Sleep promotes relational overlapping memories for long-term generalization

Matorina, Nelly[1], & Poppenk, Jordan[1,2,3*]

1. Department of Psychology, Queen's University
2. Centre for Neuroscience, Queen's University
3. School of Computing, Queen's University

* Corresponding author. 62 Arch St., Kingston, Ontario, Canada K7L 3N6, 613-533-6009, jpoppenk@queensu.ca,

Abstract

Human memory for recent events is believed to undergo reactivation during sleep. This process is thought to be relevant for the consolidation of both individual episodic memories and gist extraction, the formation of generalized memory representations from multiple, related memories. Which kinds of gist are actually enhanced, however, is the subject of less consensus. To address this question, we focused our design on four types of gist: inferential gist (relations extracted across non-contiguous events), statistical learning (regularities extracted from a series), summary gist (a theme abstracted from a temporally contiguous series of items), and category gist (characterization of a stimulus at a higher level in the semantic hierarchy) described in recent reviews. Sixty-nine participants completed memory encoding tasks addressing these types of gist and completed corresponding retrieval tasks the same evening, the morning after, and one week later. Inferential gist in a transitive inference task improved over the course of a week, whereas all other forms of memory were associated with decay. Higher proportions of REM and more spindles were associated with worse performance in a statistical learning task collapsed across time and after one week, respectively. We conclude that memory overlap is an insufficient condition for long-term generalization. Only structured relational memories, rather than associative memories (category and summary gist) or repeated co-occurrence (statistical learning) were enhanced over the course of a week. Furthermore, REM sleep may be involved in schema disintegration, which inhibits participants' ability to separate a stream of shapes into discrete units.

To better understand our world, we review the information contents of our experiences for patterns, extracting from those experiences the gist, or essential meaning (Brainerd & Reyna, 1990). "Gist extraction" of this kind can take time, and even sleep to emerge: participants who sleep typically demonstrate greater gist memory than those remaining awake for an equivalent period. This pattern has been observed in a variety of common paradigms, including transitive inference (Ellenbogen et al., 2007), statistical learning (Durrant et al., 2011), probabilistic category learning (Djonlagic et al., 2009), and false memory (Payne et al., 2009). One perspective on this gradual emergence of gist is that it reflects the qualitative reorganization of memories during sleep, where new memories emerge that were never learned directly (Landmann et al., 2014); and several researchers have proposed theoretical frameworks that this occurs either during slow-wave sleep (SWS; Durrant, 2011) or rapid eye movement (REM) sleep (Walker and Stickgold, 2010). However, not all forms of "gist" stand to benefit from such reorganization. For example, remembering the gist of a photograph (such as whether it was taken indoors or outdoors) does require extraction of category from the image, but the relation of the image to others is not essential for this to occur. We therefore asked the question: do common sleep mechanisms underlie extraction of various forms of gist, and do the various "gists" benefit differentially?

During SWS, slow oscillation field potentials (< 1 Hz) are temporally synchronized with thalamo-cortical spindles (7 – 15 Hz) and hippocampal ripples (140-220 Hz) at low cholinergic activity (Diekelmann & Born, 2010; Rasch & Born, 2013). This process is thought to reactivate memories in the hippocampus, promoting subsequent memory consolidation (Marshall & Born, 2007). Consistent with this "mere reactivation" view, Walker and Stickgold (2010) proposed that SWS supports veridical memory consolidation that keeps individual memories distinct. Under this framework, SWS should preferentially amplify gists that are internal to a stimulus, such as remembering category gist in the photograph example above, by virtue of better distinguishing it from other images seen the same day.

An alternative to this model, the information overlap to abstract (IOtA) theory, accounts for gist extraction through a process of repeated reactivation of overlapping memory elements. In combination with synaptic downscaling during SWS, this leads to strengthening of these shared elements (Lewis and Durrant, 2011). Consistent with this theory, several studies measuring gist extraction have shown that SWS is involved in gist extraction: duration of SWS correlated with improvement across sessions in a statistical learning task (Durrant et al., 2011); alpha EEG power during SWS was correlated with insight in a number reduction paradigm (Yordanova, Wagner, Born, & Verleger, 2012); and participants had more slow-wave activity after integrating words into an existing vocabulary (Tamminen, Ralph, & Lewis, 2013). These effects were specific to SWS and were not found with REM sleep or total sleep time (Durrant et al., 2011; Yordanova et al., 2009). Under the IOtA account, gist requiring integration of memories, such as transitive inference (a task requiring the stitching together of pair relations to make inferences above hierarchy), should preferentially benefit from SWS.

Adding further complexity, Walker and Stickgold (2010) proposed that it is REM that serves to integrate memories into associative networks. In particular, they proposed REM involvement in three forms of memory integration, all of which contribute to the construction of higher-order schemas: unionization of recent related items, assimilation of new items into established networks, and the abstraction of general rules. They argue that cortico-cortical processing in association areas during REM sleep, in combination with reduced noradrenaline, increased acetylcholine levels (Hasselmo, 1999) and theta wave oscillations (Jones & Wilson, 2005) might facilitate associative linking in different cortical

areas (Walker & Stickgold, 2010). They also suggest that gist extraction may take place across several nights of sleep after initial memory consolidation. Along these lines, higher immediate performance in a probabilistic learning task predicted that a greater percentage of the subsequent night spent would be spent in REM sleep (Djonlagic et al., 2009), suggesting that REM is upregulated when a sufficient amount of learning has occurred and promotes reactivation of the same circuits involved in learning.

Here, we took the novel approach of juxtaposing the effects of sleep on four forms of gist, each reflecting an apparently different gist extraction operation: inferential gist, which requires extraction of relations across non-contiguous events; statistical learning, in which regularities must be extracted from a series; summary gist, in which a theme is abstracted from a temporally contiguous series of items; and category gist, requiring characterization of a stimulus at a higher level in the semantic hierarchy. We define these four types of gist in Box 1, and provide visual depictions of the tasks we used to measure them in Figure 1. We reasoned that the role of REM or SWS in gist memory would be best elucidated by juxtaposing the effect of these sleep stages on gists that require extracting rules from relations (e.g., inferential gist and statistical learning), against those that require gist extraction from sets (summary and category gist; Stickgold and Walker, 2013). We anticipated that gist of each kind would either increase over time, or decay at a slower rate than detail memory measures that we gathered for sake of comparison. We did not have strong predictions as to whether SWS or REM would play more of a role in gist memories related to sets or relations.

We also took an individual differences approach: because false memory, transitive inference, and probabilistic learning have all been shown to have sleep effects compared to wake, we wanted to probe these effects deeper by looking only at individual differences in sleep stages. Having all of our participants in the sleep group increased our power to detect these effects. We predicted that participants with either more SWS or REM sleep would have better gist memory over time. We preregistered our hypotheses on the Open Science Framework (OSF; Matorina & Poppenk, 2019).

Method

Participants

104 participants were recruited in Kingston, Canada using posters, Facebook advertisements and web posts on Reddit, Kijiji and Craigslist. Participants were required to be: between 22-35 years of age (to avoid potential developmental effects); right-handed, an English native-speaker, have normal or corrected-to-normal vision and hearing, have no contraindications for MRI scanning, have no history of neurological disorders, sleep disorders, or recurrent mental illness that included medication, and not currently be taking psychotropic medications. Recruits meeting these criteria were invited to undergo an in-person screening session in which demographic eligibility was confirmed; a simulated MRI scanner was used to rule out claustrophobia and inability to keep still; and ability to perform above chance on a simple recognition memory test was established. Only participants who indicated that they regularly slept at least 5 hours a night on a computer administered questionnaire were included. Further details on the in-person screening session are provided in *Supplemental Materials*. Of the original sample, 69 participants passed our eligibility criteria and returned to complete the full experiment. The average age of participants who did so was 26.83 years (*SD* = 4.21 years). They described themselves as men (*n* = 30), women (*n* = 38), and other (*n* = 1). We used G*Power 3.1 to calculate a sensitivity power analysis, and found that based on our sample size of 69

participants and models with 3 predictors, we had 80% power to detect a small (0.09) effect size.

Procedure

After their screening session, participants wore a sleep EEG device (Sleep Profiler, Advanced Brain Monitoring, Carlsbad, CA) for a habituation night, during which they became accustomed to wearing and operating it. After the habituation night, on the basis of the recorded data, participants received corrective training as necessary.

