Sleep and testing both strengthen and distort story recollection

Dan Denis[1], Carissa DiPietro[1], R. Nathan Spreng[2], Daniel L. Schacter[3], Robert Stickgold[4,5], and Jessica D. Payne[1]

[1] Department of Psychology, University of Notre Dame

[2] Montreal Neurological Institute, McGill University

[3]Department of Psychology, Harvard University

[4] Department of Psychiatry, Beth Israel Deaconess Medical Center

[5] Department of Psychiatry, Harvard Medical School

Author Note

Dan Denis https://orcid.org/0000-0003-3740-7587

Carissa DiPietro https://orcid.org/0000-0002-4996-7277

R. Nathan Spreng https://orcid.org/ 0000-0003-1530-8916

Daniel L. Schacter https://orcid.org/ 0000-0002-2460-6061

Robert Stickgold https://orcid.org/ 0000-0003-3971-744X

Jessica D. Payne https://orcid.org/ 0000-0003-3643-0574

This study was supported by a National Institutes of Health grant (NIH MH048832) awarded to RS and a National Institute on Aging grant (R01 AG008441) awarded to DLS. JDP, NS, DLS, and RS conceived and designed the study. JDP led the collection of data. DD analyzed the data. DD and CD led the writing of the manuscript. All other authors contributed to the writing process. We would like to thank Jon Chamberlain for assistance in data collection, and

Sara Kim for helpful comments on the manuscript. We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Dan Denis, University of Notre Dame, E439 Corbett Family Hall, Notre Dame, Indiana, 46556. Email: ddenis@nd.edu.

Abstract

Over time, memories lose episodic detail and become distorted, a process with serious ramifications for topics such as eyewitness identification. What are the processes which contribute to such transformation over time? We investigated the roles of post learning sleep and retrieval practice in memory accuracy, transformation, and distortion, using a naturalistic story recollection task. Undergraduate students listened to a recording of the “War of the Ghosts”, a Native American folktale, and were assigned to either a retrieval practice or listen only study condition, and either a sleep or wake delay group. Sleep and retrieval practice independently resulted in more story elements being recalled accurately, and fewer importations of non-story elements, than the wake, no retrieval practice group. However, sleep and retrieval practice also led to more inferences of non-presented, but story related information. These findings suggest that both sleep and retrieval practice contribute equally to narrative memory stabilization and distortion.

Statement of relevance

When we recall a memory, rarely is it a verbatim reinstatement of the original experience. Over time, our memories are transformed. We lose episodic detail and replace it with gist-based knowledge. Our memories also become distorted and can become infused with misinformation. Our study looked at how two mechanisms, namely sleep and post-learning testing, influence memory transformation for a story over 12 hours. We found both processes increased memory accuracy, but also introduced distortion. Both sleep and testing led to more inferences, where participants remembered plausible story elements that were not mentioned in the original text. This work shows how memories are transformed over time, and raise important implications for cases where clear, accurate recall of events is important, such as eyewitness testimony. This work has ramifications in teaching settings, where testing is routinely used to assess accuracy of education materials.

Sleep and retrieval testing both strengthen and distort story recollection

Human memories are inevitably transformed and distorted over time, causing us to remember past events quite differently than the true original experience. Episodic detail is lost, with the "gist" of the narrative being preserved (Brainerd & Reyna, 2001; Lutz et al., 2017; Zeng et al., 2021). Distortions of memory also appear, with memory for different but overlapping events combined, or forgotten details being replaced or fabricated in order to fit a comprehensive narrative in tune with our cultural schemas (Bartlett, 1932; Loftus, 2005; Roediger & McDermott, 2000; Webb et al., 2016). Such transformative processes confer a flexibility and generalizability to memory that may at times be more advantageous than a veridical reinstatement of the original experience. However, this flexibility comes at the cost of memory accuracy, which has very important real-world implications when considering the reliability of eyewitness testimony (Fruzzetti et al., 1992; Roediger & McDermott, 2000; Wetmore et al., 2015). As such, understanding when and how these transformations of memory occur remains a central question in the cognitive psychology of human memory.

