item in a six-item set and be

counted as a swap error. The accuracy error for this item would therefore not be considered. The same error for an item in a single-item set would not have another item to swap with and the accuracy error would have been considered. To compensate for this bias, swap errors were calculated by taking into consideration six locations, even for sets of one item (i.e., the same locations out of which the true item locations were drawn from, as described in Materials, were used in this stage).

The benefit of sleep was calculated for each item as the difference between the accuracy error in the post-nap spatial-memory test subtracted from the error in the pre-sleep test (i.e., positive values represent memory improvement). For the basic analysis of the effects of cuing and set-size (Figure 2), item benefits were averaged for each set-size and cuing status within participant. To confirm that our results were not due to differences in pre-sleep spatial error rates between sets of different sizes, we complemented our main analysis with two additional ones (Supplementary Figure 1). For the first analysis, we regressed out pre-sleep error rates by calculating the linear relationship between pre-sleep errors and forgetting (post-sleep – pre-sleep spatial accuracy). Then we subtracted each spatial forgetting score from the spatial forgetting expected from this linear relationship (i.e., the residual) and added back the mean raw spatial forgetting value to produce an adjusted accuracy error score. The results obtained using this corrected score are presented in Supplementary Figure 1a.

For the second analysis, 50% of trials were randomly chosen and considered for analysis for each participant and each set size. We next considered the pre-sleep accuracy error rates (averaged across individual items) for the datasets generated using this method, and eliminated all datasets in which the set-size main effect had $p<0.5$. Using this method, we collected 500 subsampled datasets in which the pre-sleep data was negligibly affected by set-size. Finally, we ran a repeated-measures ANOVA to calculate the cuing effect and the cuing by set-size interaction for these 500 datasets. The *p*-values obtained by these ANOVAs are presented in Supplementary Figure 1b.

For the ICC analysis (Figure 3a), significance of the difference between the benefit for cued and non-cued sets was assessed by applying a permutation test. The ICC was first computed for cued and non-cued sets separately and the difference between these measures was calculated. Then, the data was shuffled 10^6 times, so that the assignment of each individual item benefit to a specific set was random in each permutation. The shuffling was done separately for the cued and non-cued sets and the ICC was calculated for both. Then, *p*-values were assessed by comparing differences between real ICC values with the distribution of random differences.

For the spectral analysis (Figure 4a, b), we used a 300-ms time interval starting 500 ms before cue onset as the baseline. We first identified large clusters in the time-frequency response for electrode Cz by using *Z*-score baseline correction on the averaged data across participants, conditions (i.e., set-sizes), and repetitions, and choosing clusters with *Z*-scores above 1.96 (i.e., $p<0.05$, two-way). We next calculated spectrograms for each sound presentation, averaged them across condition, and considered the change in percentage relative to baseline. Using the two identified clusters (i.e., the delta range cluster and the sigma range cluster), we calculated an average modulation across time and frequency for each trial and each cluster, and grouped them according to condition (Figure 4b). The same 300-ms interval was used for baseline correction for spindle probabilities (Figure 4c).

Reference List

1. Heine R. *Über wiedererkennen und rückwirkende hemmung*. Johann Ambrosius Barth (1914).

2. Rasch B, Born J. About sleep's role in memory. *Physiological Reviews* **93**, 681-766 (2013).

3. Diekelmann S, Born J. The memory function of sleep. *Nature Reviews Neuroscience* **11**, 114-126 (2010).

4. Pavlides C, Winson J. Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *The Journal of Neuroscience* **9**, 2907-2918 (1989).

5. Wilson MA, McNaughton BL. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676-679 (1994).

6. Olafsdottir HF, Bush D, Barry C. The role of hippocampal replay in memory and planning. *Current Biology* **28**, R37-r50 (2018).

7. Deuker L*, et al.* Memory consolidation by replay of stimulus-specific neural activity. *The Journal of Neuroscience* **33**, 19373-19383 (2013).

8. Schapiro AC, McDevitt EA, Rogers TT, Mednick SC, Norman KA. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications* **9**, 3920 (2018).