Participants conducted an initial study phase during an evening visit, which was followed by a test phase (Pre-Sleep test). After sleeping at home while measuring their own sleep using a take-home sleep EEG device, they returned the next morning (12 hours after their initial session) for another memory test (Post-Sleep test). A week later, they returned for a final memory test (Day 7). They also completed an MRI session several weeks after testing.

The study session and each test visit included a study or test phase from each of four memory tasks (Fig. 1). These included a transitive inference (TI) task (inferential gist), a Deese, Roediger and McDermott (DRM) task (summary gist), a visual statistical learning (VSL) task (statistical learning), and a word-scene associative memory task (category gist). These tasks are described in detail below. After each study and test session, participants were encouraged to take a break. In addition, we administered an operation span task (OSPAN) at each session to measure participant fatigue. For TI and VSL tasks, participants are asked to learn a set of relationships among items. Due to the study time for these tasks, we could not test a different set at each time point. Therefore, we tested the same sets multiple times and also measured test-retest effect for these tasks during the Pre-Sleep session (see *Supplemental Materials*).

As the data collection effort involved collaboration with researchers investigating a variety of individual differences within our participant group, our participants also completed further experimental sessions involving objectives unrelated to the current study goals to be reported elsewhere.

Tasks

Visual statistical learning (VSL). VSL (Turk-Browne, Junge, & Scholl, 2005) provides a measure of statistical learning, which has been defined as implicit learning of patterns that are automatically segmented from a continuous environment (such as co-occurring shapes in a sequence; Turk-Browne, Scholl, Johnson, & Chun, 2010). To implement this task, we modified a publicly-shared script (described by Turk-Browne et al., 2005) obtained from the *Millisecond Test Library*. During learning, participants were shown four blocks of a stream of red and green shapes over 312 trials (each appearing for 400 ms with 200 ms post-trial duration) and instructed to attend to shapes of one color or the other, pressing a key whenever a shape of that color repeated. Two blocks were shown for each stimulus set, which were interweaved. At test, in each of 64 trials, participants were presented with two short series of three shapes familiar from the learning phase (each appearing for 400 ms with an inter-series interval of 1s) and instructed to select the temporal sequence that seemed more familiar. Some triplets were sequences seen in the previous learning task (e.g., ABC), and others were novel sequences (e.g., AEG). In our adaptation of this task, half of target triplets involved shapes that always appeared in the same order during training (i.e., deterministic sequences; e.g., ABC), whereas the rival sequence was never encountered (e.g., BAC). The other half of triplets sometimes also appeared in the rival sequence (i.e., non-deterministic sequences; e.g., ABC was

presented at a 3:1 ratio to BAC). The sequence that was presented more frequently was always considered the correct answer. An inter-stimulus interval (ISI) of 1s separated test trials.

Transitive inference (TI). The TI task (Ellenbogen, 2007) provides a measure of inferential gist, or the ability to make inferences across paired associates that form an implicit hierarchy (e.g., A > B > C > D). To implement this task, we again modified a script (described by Frank et al., 2004) obtained from the *Millisecond Test Library*. During study, participants were introduced to the hierarchy by being presented with four pairs of symbols (i.e., "premise pairs"; e.g., A and B), and asked to guess the winning symbol (e.g., A > B; Fig. 1B). For each premise pair, the winning symbol was counterbalanced on the left and right sides. Participants were given feedback for 150ms about the correctness of their guess after each study trial. Participants studied two stimulus sets, which were divided by a mandatory 12 second break with an option to resume anytime thereafter. After reaching a criterion of 75% correct responses or completing 200 trials, they completed 180 test trials without feedback testing both the original premise pairs as well as inference pairs (e.g., B > D). Participants completed 20 test trials per premise and inference pair (five premise pairs, two first order inference pairs (B > D and C > E), one second order inference pair (B > E), and one non-inference pair (A>F)). Test trials advanced following a response by the participant. Inference pairs are higher-order relationships that must be inferred from the relationships between lower-order relationships (e.g., correctly identifying that A > C would require inference from the premise pairs A > B and B > C). Correctly inferring such higher order relationships is taken as evidence for the formation of a superordinate hierarchy.

Word-scene associative memory (AM). The goal of the AM task was to provide gist and detail memory measures that were internal to an individual stimulus, rather than linked across a series or set of items. Our approach was to use single words to cue scene associates that could be characterized in various levels of detail. We probed four measures of memory for each studied item: super-category (general semantic category of scene, e.g., "hotel"), category (room type within category, e.g., "lobby"), instance (specific exemplar of a lobby), and detail (exact contents of the scene). Scene associates belonged to three super-categories (e.g., *house, restaurant, school*), which were all divided into 3 categories (e.g., *bathroom, bedroom and kitchen* for the 'house' category). Each category contained 3 instances (e.g., the room-type category *kitchen* contains 3 separate kitchens). Finally, each instance contained 3 detail variants. One detail image was the one originally studied and the other two have been edited to replace one object with a novel object. During the study phase, participants viewed each word-scene pair for 8 s under incidental learning instructions designed to bind the word cue and picture associate: in particular, to decide whether the "name" each room had been given could plausibly have been assigned based on an object found in that room. A 0.5 s ISI separated each trial. During test, participants were cued with the word associate for 1 s, the answered a cascade of two-option forced-choice questions about its associate, deciding on whether it is a word associated with a noise or room image, super-category, category, instance, then details. Each probe in this series of four probes was presented for 3.5 s with a pause of 0.3 s between each probe, and a 1 s ISI between each trial. For each multiple-choice question, lures were selected within the correct answer to the prior probe (e.g., if the correct super-category of an associate to a word was "house", the options for the category probe would be *bathroom, bedroom* and *kitchen*). Each probe was presented in visual format: e.g., house, restaurant and school icons were presented. In the instance probe, all options were

shown using a variant of the images that excluded the feature to be assessed in the final "detail" probe, so as not to reveal the correct answer in that final stage. During each of the three test sessions, we tested participants on three previously-untested super-categories to avoid test-retest effects.

Deese, Roediger and McDermott (DRM). The DRM task provides a measure of summary gist, the ability to extract meaning common to a set of related items. The difference between summary gist and inference gist is described by Stickgold and Walker (2013) as being akin to the difference between sets and relations, respectively. To implement this task, we again modified a publicly-shared script obtained from the *Millisecond Test Library*. During study, participants were asked to learn 27 lists of 14 related words each. Words were presented with a stimulus onset asynchrony of 1.5 seconds, with each list separated by 6 s break during which participants were asked to solve simple math problems. During test, participants were presented with 27 target words (old words from the study list), 9 critical lures (related words associated with the theme of each word list) and 36 unrelated lures (new words from an unstudied category). These words were presented one at a time in random order and trials only progressed when participants made a response. Based on the suggestion by Stahl and Klauer (2008) to distinguish gist and detail traces, participants were given response options, 'remember', 'related', or 'new'. Stahl and Klauer used these to model group estimates of several memory subtypes; however, as this approach failed when running models for individual participants, we instead modeled summary gist using a d'-based analysis (see *Supplemental Materials*).

Operation span (OSPAN). Lopez, Previc, Fischer, Heitz, and Engle (2012) found that the simulated flight performance of sleep-deprived Air Force pilots was predicted by OSPAN. Hence, it is often used as a suitable proxy for performance-related fatigue. Here, we used the OSPAN task to measure and control for variability in participant alertness across each test phase, using a task described by Unsworth, Heitz, Shrock, and Engle (2005) and obtained from the *Millisecond Test Library*. The task consisted of 5 practice letter questions, 15 practice math questions and 15 testing trials. Practice trials had no time limit and featured a 0.5s ITI. For practice letter questions, three to seven letters were shown in series on the screen. Each letter was presented for 1 second with a 200ms ISI. After presentation, participants were asked to enter the letters that they had seen in the correct sequence on the onscreen keyboard. Participants received feedback on their answers for 2 seconds. For practice math questions, participants were asked to complete math operations (e.g., (1*2)+1). They were instructed to complete the math problem as quickly as possible. Once they had an answer, they clicked the button to progress to the next screen. A possible answer was displayed on screen and participants were asked to indicate whether the answer is 'true' or 'false'. Practice trials were used to calculate the mean time that it takes a participant to solve math problems. For test trials, letter and math questions were intermixed, and there was a limit of the participant's average math problem time plus 2.5 SDs before the trial progressed on its own (to discourage letter rehearsal). Each test block consisted of both letter and number trials. There were three blocks of each of the five set sizes (i.e., there were three repetitions of 3, 4, 5, 6, and 7 letter and number sizes). The shortened OSPAN consisted of five practice letter questions, five practice math questions and one block of each of four set sizes (i.e., one repetition of 3, 4, 5, and 6 letter and number sizes).