Perhaps the most historic example of memory distortion comes from Bartlett's (1932) *Remembering: A study in Experimental and Social Psychology*, and his famous "War of the Ghosts" experiments. The War of the Ghosts is a Native American folktale written in a disjointed, non-Western style unfamiliar to most college students (Bartlett, 1932). Bartlett had participants read the story twice before testing their memory for the story over varying time delays. He found that over time, not only did forgetting occur, but memory also become distorted and infused with misinformation. Such distortions in recall became magnified over time and with repeated recall attempts, as participants held onto the schema or gist while replacing story details in order to keep a complete narrative. The importations could be quite dramatic, especially if the delay between recall attempts was long (Bartlett, 1932). Since Bartlett's initial study, his results have been replicated under stricter experimental conditions (Bergman & Roediger, 1999; Wagoner & Gillespie, 2014). What factors influence these time-dependent transformations?

Simply attempting to recall a memory shorty after the original event can bolster memory. This is the basic tenet of the testing effect, that retrieval practice following learning strengthens memory in a manner that becomes clear at a later, subsequent test (Adesope et al., 2017; Roediger & Butler, 2011). Numerous studies have demonstrated the benefits of retrieval practice on memory accuracy (see Rowland, 2014 for a detailed review and meta-analysis), and it's principles have been widely applied in classroom settings to improve students' learning outcomes (Schwieren et al., 2017). However, to our knowledge, no studies have examined how retrieval practice may impact distortions of memory.

A period of sleep following learning can also benefit memory retention (compared to a similar period awake, e.g., Payne, 2011; Payne et al., 2008; Rasch & Born, 2013). Rather than simple passive protection from interference, sleep actively strengthens and restructures memories, allowing insights to be gained (Sanders et al., 2019; Wagner et al., 2004), inferences to be made (Ellenbogen et al., 2007; Huguet et al., 2019), and integration and abstraction to occur (Tamminen et al., 2010).

It remains unclear whether sleep and retrieval interact in terms of their impact on memory. Some research suggests that pre-sleep retrieval practice eliminates the typical benefit of sleep compared to wake (Abel et al., 2019; Antony & Paller, 2018; Bäuml et al., 2014). On the other hand, other work has found that a benefit of sleep *only* occurs if a pre-sleep retrieval test is included (Schoch et al., 2017), and only for items correctly recalled during pre-sleep testing (Denis et al., 2020; Fenn & Hambrick, 2013; Muehlroth et al., 2020).

The current study used the War of the Ghosts story to examine the effect of sleep and retrieval practice on story recall, transformation, and distortion. The War of the Ghosts story has several advantages as a tool for studying memory. Story recall represents the kind of recollection that people engage in naturalistically, in which memory for the “gist” of an event often replaces veridical recall. It enables us to study not only accurate retrieval of facts, but also the distortions and importations of new information that result from schema-based representations.

Method

Participants

A total of 114 Harvard University undergraduate students (M_{age} = 19.9, SD = 2.8, 62 females) participated in the study and were compensated for their time with either course credit or $20. Because the research design and effects of interest were novel, it was not possible to estimate effect size from previous studies in order to perform an *a priori* power analysis. However, we note that our sample size is comparable to other sleep and memory studies (Klinzing et al., 2021; Noack et al., 2021; Payne, Stickgold, et al., 2008; Scullin & McDaniel, 2010).

All participants were native English speakers with normal or corrected-to-normal vision and free from any psychiatric disorders, sleep, drug, or alcohol problems. Participants were not taking any medications affecting the central nervous system. Informed consent was taken from all participants, and the study was approved by the Harvard University and Beth Israel Deaconess Medical Center Internal Review Boards.

Materials

All participants completed a 3-night sleep log and the Stanford Sleepiness Scale (a one-item questionnaire used to assess participant’s alertness in the moment) at the start of each session (Hoddes et al., 1972). A recorded version of the Native American folktale “The War of the Ghosts” served as the to-be-remembered material (Bartlett, 1932).