9. Tambini A, Davachi L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19591-19596 (2013).

10. Alm KH, Ngo CT, Olson IR. Hippocampal signatures of awake targeted memory reactivation. *Brain structure & function* **224**, 713-726 (2019).

11. Oudiette D, Paller KA. Upgrading the sleeping brain with targeted memory reactivation. *Trends in Cognitive Sciences* **17**, 142-149 (2013).

12. Rudoy JD, Voss JL, Westerberg CE, Paller KA. Strengthening individual memories by reactivating them during sleep. *Science* **326**, 1079 (2009).

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

14. Antony JW, Gobel EW, O'Hare JK, Reber PJ, Paller KA. Cued memory reactivation during sleep influences skill learning. *Nature Neuroscience* **15**, 1114-1116 (2012).

15. Schönauer M, Geisler T, Gais S. Strengthening procedural memories by reactivation in sleep. *Journal of Cognitive Neuroscience* **26**, 143-153 (2014).

16. Schreiner T, Rasch B. Boosting vocabulary learning by verbal cueing during sleep. *Cerebral Cortex* **25**, 4169-4179 (2015).

17. Shanahan LK, Gjorgieva E, Paller KA, Kahnt T, Gottfried JA. Odor-evoked category reactivation in human ventromedial prefrontal cortex during sleep promotes memory consolidation. *eLife* **7**, (2018).

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

19. Diekelmann S, Biggel S, Rasch B, Born J. Offline consolidation of memory varies with time in slow wave sleep and can be accelerated by cuing memory reactivations. *Neurobiology of Learning and Memory* **98**, 103-111 (2012).

20. Antony JW, Cheng LY, Brooks PP, Paller KA, Norman KA. Competitive learning modulates memory consolidation during sleep. *Neurobiology of Learning and Memory* **155**, 216-230 (2018).

21. Vargas IM, Schechtman E, Paller KA. Targeted memory reactivation during sleep to strengthen memory for arbitrary pairings. *Neuropsychologia* **124**, 144-150 (2019).

22. Creery JD, Oudiette D, Antony JW, Paller KA. Targeted memory reactivation during sleep depends on prior learning. *Sleep* **38**, 755-763 (2015).

23. Faulkenberry TJ. Computing bayes factors to measure evidence from experiments: An extension of the bic approximation. *Biometrical Letters* **55**, 31-43 (2018).

24. Wagenmakers EJ. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review* **14**, 779-804 (2007).

25. Raftery AE. Bayesian model selection in social research. *Sociological methodology* **25**, 111-164 (1995).

26. Antony JW, Piloto L, Wang M, Pacheco P, Norman KA, Paller KA. Sleep spindle refractoriness segregates periods of memory reactivation. *Current Biology* **28**, 1736-1743 e1734 (2018).

27. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**, 30-46 (1996).

28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* **15**, 155-163 (2016).

29. Cairney SA, Lindsay S, Sobczak JM, Paller KA, Gaskell MG. The benefits of targeted memory reactivation for consolidation in sleep are contingent on memory accuracy and direct cue-memory associations. *Sleep* **39**, 1139-1150 (2016).

30. Cousins JN, El-Deredy W, Parkes LM, Hennies N, Lewis PA. Cued reactivation of motor learning during sleep leads to overnight changes in functional brain activity and connectivity. *PLoS biology* **14**, e1002451 (2016).

31. Cairney SA, Guttesen AAV, El Marj N, Staresina BP. Memory consolidation is linked to spindle-mediated information processing during sleep. *Current Biology* **28**, 948-954.e944 (2018).

32. Antony JW, Schonauer M, Staresina BP, Cairney SA. Sleep spindles and memory reprocessing. *Trends in Neurosciences* **42**, 1-3 (2019).

33. Polyn SM, Norman KA, Kahana MJ. A context maintenance and retrieval model of organizational processes in free recall. *Psychol Rev* **116**, 129-156 (2009).