Apparatus

Participants completed three sessions individually in a testing room. One task was executed using MATLAB with Psychtoolbox (Kleiner et al., 2007) and SuperPsychToolbox (Mountjoy & Poppenk, 2015). The remaining three tasks (OSPAN, DRM, VSL, and TI) were executed using Inquisit (Version 5.0.6.0, 2016). For sleep stage measures, we used the Sleep Profiler, a single-channel electroencephalography (EEG) device worn on the forehead that records at 256 Hz from three sensors at approximately AF7, AF8, and Fpz (Lucey et al., 2016; Sleep Profiler Scoring Manual, 2015). The device applies a 0.1 Hz low-frequency filer and a 67 Hz high-frequency filter. Expert review of Sleep Profiler data and concurrently-collected data from a full polysomnography (PSG) net (which is broadly regarded as a "gold standard" for sleep data collection) has previously resulted in strong agreement for total sleep time (*ICC* = 0.96) and REM sleep (*ICC* = 0.92), and poorer agreement for Stage 1 (*ICC* = 0.66) and Stage 3 (*ICC* = 0.67; Lucey et al., 2016). Agreement for Stage 3 increased when combined with Stage 2 into a non-REM (NREM) measure (*ICC* = 0.96).

The Sleep Profiler system also features automated sleep-staging software, which we manually validated using a subset of raw sleep data (described below) gathered using our Sleep Profiler. Briefly, as described by Levendowski and colleagues (2017), the software first rejected 30-second epochs where the absolute amplitude was ≥ 500 μV, applied a notch filter, and then an infinite impulse response band pass-filter to obtain 16 Hz samples of the power values for delta (1-3.5 Hz), DeltaC (delta power corrected for ocular activity), theta (4–6.5 Hz), alpha (8–12 Hz), sigma (12–16 Hz), beta (18–28 Hz), and EMG bands (> 40 Hz with a 80 Hz, 3 dB rolloff). Another set of power values was derived after application of a 0.75-Hz high-pass filter. Both filtered and unfiltered power spectra were used to stage sleep. If at least 15 s of valid data was available, AF7-AF8 channels were used for scoring, followed by AF7-Fpz and AF8-Fpz. When either of the latter were used, power spectra were increased to compensate for signal attenuation due to shorter interelectrode distances.

Power spectra averaged from 16 to 4 Hz were used to detect sleep spindles, which were characterized by spikes in absolute and relative alpha and sigma power that met empirical thresholds. Spindles were at a minimum 0.25 Hz in length with no maximum. To reduce misclassifying spindles, beta and EMG power bands had to be simultaneously surpassed relative to the alpha and sigma power.

For details on acquisition and analysis of hippocampal volumes, see *Supplemental Materials*.

Sleep Profiler validation

We selected a random subset of ten participants and used their raw sleep EEG data for validation of the Sleep Profiler auto-staging. Two independent raters rated the data and we took the average of their ratings. Prior to scoring, we pre-registered on OSF that if intra-class correlations (ICCs) were good (between 0.75 and 0.9) or excellent (over 0.9) for SWS, REM sleep, and TST (Koo & Li, 2016), we would use the sleep stage values derived from the automated scoring system. If they were below 0.75, then we would manually score the sleep data. We used an ICC measure that estimates the degree of consistency across measurements rather than absolute values (McGraw & Wong, 1996).

ICCs between the automated scoring system and two independent raters, as well as inter-rater ICCs, are given in Table 1. Human rater agreement with one another was excellent for TST and REM, whereas their agreement with the algorithm was lower, at a level closer to 0.8 (good). Human rater agreement with the algorithm was also good for SWS, whereas SWS agreement among human raters was only moderate. Having observed consistently good algorithm classification performance (with all ICCs above a

cutoff of 0.75), we therefore used the sleep stage values derived from the automated scoring system for all of our analyses. We did, however, perform a manual quality control inspection of the sleep data to remove 1) any recordings that contain no SWS or no REM, 2) partial recordings that did not contain the full night of sleep, and 3) recordings where a participant's total sleep time was less than 50% of how much they typically reported sleeping on weekdays. This resulted in the omission of data from four participants.

Data analysis

As preregistered, we excluded data falling three median absolute deviations above or below the mean of each variable. To investigate relationships among gist and detail variables, we ran Pearson correlations. To investigate predictors of patterns of change in gist and detail over the course of a week, we used multi-level modelling (MLM) implemented in the nlme package in R (R Core Team, 2016; Pinherio, Bates, DebRoy, Sarkar, & Team, 2007). Multi-level modelling allowed us to estimate an individual change function in each participant, as well as predictors of this change function. Therefore, this analysis is especially suitable for looking at individual differences in change over time (Nair, Czaja, & Sharit, 2008). In the models, we estimated each participant's change over three time points at Level 1 and individual differences at Level 2. Our predictors were Time, as well as Level 2 predictors related to individual differences in sleep and hippocampal volumes. Time was coded as 0 for Pre-Sleep, 1 for Post-Sleep, and 2 for Day7. In all multi-level models, we included grand-mean centered hippocampal volumes (aHPC and pHPC) or sleep stages (proportion of SWS or REM) as predictors. We also included fatigue (OSPAN) as a control variable. Outcome variables were measures of gist memory on four different tasks at 3 time points: in the same evening as study (Pre-Sleep), the next morning (Post-Sleep), and one week later (Day7). To balance control for multiple comparisons with experimental power, we selected a 0.05 False Discovery Rate (Benjamini & Hochberg, 1995). Lastly, to investigate whether gist and detail measures decayed at different rates, we conducted two MANOVAs. For the AM and DRM tasks, we conducted a MANOVA for 3 time points and 3 dependent variables, followed up by ANOVAs. For the TI task, we conducted a MANOVA for 3 time points and 2 dependent variables, followed up by ANOVAs.

Results

The goal of our study was to investigate the influence of sleep, including specific sleep stages, on different forms of gist, juxtaposing those requiring rule extrapolation from relations against those requiring gist extraction from sets. We approached this question both within-subjects – evaluating relative change in memory over time – as well as between-subjects, evaluating the predictive power of sleep variables over gist extraction. We began our between-subjects analysis by investigating correlations among our memory variables to identify possible correspondence among memory types. We then compared gist and detail memory variables within the same tasks using a MANOVA. Afterwards, we investigated patterns of change in gist and detail memory as a function of time and sleep stages. As pre-registered, we also ran a version of the model incorporating hippocampal volumetric predictors (see *Supplemental Materials*), but as these were not predictive of memory performance, we re-ran our model without a hippocampal factor and present it in this form for simplicity. We investigated models with both interactions over time, as well as models collapsed over time. Descriptive sleep statistics are given in Figure3A. Before addressing our research questions, we first discuss baseline models.

Memory predictors

We analyzed only variables in which participants achieved above-chance and off-ceiling performance as of Pre-Sleep, to ensure that our measures were sensitive to change in both a positive and negative direction. All performance measures were off ceiling and off floor. One-sample t tests showed that all of the outcome variables in AM, TI, and VSL were significantly above chance and below ceiling as of Pre-Sleep, all *p*s < .005. One-sample t tests showed that memory for summary gist (DRM) was significantly above 0 for Pre-Sleep, $t(68) = 5.638$, $p < .001$, but not at Post-Sleep, $t(67) = 1.202$, $p = .234$. At Day7, summary gist (DRM) was significantly *below* a chance score of 0, $t(66) = -8.800$, $p < .001$. Memory for individual words (DRM) was not significantly above 0 for any time points, so we excluded this variable from subsequent analyses.