Procedure and Design

Participants were assigned to one of four experimental groups (**Figure 1**): 1) a wake + retrieval practice group (n = 30); 2) a wake, listen-only group (n = 29); 3) a sleep + retrieval practice group (n = 27), and 4) a sleep, listen-only group (n = 28). Participants listened to the War of the Ghosts story twice in succession, either in the morning (9am), or the evening (9pm) depending on the sleep/wake group assignment (i.e., wake groups listened in the morning and sleep groups listened in the evening). All participants then completed a 15-minute distractor task consisting of simple arithmetic problems. Next, half of the participants were dismissed (i.e., those in the two listen-only groups; wake, listen-only group; sleep, listen-only group), while the other half of participants were given a blank sheet of notebook paper and were instructed to engage in free recall by recollecting the story as accurately as possible (i.e., those in the two retrieval practice groups; wake + retrieval practice group; sleep + retrieval practice group). All participants returned 12 hours after their first session for a final recall test, following either a day of wakefulness (wake groups) or after a night of sleep (sleep groups). For half of the participants, it was their first time recalling the story (listen-only groups), while for the other half, it was their second recall of the story (retrieval practice groups).

During each recall test, participants were encouraged to use the exact same words as the ones heard in the story (veridical recall), and to write down as many facts and events as possible. Participants were given 10min for each recall test and were encouraged to use the full time.

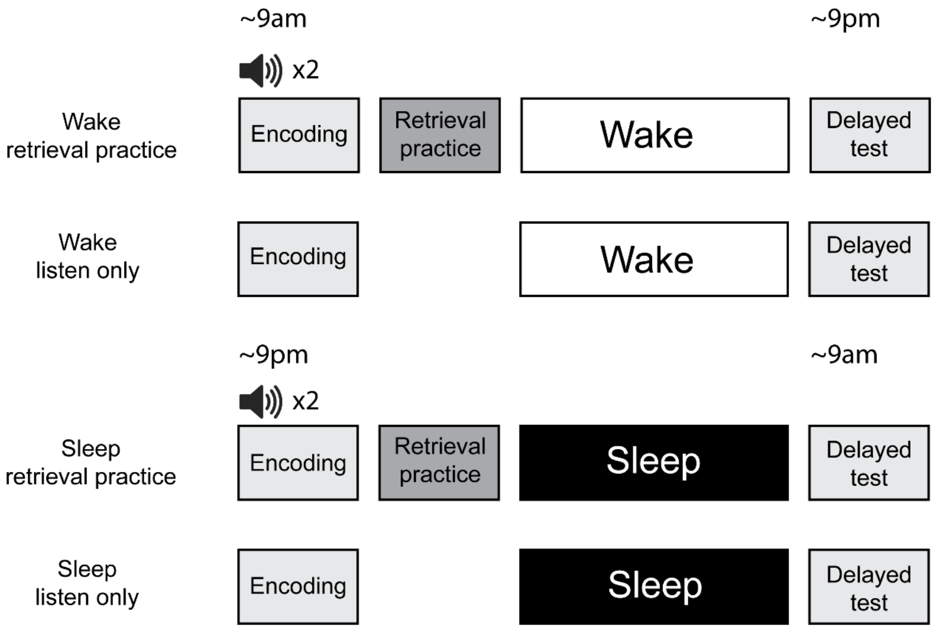

Figure 1. Experimental design. During encoding, participants listened twice to an audio recording of The War of the Ghosts. After a 15-minute distractor task, half of the participants were given a sheet of paper and were asked to recall the story in as much detail as possible. The other half were dismissed after listening to the story. After a 12-hour delay, all participants returned to the lab and were required to recall the story in as much detail as possible. In half of the participants, the delay included a day of wakefulness. In the other half, the delay included a night of sleep.

Scoring

The War of the Ghosts story was subdivided into 42 propositions following previously established methods (Bergman & Roediger, 1999; Mandler & Johnson, 1977). Recalled material was transcribed and divided into units of propositions. These transcripts were then scored by two researchers who were blind to experimental condition. Sentences from the participant transcripts were matched to the proposition or propositions in the story that were most similar in content. We identified propositions from the recalled material that were fit into five categories: accuracy, gist, omission, distortion, and importation. Distortions were further subdivided into three subcategories; inference, normalization, and incorrect placement. The exact rule followed for each proposition category, and an accompanying example, can be found in **Table 1.** Inter-rater reliability between the primary and secondary scorer showed a Cohen’s kappa of .84. All data reported below are those of the primary scorer.