34. Morton NW, Polyn SM. A neurocognitive theory of episodic and semantic interactions during memory search. (Under Review).

35. Anderson JR, Bower GH. Recognition and retrieval processes in free recall. *Psychological Review* **79**, 97-123 (1972).

36. Polyn SM, Norman KA, Kahana MJ. A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review* **116**, 129-156 (2009).

37. Howard MW, Kahana MJ. When does semantic similarity help episodic retrieval? *Journal of Memory and Language* **46**, 85-98 (2002).

38. Smith DM, Bulkin DA. The form and function of hippocampal context representations. *Neuroscience & Biobehavioral Reviews* **40**, 52-61 (2014).

39. Yonelinas AP, Ranganath C, Ekstrom AD, Wiltgen BJ. A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*, (2019).

40. Raaijmakers JG, Shiffrin RM. Sam: A theory of probabilistic search of associative memory. In: *Psychology of learning and motivation* (ed^(eds). Elsevier (1980).

41. Hasselmo ME. Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences* **3**, 351-359 (1999).

42. Gais S, Born J. Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2140-2144 (2004).

43. Klinzing JG, Kugler S, Soekadar SR, Rasch B, Born J, Diekelmann S. Odor cueing during slow-wave sleep benefits memory independently of low cholinergic tone. *Psychopharmacology* **235**, 291-299 (2018).

44. Bendor D, Wilson MA. Biasing the content of hippocampal replay during sleep. *Nat Neurosci* **15**, 1439-1444 (2012).

45. Manning JR, Norman, K. A., & Kahana, M. J. . The role of context in episodic memory. In: *The cognitive neurosciences, fifth edition* (ed^(eds Gazzaniga M). MIT Press (2014).

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

47. Huber R, Tononi G, Cirelli C. Exploratory behavior, cortical bdnf expression, and sleep homeostasis. *Sleep* **30**, 129-139 (2007).

48. Huber R, Ghilardi MF, Massimini M, Tononi G. Local sleep and learning. *Nature* **430**, 78-81 (2004).

49. Colrain IM. The k-complex: A 7-decade history. *Sleep* **28**, 255-273 (2005).

50. Voss U, Harsh J. Information processing and coping style during the wake/sleep transition. *Journal of Sleep Research* **7**, 225-232 (1998).

51. Blume C*, et al.* Preferential processing of emotionally and self-relevant stimuli persists in unconscious n2 sleep. *Brain and Language* **167**, 72-82 (2017).