To explore possible relationships and redundancies among our behavioural memory measures, we inspected a correlation matrix, focusing on relationships involving measures from the same task (Fig. 3A), and to a more limited extent, across tasks. Because the goal of this step was to limit our predictors rather than form inferences about tasks, we did not apply multiple-corrections comparisons in this step. As there were strong correlations in all tasks, to simplify our analyses, we combined gist measures in our different tasks: first-order and second-order inference memory into an inferential gist measure; non-deterministic and deterministic sequence memory into a statistical learning measure; and super-category, category, and summary gist into an associative gist measure. As a result of this step, the remainder of our analyses concerned only three variables: summary gist, inferential gist, and associative gist.

Special consideration is owed to collapsing of the VSL task. Fuzzy Trace Theory, one of the prominent frameworks distinguishing between gist and detail memory, defines detail as information that was tested in the exact form it was learned (Brainerd & Reyna, 1990). Given this definition, we initially felt deterministic sequence memory could be considered a detail measure and non-deterministic sequence memory a gist measure, which is the framework we followed in our preregistration. However, as both deterministic and non-deterministic sequences require extracting patterns from a sequence, and as we found them both to be correlated in the analysis above, we feel they are both better-represented as gist measures. Non-deterministic sequences additionally require participants to sift through more noise than deterministic sequences, so can be considered higher on a hierarchy of gist. A better detail measure for the task would have been a measure of participants' memory for individual shapes (not measured in the current analysis).

Hence, we were left with three putative detail memory predictors: a "detail" probe in our AM test (scene-cued object recognition), memory for premise pairs (TI recognition of explicitly studied shape relations), and individual word recall (DRM). As we excluded individual word recall from analysis due to low performance (as discussed above), We were left with an AM and TI measure only. We performed MANOVA analyses to assess divergence of gist and detail over time for these two tasks. For AM, there was a significant MANOVA for Time differences across associative gist, instance and detail memory, $F(2,179) = 20.157$, $p < .001$. Follow-up ANOVAs indicated significant decay rates in associative gist, $F(2,179) = 66.998$, $p < .001$, and instance memory, $F(2, 179) = 27.959$, $p < .001$, but not detail memory, $F(2, 179) = 2.604$, $p = .077$. For TI, the MANOVA for time differences across premise pair memory and inferential gist memory was not significant, $F(2, 195) = 1.692$, $p = .151$.

Baseline Multi-Level Models. We next constructed a baseline model to compute Intraclass Correlation Coefficients, which indicate the amount of variance in dependent

variables that can be attributed to individual differences (and that can therefore be reasonably attributed to individual difference predictors, such as sleep). Variables with 10% or more individual differences variance are considered to have meaningful individual differences. The ICCs were $\rho = 0.825$ for statistical learning (VSL), $\rho = 0.000$ for associative gist (DRM and AM), and $\rho = 0.662$ for inferential gist (TI). This means that, across different memory measures, between 0 and 83% of the variance in memory variables can be attributed to individual differences. A low intraclass correlation coefficient for associative gist may be due to large within-subject changes over time. Because most of the models are above threshold, we consider there to be significant individual differences in the memory variables we measured. We therefore proceeded with multi-level modelling as planned.

Gist Memory over Time. Category gist significantly decayed over time, from Pre-Sleep to both Post-Sleep and Day7 (*p*s < .001). Statistical learning significantly decayed between Pre-Sleep and Day7 ($p = .022$), but there was no significant decline as of Post-Sleep ($p = .652$). By contrast, Inferential gist was significantly enhanced over time from Pre-Sleep to Post-Sleep ($p = .007$) and Day7 ($p = .0005$). After correcting for FDR, we did not find any significant effects of OSPAN, *ps* > .009.

Sleep. Participants reported an average weeknight sleep of 7 hours and 14 min (*SD* = 62 min). Average total sleep time for Pre-Sleep test was 5 hours and 59 min (*SD* = 91 min). Regarding our confirmatory analyses, proportion of REM negatively predicted statistical gist collapsed over time, or controlling for time, $b = -0.857$, $t(59) = -2.907$, $p = .0051$. After correcting for FDR, we didn't find any significant SWS or TST effects, *p*s > 035. Regarding our exploratory analyses, spindles across all sleep stages negatively predicted statistical learning at Day7, $b = -0.000$, $t(118) = -3.327$, $p = 0.0012$. We did not find any significant NREM or Stage 2 effects, *p*s > .015. For the purposes of future hypotheses, we note a positive trend toward REM predicting inferential gist collapsed over time, $b = 0.758$, $t(59) = 1.776$, $p = .081$. This may reflect a small effect that we did not have enough power to detect, and should only be treated as an exploratory report that requires future studies with larger sample sizes to confirm the effect. We also note a positive trend towards proportion of NREM sleep positively predicting statistical learning, $b = .671$, $t(63) = 2.506$, $p = .015$.

Discussion

The goal of this study was to investigate the relationship of time and sleep stages on various forms of gist memory. We found a clear dissociation between inferential gist, which was enhanced over the course of a week, and the decay associated with statistical learning and associative gist. Thus, inferential gist memories may undergo qualitative changes over the course of (at least) a week to allow for the formation of new memory content that was never learned directly. We also found that REM sleep negatively predicted performance on statistical learning collapsing over time. Thus, REM sleep may inhibit participants' ability to extract a pattern from a series.

Taking Lewis & Durrant's (2011) model of overlapping memory representations as a general gist extraction process, we conclude that memory overlap is an insufficient condition for long-term generalization. In our study, only structured relational memories (inferential gist), where each item was related to each other item, were enhanced. This type of memory contrasts with associative memories (associative gist) and repeated co-occurrence in statistical learning, which could involve discrete sets of overlapping memories organized into triplets. Inferential gist likely involves the formation of an

overarching structural or schematic framework, such that any inference or premise pair can be referenced to this framework to determine the correct response. This may be a more similar structure to insight tasks (Wagner, Gais, Haider, Verleger, & Born, 2004) and artificial grammars (Gomez, Bootzin, & Nadel, 2006).

Our finding that inferential gist is enhanced over sleep and time is consistent with previous research comparing sleep and wake groups (Ellenbogen et al., 2007; Lau et al., 2010). Although it is possible that the increase in memory is due to repeated testing, we did not find a significant test-retest effect when testing twice during the Pre-Sleep session (described in *Supplemental Materials*). Furthermore, each task was presented to participants in a counterbalanced order, and no one task was deemed more important, so it is unlikely that participants selectively encoded information from transitive inference above other tasks. In our data, we did not find any significant predictive effects of sleep stages on inferential gist at specific time points or collapsed over time. Speculatively, the absence of stimuli during sleep could be what promotes higher inferential gist memory in sleep rather than wake groups. We do note, however, that we observed a trend towards a positive REM prediction.

We also found that statistical learning was retained over a night of sleep, such that there was no significant change between Pre-Sleep test and Post-Sleep test. Previous research found higher scores in sleep compared to wake groups for probabilistic learning (Djonlagic et al., 2009), indicating that statistical learning might be protected from decay rather than enhanced over time. One distinction between statistical learning and associative gist is that sequences in statistical learning were repeated over time, perhaps allowing for the formation of a schema that protects the memory from decay. Overall, structured relational memories formed *new* memory content over sleep, repeated cooccurring memories were retained after sleep, and associative memories decayed over sleep.

Sleep Stages

Proportion of REM negatively predicted statistical learning collapsing over time, an effect not found with total sleep time. As this effect was found collapsing over time rather than interaction with a specific time point, one possible explanation is that a waking mechanism analogous to REM sleep exists that inhibits statistical learning at initial test and other time points. As mentioned above, Landmann et al. (2014) propose that REM sleep is responsible for a process called schema disintegration, which disbands existing schemas and allows for creativity. Considering this hypothesis, participants with higher proportions of REM sleep may have weaker statistical learning schemas. In contrast to previous research, we did not find any evidence that SWS predicts inferential gist, statistical learning, or associative gist (Lau et al., 2010 for transitive inference; Durrant et al., 2011 for statistical learning). We also did not find evidence to support the theory proposed by Landmann et al. (2014) that schema integration (of which inferential gist is a special case) and schema formation (statistical learning) takes place during SWS or NREM sleep. We do note, however, a positive trend towards NREM sleep predicting statistical learning. Hence, an active sleep mechanism to retain statistical learning may exist in addition to a negative (REM) inhibitory mechanism.