Table 1. Rules for scoring each proposition in the story

Proposition category	Rule	Example
Accurate	Propositions that were recalled verbatim, or verbatim with minor omission that did not result in a loss of detail	Verbatim: “They escaped to the shore, and hid behind some logs” With minor omission: “They escaped and hid behind some logs”
Gist	Contained all elements of a proposition but used distinctly different wording, but the meaning was essentially correct	“I will not go along. I might be killed. My relatives don’t know where I have gone” changed to “But my family won’t know where I have gone and I might die and they’d never know”
Omission	Omission of key details of the proposition	“now canoes came up, and they heard the sound of paddles” changed to “canoes were approaching, and they started to hear noises”
Inference	Wrote an event in a proposition that was not explicitly stated, but could reasonably be assumed to have occurred between two events	Writing that one of the boys got into the canoe, when it is never explicitly stated that he traveled with the warriors
Normalization	Conventionalization of a story element	Recalling that the two young men went on a fishing trip, rather than going down to the river to hunt seals
Incorrect placement	The proposition contained an element from a different portion of the story	Referring to the warriors as ghosts at the beginning of the story, when they are not referred to as ghosts until the middle
Importation	An entirely new element was added to the story	Recalling that the warriors fought with guns (the story only mentions bows and arrows)

Results

Ruling out time of day influences

First, we checked to see whether time of day affected recall by comparing memory metrics (accuracy, gist, omission, distortion, and importation) between the sleep and the wake group who engaged in retrieval practice. There were no differences in initial free recall performance based on delay condition for any of the metrics (all *p*’s > .08, all *d*’s < 0.50; **Table**

S1). Similarly, there was no difference between the sleep and wake groups in terms of total number of propositions produced at the initial (*t* (54.80) = 0.96, *p* = .34, *d* = 0.26) test. At the delayed test, a 2 (delay group; sleep, wake) x 2 (study condition; retrieval practice, listen-only) ANOVA found a significant main effect of delay group on total number of propositions produced (*F* (1, 110) = 9.43, *p* = .003, ηp^2 = .08), with a higher number of propositions in the sleep groups (M = 29.13, SD = 4.11) compared to the wake groups (M = 26.51, SD = 5.06). The main effect of study condition, and the interaction, were both non-significant (*p*s > .07).

There were no group differences on the Stanford Sleepiness Scale, at either the initial (*t* (56) = 0.62, *p* = .54, *d* = 0.16) or final (*t* (89) = 0.60, *p* = .55, *d* = 0.12) recall test. As such, time of day did not influence memory performance, and participants were equally alert (subjectively) in the morning and in the evening.

Effects of sleep and retrieval practice on recall performance

A series of 2 (delay group; wake, sleep) x 2 (study condition; retrieval practice, listen-only) ANOVAs examined the effects of sleep and retrieval practice on accuracy, gist, omission, distortion, and importations during story recollection at the delayed test (**Table 2; Figure 2**).

Table 2. Percentage of propositions exhibiting each memory category at the final recall test

	Wake listen-only	Sleep listen-only	Wake + retrieval practice	Sleep + retrieval practice
	M (SD)	M (SD)	M (SD)	M (SD)
Accuracy	5.16 (5.32)	10.24 (6.32)	9.71 (6.49)	9.57 (6.51)
Gist	32.20 (10.63)	30.74 (8.06)	33.04 (11.39)	31.30 (9.07)
Omission	17.69 (6.10)	16.66 (7.04)	16.86 (7.08)	16.49 (5.68)
Distortion (inference)	7.70 (6.16)	11.91 (6.04)	11.85 (6.72)	11.99 (7.95)
Distortion (normalization)	6.40 (4.08)	6.45 (4.21)	6.27 (4.78)	4.37 (3.79)
Distortion (incorrect placement)	1.85 (2.54)	2.52 (2.49)	2.14 (2.51)	2.97 (2.58)
Importation	29.00 (11.48)	21.48 (8.89)	20.13 (11.63)	23.30 (9.39)