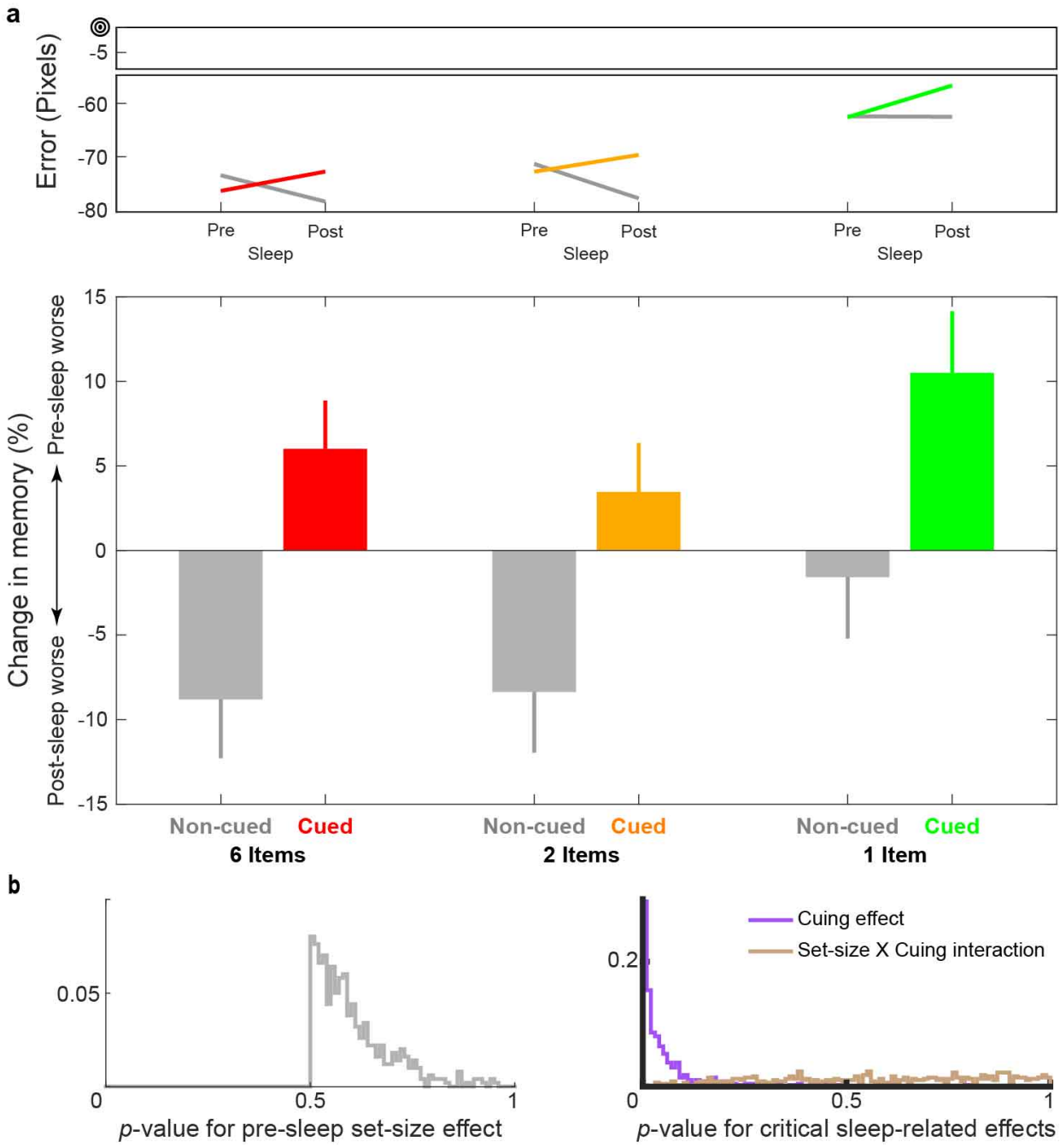
52. Andrillon T, Pressnitzer D, Leger D, Kouider S. Formation and suppression of acoustic memories during human sleep. *Nature Communications* **8**, 179 (2017).

53. Iber C, Ancoli-Israel S, Chesson A, Quan S. *Aasm manual for scoring sleep* (2007).

54. Delorme A, Makeig S. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods* **134**, 9-21 (2004).

Supplementary Materials include:

1. Supplementary Figure 1
2. Supplementary Table 1



<u>Supplementary Figure 1</u>: The effect of cuing and the null interaction effect between cuing and set-size are not a results of pre-sleep differences in errors between set-sizes. (a) To rule this option out, we regressed the pre-sleep error rates out. Considering the corrected data, the cuing effect is still significant (p<0.001) and the cuing-set-size interaction is not (p>0.89). (b) To complement this method, we subsampled out data in a way that would minimize the effects of pre-sleep difference (i.e., we chose

subsampled dataset that had a pre-sleep set-size difference with $p>0.5$, as shown in the histogram on the left). We then calculated *p*-values for the cuing effect and the cuing-set-size interaction for these datasets (histograms shown on right) and showed that the cuing effect was consistently significant whereas the interaction was consistently not.

Supplementary Table 1: No significant correlations between physiological measures and cuing benefit

	Delta power modulation		Sigma power modulation		Spindle probability modulation	
	r	Uncorrected *p*-value	r	Uncorrected *p*-value	r	Uncorrected *p*-value
One-item sets	0.03	0.14	0.02	0.26	0.02	0.23
Two-item sets	0.04	0.04	-0.02	0.43	-0.03	0.16
Six-item sets	0	0.89	0.06	0.09	0.01	0.77