In our exploratory analyses, we found that spindles negatively predicted statistical learning at Day7. In other words, if participants had more spindles the night following encoding, they performed more poorly on statistical learning one week later. Tamminen, Payne, Stickgold, Wamsley, and Gaskell (2010) found evidence that spindles are associated with the integration of new memories. Given this framework, it's possible that

integration processes associated with spindles inhibited participants' ability to retain statistical learning.

Relationships among Gist and Detail Memory Measures

Looking at the relationships among gist and detail measures, we found within-task correlations across gist and detail memory in all of our tasks. Hence, gist memories may be dependent on detail memories (or vice versa). Integrating our results with previous classification systems, Stickgold and Walker (2013) distinguish between gist extraction from sets, and rule extrapolation from relations. Inferential, statistical learning would fall under rule extrapolation, whereas category gist and summary gist would fall under gist extraction. Rule extraction measures across tasks (inferential gist and statistical learning) are not correlated. However, gist extraction measures (category and summary gist) are correlated. Thus, it does not seem that individual differences in ability to extract rules is consistent within individuals, but individual differences in the ability to extract gist is consistent.

We also found correlations between premise pair memory and associative memory measures, suggesting an associative cognitive component involved in learning premise pair relationships (two shapes in TI, words and scenes in AM). This idea is also consistent with the lack of a correlation between premise pairs and detail memory, as during detail questions participants had to distinguish only the correct detail arrangement they had seen, and did not need to retrieve the word-scene association. Lastly, summary gist memory was correlated statistical learning and associative memory, which were not correlated with one another. Summary gist has a gist extraction component, as well as a temporal component as during study, individual words move quickly on the screen. VSL required participants to watch a quickly moving sequences of shapes and consolidate them immediate into patterns, which may be the cognitive component correlated to summary gist. Of note, DRM is a verbal task, while AM and VSL are visual tasks, so the correlations are not likely to be due to visual compared with verbal memory types.

Brainerd and Reyna (2005) suggested that gist traces are more resistant to forgetting than detail traces. Comparing gist and detail traces within the same task, associative gist (AM and DRM) decreased by 77%, whereas instance memory (AM) decreased by 16% and detail (AM) by 9%. Our omnibus MANOVA was significant, with follow-up test indicating significant decreases only in associative gist and instance memory. Hence, associative gist traces are more rapidly forgotten over time than detail traces and detail may be preferentially consolidated in this task. This goes against our earlier hypothesis that gist memory would be more resistant to forgetting than detail memory. Our omnibus MANOVA for transitive inference was not significant, so we did not find any evidence that inferential gist memory (TI) is more resistant to decay than premise pairs (TI).

Hippocampal Volumes

After correcting for FDR, we did not find any significant hippocampal effects. Previous studies have found activation in the right aHPC during overlapping transitive inference pairs (Heckers, Zalesak, Weiss, Ditman, & Titone, 2004), as well activation in the aHPC in more distant (compared to more proximal) premise pairs (Collin et al., 2015). Thus, although inferential gist is likely to be encoded in the aHPC, larger aHPC volumes do not seem to predict better inferential gist memory. Previous research also found activation in the right hippocampus during VSL (Turk-Browne, Scholl, Chun, Johnson, 2009). A critical distinction between these studies and ours is that we measured individual

differences in hippocampal volumes rather than activation in the aHPC and pHPC, which could explain these differences.

The current design did not allow us to pinpoint the dynamic changes in hippocampal activation and sleep over time. Although we were able to behaviorally measure different kind of gist and relate those to individual differences in hippocampal volumes, there may be time-dependent changes in the activation of the aHPC and pHPC. For instance, in a recent paper, Dandolo and Schwabe (2018) found that activity in the aHPC significantly decreased over time and was related to memory specificity at encoding, whereas pHPC activity remained the same. Tompary and Davachi (2018) found that feature overlaps emerged in the hippocampus over the course of a week. Individual differences in hippocampal volumes do not allow us to see these time-related changes. Future studies could test within an fMRI over three sessions to relate activation in the aHPC to pHPC to time-dependent memory changes. Furthermore, diffusion tensor imaging (DTI) scans could determine the degree to which water diffuses within the aHPC, whether this density is related to gist memory and whether degree of diffusion is related to aHPC volume.

Limitations

One consideration in our design is that gist or detail memory traces could have been retrieved based on how memory questions were phrased in different tasks. For instance, in DRM and AM, participants were explicitly asked to select the correct answer, exactly how it was studied, thus being probed for a detailed response. However, in VSL, participants were asked which sequence was more familiar, thus possibly eliciting a gist memory trace. In future studies, when comparing different kinds of gist, memory tasks should be worded such that they are asking for a similar kind of representation. We chose the wording of memory questions in this study to be consistent with previous studies, to easily compare our results with previous research. Altering the phrasing of VSL questions to “Which exact sequence did you see?”, for example, could be associated with a number of other unexpected changes.

Conclusion

We conclude that memory overlap is an insufficient condition for long-term generalization. Only structured relational memories, rather than associative memories (category gist) or repeated co-occurrence (statistical learning) were enhanced over the course of a week. Furthermore, REM sleep may be involved in schema disintegration, which inhibits participants’ ability to separate a stream of shapes into discrete units.

Box 1. Definitions

Gist Memory	A coarse-grained memory representation that captures the essential underlying features of an experience.
Detail Memory	A fine-grained memory representation that captures specific surface-level features of an experience.
Types of gist	
Statistical Learning	Implicit learning of patterns or regularities that are automatically segmented from a continuous environment.
Inferential Gist	Extraction of relations across non-contiguous events, or extraction of higher-order relationships from lower-order ones.
Summary Gist	Abstraction of a theme from a temporally contiguous series of items. Summary gist encapsulates the circumstance commonly observed in a false memory paradigm, such as the word list task employed here, where a category-consistent word is erroneously believed to be present. Multi-item gist extraction or multi-item generalization have been alternate terms used to describe this idea by Stickgold and Walker (2013).
Category Gist	Characterizion of a stimulus at a higher level in the semantic hierarchy than was experienced.
Related concepts	
Gist Extraction	Scanning of information and extraction of patterns or essential meaning. This can be an immediate or gradual process.
Schema	We use the distinction put forth by Robin and Moscovitch (2017): a gist representation is specific to a single episode, whereas a schema is a more abstract representation from similar episodes. Participants in our study experience only one study phase, so we refer to the resulting memories as gist memories.