Note. M = mean, SD = standard deviation

For accuracy, there was a significant main effect of delay group (sleep/wake) (*F* (1, 110)

= 4.55, *p* = .035, ηp^2 = .04), with higher accuracy after sleep (M = 9.91%, SD = 6.36%) compared to wake (M = 7.47%, SD = 6.33%). There was no main effect of study condition (retrieval practice/listen-only) (*F* (1, 110) = 3.07, *p* = .08, ηp^2 = .03), but there was a significant delay group (sleep/wake) * study condition (retrieval practice/listen-only) interaction (*F* (1, 110) = 5.07, *p* = .026, ηp^2 = .04). Follow-up tests revealed that accuracy was significantly higher after sleep compared to wake only for those participants who did not engage in retrieval practice (*t* (52.79) = 3.28, *p* = .002, *d* = 0.87). When participants engaged in retrieval practice, there was no difference in accuracy between the sleep and the wake groups (*t* (54.33) = -0.08, *p* = .94, *d* = 0.02). This effect was driven by the presence of a post-learning retrieval practice session significantly improving accuracy across a day of wake (*t* (55.54) = 2.95, *p* = .005, *d* = 0.77), without changing accuracy across a night of sleep (*t* (52.76) = -0.39, *p* = .70, *d* = 0.11). We note that the interaction remained significant after removal of an outlier (> 1.5 * inter-quartile range) in the sleep, listen-only group (*F* (1, 109) = 4.07, *p* = .046, ηp^2 = .04). There were no main effects or interactions between delay group and study condition on the proportion of gist propositions or omissions (all *p* > .38).

With regards to distortion, the main effects of both delay group (sleep/wake) and study condition (retrieval practice, listen-only) were non-significant (*p* > .17). There was, however, a significant interaction between delay group (sleep/wake) and study condition (retrieval practice, listen-only) (*F* (1, 110) = 4.08, *p* = .046, ηp^2 = .04). This interaction was driven by significantly more distortions in the sleep group compared to the wake group when there was no post-learning retrieval practice session (*t* (52.89) = 2.98, *p* = .004, *d* = 0.79). We probed this finding further by comparing the three subtypes of distortion between the sleep and wake groups (**Figure 3**). The delay groups differed only in the proportion of inferences (*t* (54.98) = 2.61, *p* = .012, *d* = 0.69),

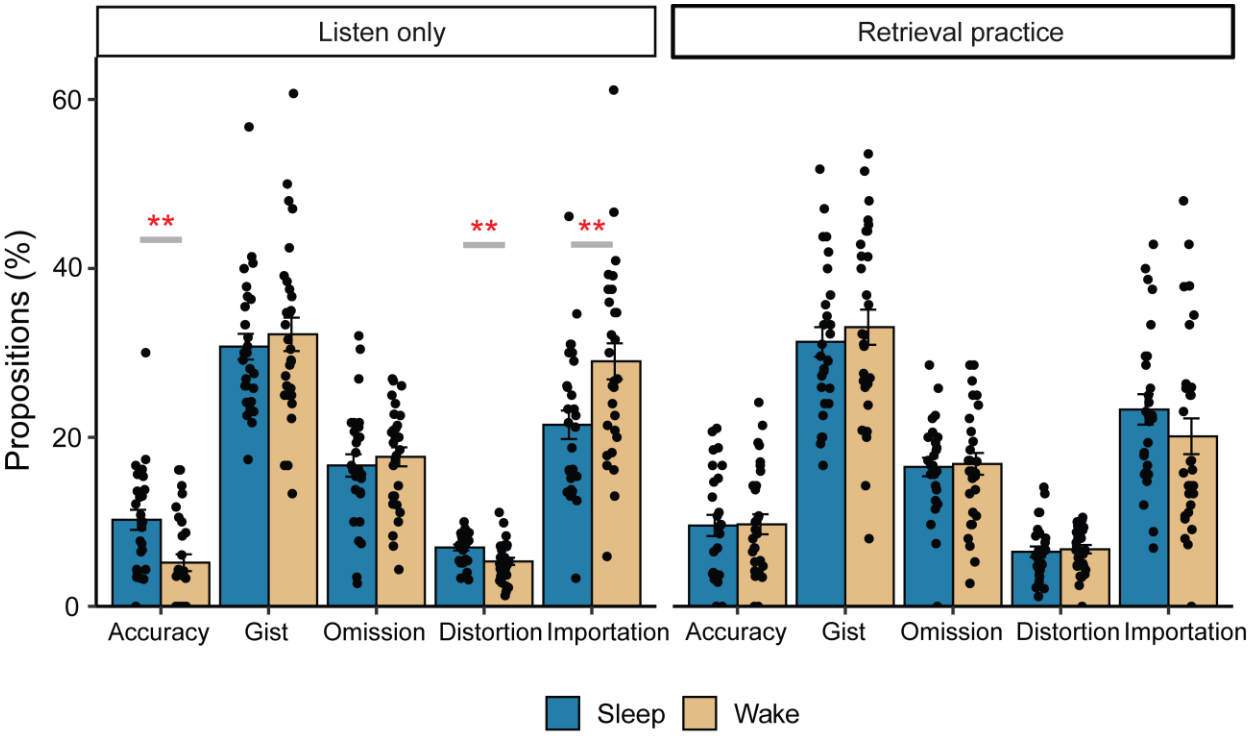

Figure 2. Percentage of propositions for each memory category at the delayed test. Dots represent individual data points, and error bars display the standard error. ** = *p* < .01.

with sleep participants making more inferences than wake participants. The proportion of inferences made by the sleep group did not differ depending upon whether participants engaged in retrieval practice or not (*t* (48.53) = 0.04, *p* = .97, *d* = 0.01); however, wake participants made significantly more inferences in the retrieval practice condition, compared to the listen-only condition (*t* (56.84) = 2.48, *p* = .016, *d* = 0.65). There were no delay group (sleep/wake) differences for normalizations (*t* (54.76) = 0.05, *p* = .96, *d* = 0.01) or incorrect placements (*t* (54.99) = 1, *p* = .32, *d* = 0.27).

For importations, the main effect of delay group (sleep/wake) was not significant (*F* (1, 110) = 1.23, *p* = .27, ηp^2 = .01), and the main effect of study condition (retrieval practice/listen-only) was marginal (*F* (1, 110) = 3.60, *p* = .06, ηp^2 = .03), showing numerically more importations in listen-only groups (M = 25.31%, SD = 10.88%) compared to retrieval practice groups (M = 21.63%, SD = 10.65%). There was, however, a significant interaction between delay group (sleep/wake) and study condition (retrieval practice, listen-only) (*F* (1, 110) = 7.44, *p* = .007, ηp^2 = .06). Follow-up tests revealed there to be significantly fewer importations in the sleep group compared to the wake group when participants did not engage in retrieval practice (*t* (52.56) = -2.77, *p* = .008, *d* = 0.73). On the other hand, when participants engaged in retrieval practice, the difference in importations between the wake and sleep group disappeared (*t* (54.39) = 1.14, *p* = .26, *d* = 0.30). Mirroring the results for accuracy, this effect was driven by the proportion of importations reducing across a wake delay for those in the retrieval practice condition, compared to those in the listen-only condition (*t* (55.54) = -2.95, *p* = .005, *d* = 0.77), whereas importations rates were equivalent across sleep, regardless of whether a post-learning retrieval practice session was present or not (*t* (52.76) = 0.39, *p* = .70, *d* = 0.11). The interaction remained significant when an outlier was removed from the wake, listen-only group (> 1.5*inter quartile range; *F* (1, 109) = 6.39, *p* = .013, ηp^2 = .06).