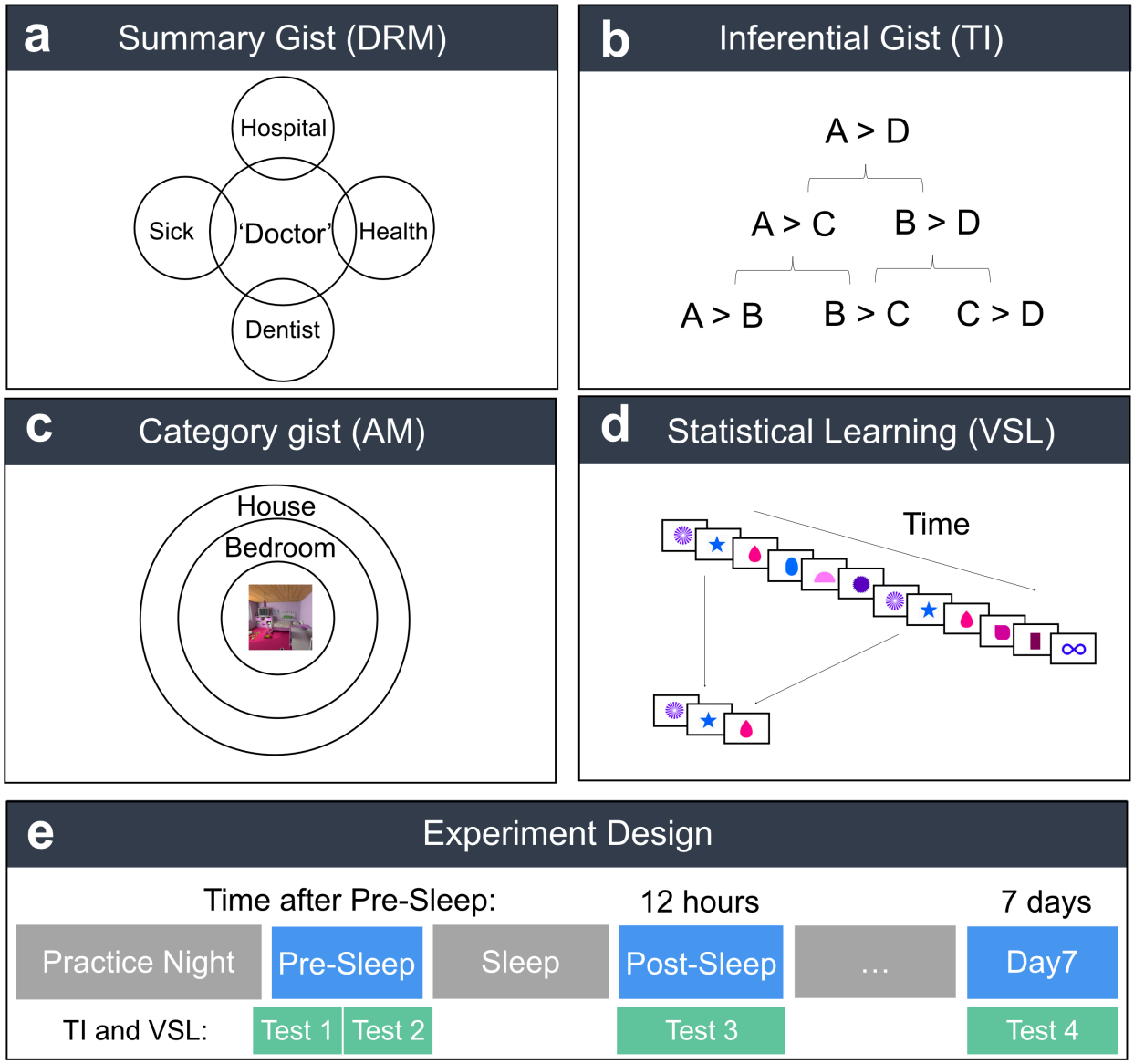


Figure 1. Multiple conceptualizations of gist. a) Conceptual illustration of overlapping meaning between studied words and the unstudied "gist" word in the DRM task. b) Conceptual illustration of studied premise pairs on the bottom row, unstudied first-order inferences on the second row, and unstudied second order inferences on the top row. c) Conceptual illustration of gist internal to an individual stimulus. Participants studied word-scene pairs, and were then asked about the scene's category (e.g., bedroom) and super-category (e.g., house). d) The left-hand panel shows a sequence presented over time, and one triplet which is repeated in the sequence. e) Study design. Participants first completed a practice night with the Sleep Profiler. Then they completed a study and first test session (Pre-Sleep test), which included two test-retests, Test 1 and Test 2 for TI and VSL tasks. Twelve hours and 7 days later participants completed a second (Post-Sleep Test) and third (Day 7) test session, respectively.

Table 1
Intra-class Correlations between the Sleep Profiler Automated Scoring System and two Independent Raters

Sleep stages	Intra-class correlations (ICCs)	Inter-rater ICCs
TST	0.855	0.982
SWS	0.819	0.669
REM	0.798	0.988

Table 2
Memory performance at Pre-Sleep.

Memory Variable	Mean accuracy	Standard Deviation
Premise Pairs	0.685	0.136
First Order Inferences	0.582	0.232
Second Order Inferences	0.623	0.277
Deterministic Sequences	0.676	0.205
Non-Deterministic Sequences	0.659	0.177
Super-Category	0.770	0.134
Category	0.818	0.128
Instance	0.914	0.095
Detail	0.662	0.125
Summary Gist	0.662	0.949

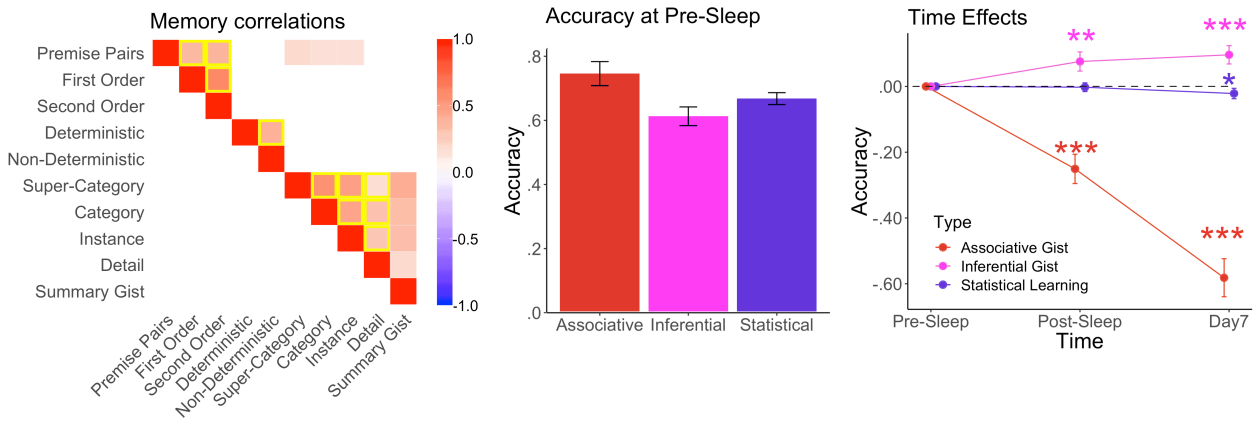
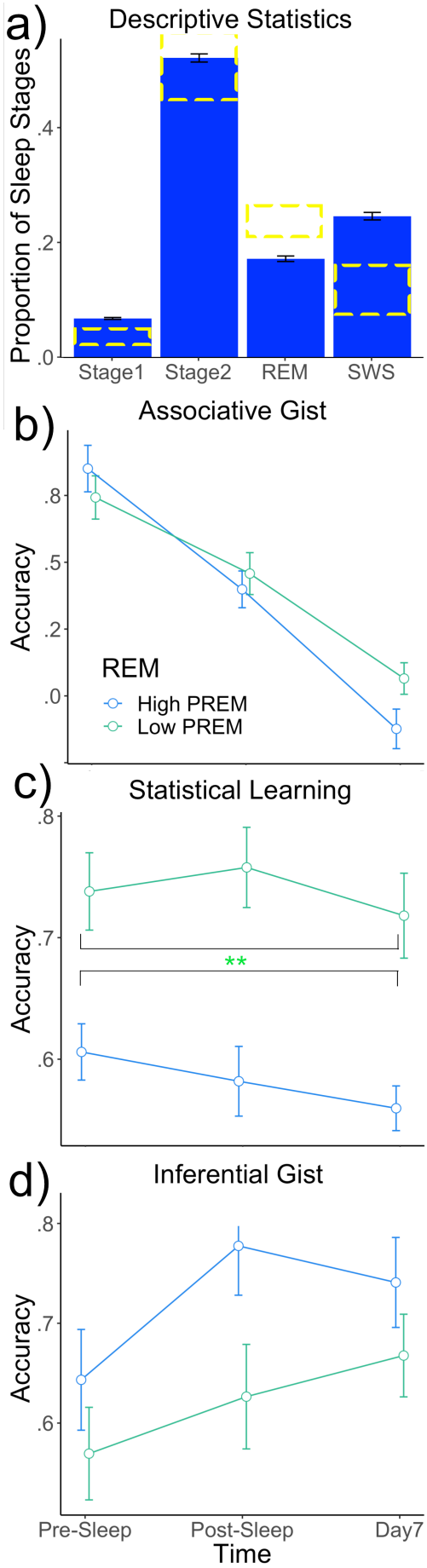


Figure 2. a) Correlations among Dependent Variables. Yellow boxes indicate potential relationships among variables that are within the same task. Missing values indicate non-significant relationships. b) Time effects for all three dependent measures. Associative memory significantly decreased from Pre-Sleep test to both Post-Sleep test and Day7. Inferential gist increased from Pre-Sleep test to both Post-Sleep test and Day7. Statistical learning significantly decreased from Pre-Sleep test to Day7.

Figure 3. a) Bar plot showing descriptive statistics in this study compared to general population norms derived from Carskadon and Dement (2011). b-d) High and low REM groups across statistical learning, associative gist, and inferential gist measures. Proportion of REM significantly negatively predicted statistical learning collapsing over or controlling for time.