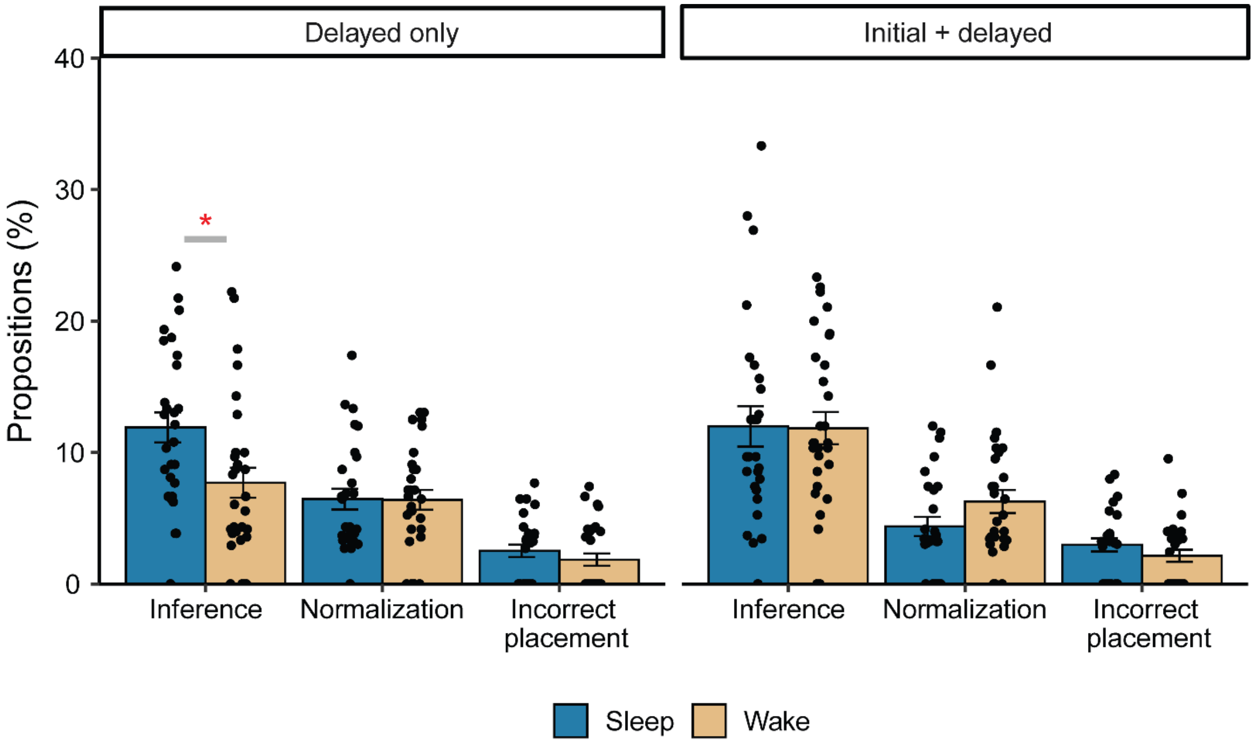

Figure 3. Percentage of propositions showing each distortion subtype at the delayed test. Dots represent individual data points, and error bars display the standard error. * = *p* < .05.

These results suggest an effect of sleep on increasing accuracy *and* inference, while reducing importations, compared to wake. However, this was only true in the absence of retrieval practice. We next sought to quantify the relative effects of sleep and retrieval practice on memory performance (**Figure 4**). Using the wake, listen-only group as a baseline (i.e., the group with neither sleep nor retrieval practice), we compared the size of improvements gained via either sleep (sleep, listen-only group), retrieval practice (wake, retrieval practice group) or both (sleep, retrieval practice group) by subtracting the average wake, listen-only final recall test score from each participant's score in the other three groups. (See elsewhere for examples of this approach: Denis et al., 2020; Talamini et al., 2008). Compared to the wake, listen-only group, both sleep and retrieval practice led to a significant improvement in accuracy (all *p*'s < .001, all *d*'s > 0.69) and significantly more inferences (*p*'s < .003, *d*'s > 0.60), while significantly reducing the proportion of importations (*p*'s < .001, *d*'s > 0.75). The size of these effects was equivalent, with sleep and retrieval practice providing nearly identical changes in memory (accuracy increase: sleep M (SD) = 5.07% (6.32%), retrieval practice M (SD) = 4.55% (6.49%); inference increase: sleep M (SD) = 4.21% (6.04%), retrieval practice: M (SD) = 4.15% (6.72%); importation decrease: sleep M (SD) = -7.52% (8.89%), retrieval practice: M (SD) = -8.88% (11.63%), all *p* > .60, all *d* < 0.14). The influences of sleep and retrieval practice were not additive, with the sleep + retrieval practice group not showing any differences compared to the sleep, listen-only or wake + retrieval practice groups (all *p*'s > .25, all *d*'s < 0.31).