Supplemental Materials

Description of pre-registered hippocampal analysis

In-person screening session

Participants were advised that the session would last approximately two hours, and that they would be compensated CAD$20 for their time (or a pro-rated amount in the case of early withdrawal). This session was conducted in order to test for the following eligibility criteria in the simulated MRI scanner: does not report falling asleep after instructed to stay awake for a 20-minute eyes-open resting session; follows instructions to to stay still (operationalized as falling below the 95th percentile of low-frequency motion as measured in a reference sample); demonstrates task engagement by responding to at least 33% of encoding trials in a memory task and obtains a d-prime of at least 0.1 in a subsequent memory test. Furthermore, due to eligibility criteria for collaborator sessions, participants were also tested for acceptable reading comprehension and speed through TOWRE score of at least 26.4 and Nelson-Denny score of 2. Lastly, participants were expected to show up on time for their appointment (within 15 minutes of the scheduled time) or reschedule 12 hours in advance for the pre-screening session (with tolerance for one missed appointment without notice), illustrate attention to and comprehension of verbal and written instructions, be courteous in all interaction with the experimenter, and not indicate high usage of alcohol or recreational drugs. Thirty-seven participants failed eligibility criteria.

Test-retest effects

TI and VSL were split into sections A and B which were identical tasks with different stimuli to measure a test-retest effect during the Pre-Sleep session. A second reason for two sections was that during piloting, we found that certain stimuli would have chance scores at all levels, whereas others would be above chance for some participants. Section A was tested twice at the Pre-Sleep test, once at the Post-Sleep test, and once at Day7. Section B was tested once at each test session.

Unfortunately, due to a programming error, the test-retests for TI were variable. Instead of four Section A tests and three Section B tests, sometimes Section A was tested five or six times and Section B once to twice. However, each participant had at least one Section A or Section B test at each time point. For cases when there were two tests at the Post-Sleep test or Day7, we only included the first test. There were 22 participants with full sets (7 tests as planned) and 48 with incomplete sets (7 tests but with variable amounts in each stimulus set). To ensure this repeated testing did not help participants to acquire the materials, we conducted an independent samples t-test to investigate whether participants with more retests in the incomplete set had higher scores than those with complete sets. We found that more retests did not lead to better scores for premise pairs, $t(393) = 2.22$, $p = .027$ in the opposite direction, $d = 0.23$, first-order inferences, $t(380) = -0.68$, $p = .49$, $d = 0.07$, or second-order inferences, $t(359) = 0.14$, $p = .88$, $d = 0.02$. Degrees of freedom are given by total testing sessions across all participants in complete and incomplete groups (i.e., 176 sessions for 22 participants in the complete set and 383 sessions for 48 participants in the incomplete set).

In the VSL task, we selected distinctive shapes for Section A and used stimuli created by Turk-Browne et al. (2005) and provided by Inquisit for Section B. Section A was significantly above chance at all time points, ps < .001. However, Section B was not significantly above chance at the Post-Sleep test and Day7, $p = .55$ and $p = .58$, respectively. Therefore, we only included Section A in this dataset, with the exception of

one participant for whom there were no scores from Section A, for whom we substituted those with the second set. There was a significant test-retest effect in memory for deterministic sequences from Section A, $t(44) = -2.09$, $p = .043$, $r = .86$, indicating that as participants were tested on the material a second time during Pre-Sleep test, their scores improved. There was no significant test-retest effect in memory for non-deterministic sequences from Section A, $t(44) = 0.06$, $p = .951$, $r = .72$.

In the TI task, we selected distinctive abstract images for both stimulus sets. Sections A and B were significantly above chance for all time points, *p*s < .001. There were no significant test-retest effects for premise, first-order or second-order inference memory of either stimulus set, *p*s > .29, indicating that repeated tests in the same night were not associated with higher scores.

DRM Measures

Due to unanticipated parameter estimates very close to 0 and 1 when running models for individual participants (rather than group-level analyses that were performed by Stahl & Klauer, 2008), we decided to use another analysis approach post-hoc that was better tailored to scores for individual participants. The goal was to determine gist and detail measures from the 3x3 matrix for each participant (old, related and new responses to old, related, and new items). Dandolo and Schwabe (2018) describe two models that reflect gist and detail memory. An “Old Distinct” model expects old memory representations to be distinct from both related and new ones, indicating a precise detailed memory. Conversely, an “Old and Related Similar” model expects old and related memory representations to be quite similar, while being distinct from new items, a model which suggests that gist has been encoded but detail has not. Because one model captures only detail, and the other only gist – the two measures are orthogonal. We adapted this model to our task. Hits in the “Old Distinct” model were old responses to old items, whereas hits in the “Old and Related Similar” model were old or related responses to old or related items. False alarms in the “Old Distinct” model were related responses to old items, whereas false alarms in the “Old and Related Similar” model were new responses to old or related items.

We started by generating a 3x3 matrix for each participant (old, related, and new responses to old, related, and new items). First, we divided each item in the 3x3 matrix by the number of total items in this category to get a proportion of old, related and new responses. Then, we collapsed old and related rows and columns into a 2x2 matrix of old/related and new responses to old/related and new items. Based on the interpretation that participants were using the “old”, “related” and “new” response options as a spectrum of “oldness”, we weighted old responses by 1 and related responses by 0.5 as a way to account for confidence of oldness in each response. Finally, we replaced all mean values of 0 and 1 in the resulting 2x2 matrix with 0.00001 and 0.9999 in order to avoid hit or false alarm rates of 0 or 100%. Then, we calculated d’ values for both the “Old Distinct” (memory for individual words) and “Old and Related Similar” (summary gist) models.

MRI data acquisition

The day prior to each participants’ MRI scan, they completed a biofeedback session in the simulated MRI scanner to become better habituated to an MRI-like environment and learn to reduce their head movement (and thereby improve the signal quality of their brain images sampled the next day). During the biofeedback session, participants viewed a 45-minute documentary with a live readout of their head motion overlaid. When their head motion exceeded an adaptive threshold, the documentary was paused for several seconds

while static was played on the screen along with a loud, unpleasant noise. Following the documentary, a brief memory test was administered outside of the mock scanner to ensure participants were paying attention to the film content rather than just their motion (not analyzed here).

The next day, we used a whole-body MRI scanner (Magnetom Tim Trio; Siemens Healthcare) to gather a variety of image sequences over the course of a 1.5 hour scan. As described in our pre-registration, our hypotheses in the current study related specifically to the medial-temporal lobe anatomical analyses. To assess these predictions, we gathered high-resolution whole-brain T1-weighted (T1w) and T2-weighted (T2w) anatomical images (in-plane resolution 0.7 x 0.7 mm^2; 320 x 320 matrix; slice thickness 0.7 mm; 256 AC-PC transverse slices; anterior-to-posterior encoding; 2 x acceleration factor; *T1w* TR 2400 ms; TE 2.13 ms; flip angle 8°; echo spacing 6.5 ms; *T2w* TR 3200 ms; TE 567 ms; variable flip angle; echo spacing 3.74 ms) and an ultra-high resolution T2-weighted volume centred on the medial temporal lobes (resolution 0.5 x 0.5 mm^2; 384 x 384 matrix; slice thickness 0.5 mm; 104 transverse slices acquired parallel to the hippocampus long axis; anterior-to-posterior encoding; 2 x acceleration factor; TR 3200 ms; TE 351 ms; variable flip angle; echo spacing 5.12 ms). The whole brain protocols were selected on the basis of protocol optimizations designed by Sortiropoulos and colleagues (2014). The hippocampal protocols were modeled after Chadwick and colleagues (2014).

Hippocampal volumes

As noted in our pre-registration, we did not have any specific hypotheses regarding laterality for hippocampal variables. Therefore, before beginning data analysis, we decided to perform a Pearson's correlation of aHPC and pHPC volumes across hemispheres. If the correlations were greater than 0.9, we would merge the volumes across hemispheres. Otherwise, we would run them separately in our models. aHPC volumes across hemispheres had a correlation of 0.686, and pHPC volumes across hemispheres had a correlation of 0.458. Therefore, we analyzed left and right hippocampal volumes separately in all of our models, investigating only either left or right hippocampal volumes in any given model.

The ultra-high-resolution T2w 0.5mm isotropic medial temporal lobe scans were submitted to automated segmentation using HIPS, an algorithm previously validated to human raters specialized in segmenting detailed neuroanatomical scans of the hippocampus (Romero, Coupé, & Manjón, 2017). Three independent raters (including one of the authors) were trained on segmenting the hippocampus at the uncal apex into aHPC and pHPC segments, and achieved a Dice coefficient of absolute agreement of 80%. Two of these raters (including one of the authors) independently segmented all participants in this study using the 0.5mm T1w scans. The T2w medial temporal lobe scans were registered to the T2w whole-brain scans, which were in turn registered to the T1w whole-brain scans, and the combined transform was used to place the rater landmarks on the detailed medial temporal lobe scans. All voxels belonging to the hippocampus posterior to this coronal plane containing this point were classified as posterior hippocampus (pHPC), whereas all points anterior to and including this plane were classified as anterior hippocampus (aHPC). Finally, the total number of voxels in each subregion was multiplied by the volume of each voxel to obtain a total aHPC and pHPC volume. These steps were followed separately in each hemisphere. Freesurfer segmentation of the T1w and T2w brain data (v6.0; Fischl et al., 2004) was used to compute Estimated Intracranial Volumes, which in turn were used to control for effects of overall head size in the hippocampus volume vectors.

Results of hippocampal analysis

Average anterior hippocampal volumes were 1027 (*SD* = 158) on the left and 1100 (*SD* = 123) on the right. Average posterior hippocampal volumes were 1053 (*SD* = 119) for the left, and 1026 (*SD* = 120) on the right. Average anterior/posterior ratios were 0.98 (*SD* = 0.19) on the left, and 1.09 (*SD* = 0.22) on the right.

In this study, we collected data from a narrow age range (22-35) among healthy adults. We preregistered that if age significantly predicts either anterior and posterior hippocampal volumes in a regression, we will include age as a covariate. If age does not significantly predict either or pHPC volume, then to simplify our analyses, we will not include age. Age did not significantly predict hippocampal volumes in our sample, so we did not include age in our models ($p = .718$ and $p = .668$ for left and right aHPC volumes, $p = .078$ and $p = .089$ for pHPC volumes).

After controlling for FDR, we did not find any significant effects of aHPC volumes, pHPC volumes or aHPC/pHPC ratio when including these variables in our model, all *p*s > 0.08.

References

Abe, N., Okuda, J., Suzuki, M., Sasaki, H., Matsuda, T., Mori, E., ... & Fujii, T. (2008). Neural correlates of true memory, false memory, and deception. *Cerebral Cortex*, *18*(12), 2811-2819.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*(1), 3-47.

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Carskadon, M. A., & Rechtschaffen, A. (2011). Monitoring and staging human sleep. *Principles and practice of sleep medicine*, *5*, 16-26.

Chadwick, M. J., Bonnici, H. M., & Maguire, E. A. (2014). CA3 size predicts the precision of memory recall. *Proceedings of the National Academy of Sciences USA*, *111*(29), 10720–10725.

Collin, S. H., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature neuroscience*, *18*(11), 1562.

Dandolo, L. C., & Schwabe, L. (2018). Time-dependent memory transformation along the hippocampal anterior–posterior axis. *Nature communications*, *9*(1), 1205.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*(2), 114-126.

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory*, *16*(12), 751-755.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia*, *49*(5), 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*, *104*(18), 7723-7728.

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Segonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *Neuroimage*, *23 Suppl 1*, S69-84.

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological science*, *17*(8), 670-674.

Hasselmo, M. E. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends in cognitive sciences*, *3*(9), 351-359.

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), 153-162.

Inquisit 5 [Computer software]. (2017). Retrieved from http://www.millisecond.com.

Jones, M. W., & Wilson, M. A. (2005). Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task. *PLoS biology*, *3*(12), e402.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Baglioni, C., Spiegelhalder, K., ... & Nissen, C. (2014). The reorganisation of memory during sleep. *Sleep Medicine Reviews*, *18*(6), 531-541.

Lau, H., Tucker, M. A., & Fishbein, W. (2010). Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of learning and memory*, *93*(4), 554-560.

Levendowski, D. J., Ferini-Strambi, L., Gamaldo, C., Cetel, M., Rosenberg, R., & Westbrook, P. R. (2017). The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *Journal of Clinical Sleep Medicine*, *13*(06), 791-803.

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in cognitive sciences*, *15*(8), 343-351.

Lopez, N., Previc, F. H., Fischer, J., Heitz, R. P., & Engle, R. W. (2012). Effects of sleep deprivation on cognitive performance by United States Air Force pilots. *Journal of Applied Research in Memory and Cognition*, *1*(1), 27-33.

Lucey, B. P., Mcleland, J. S., Toedebusch, C. D., Boyd, J., Morris, J. C., Landsness, E. C., ... & Holtzman, D. M. (2016). Comparison of a single−channel EEG sleep study to polysomnography. *Journal of sleep research*, *25*(6), 625-635.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, *11*(10), 442-450.

Matorina, N., & Poppenk, J. (2019, March 14). Sleep Contributions to Hippocampal Consolidation of Gist and Detail Memory. Retrieved from osf.io/kqc7z

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30.

Millisecond Software (2015). Automated Operation Span [Computer software]. Retrieved from https://www.millisecond.com/download/library/ospan/

Millisecond Software (2015). Deese-Roediger-McDermott (DRM) False Memory Procedure (Visual) [Computer software]. Retrieved from https://www.millisecond.com/download/library/falsememories/

Millisecond Software (2015). Transitive Inference Task [Computer software]. Retrieved from https://www.millisecond.com/download/library/TransitiveInferenceTask/

Millisecond Software (2015). Automated Operation Span [Computer software]. Retrieved from https://www.millisecond.com/download/library/ospan/

Mountjoy, J., & Poppenk, J. (2015). Introducing SuperPsychToolbox: An Open-Source Tool to Facilitate Coding and Analysis of Psychology Experiments. In *Canadian Journal of Experimental Psychology, 69(*4), 332-332.

Nair, S. N., Czaja, S. J., & Sharit, J. (2007). A multilevel modeling approach to examining individual differences in skill acquisition for a computer-based task. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *62*(1), 85-96.

Stahl, C., & Klauer, K. C. (2008). A simplified conjoint recognition paradigm for the measurement of gist and verbatim memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 570.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature neuroscience*, *16*(2), 139-145.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

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

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences*, *17*, 114-123.

Romero, J. E., Coupé, P., & Manjón, J. V. (2017). HIPS: A new hippocampus subfield segmentation method. *NeuroImage*, *163*, 286–295.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L. W., Wamsley, E. J., Tucker, M. A., ... & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of learning and memory*, *92*(3), 327-334.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Team, R. C. (2007). Linear and nonlinear mixed effects models. *R package version*, *3*(57), 1-89.

Sotiropoulos, S. N., Jbabdi, S., Xu, J., Andersson, J. L., Moeller, S., Auerbach, E. J., … Behrens, T. E. (2013). Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage*, *80*, 125–143.

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

Tompary, A., & Davachi, L. (2018). Consolidation Promotes the Emergence of Representational Overlap Across Related Memories in the Hippocampus and Medial Prefrontal Cortex. *Available at SSRN 3155632*.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of cognitive neuroscience*, *21*(10), 1934-1945.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*(33), 11177-11187.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods*, *37*(3), 498-505.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. Nature, *427*(6972), 352.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. *Nature Reviews Neuroscience*, *11*(3), 218-218.

Yordanova, J., Kolev, V., Wagner, U., Born, J., & Verleger, R. (2012). Increased alpha (8–12 Hz) activity during slow wave sleep as a marker for the transition from implicit knowledge to explicit insight. *Journal of Cognitive Neuroscience*, *24*(1), 119-132.

Yordanova, J., Kolev, V., Wagner, U., & Verleger, R. (2009). Covert reorganization of implicit task representations by slow wave sleep. *PLoS One*, *4*(5), e5675.

Acknowledgements

We gratefully acknowledge Gillian Marvel for assistance with behavioural and sleep EEG data collection; Julie Tseng, Lauren DeMone, and Natalie Doan with scheduling; Julie Tseng and Don Brien with MRI data acquisition; and Justin Siu, Roland Dupras and Mike Lewis with technical support. This research was funded by Natural Sciences & Engineering Research Council Discovery Grant 03637 (J.P.), which also supported N.M. Infrastructure funding was provided by Canada Foundation for Innovation – John R. Evans Leaders Fund (J.P.), and a Queen's University Research Initiation Grant to J.P., who was supported by the Canada Research Chairs program.