As a final analysis, we examined how memory changed between the initial and delayed test in the retrieval practice groups (**Figures S1 & S2**). We performed these analyses using a series of 2 (session: initial recall, delayed recall) X 2 (delay period: sleep, wake) mixed

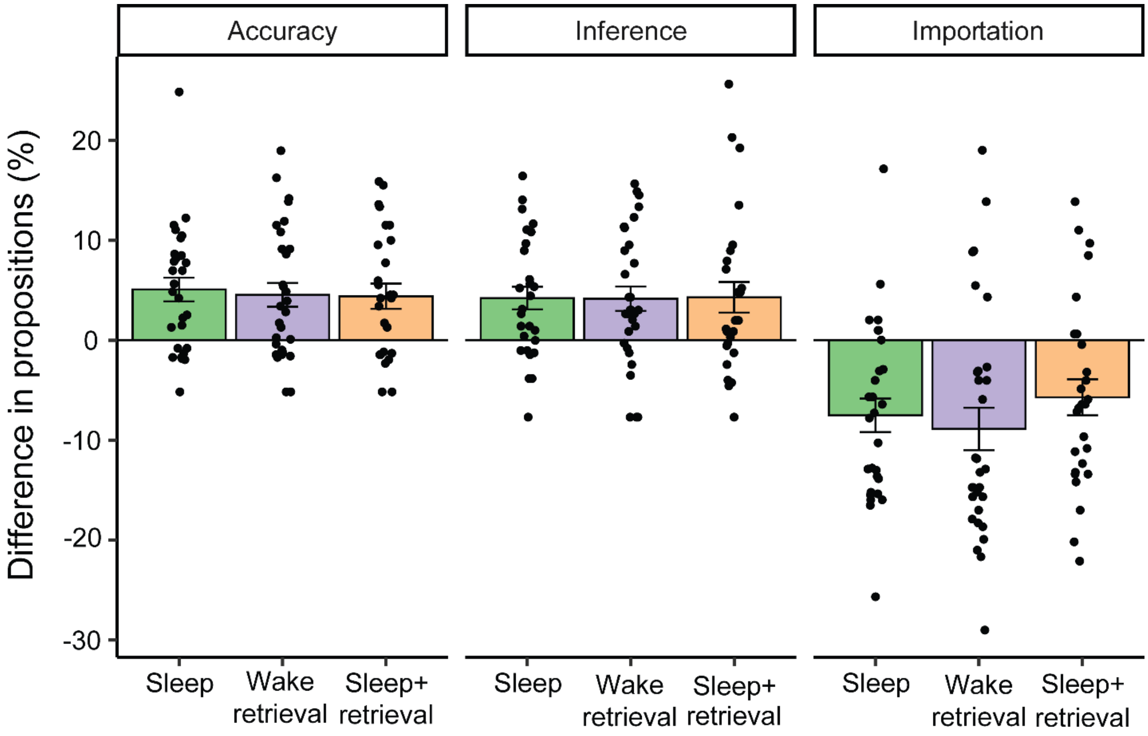

Figure 4. Relative effects of sleep and wake + retrieval practice on delayed memory performance. Difference scores were calculated by subtracting mean accuracy/inference/importation from the wake, listen-only group from each participant's score in the other three groups. As such, a positive value indicates a relative increase with sleep and/or retrieval practice, and a negative value indicates a relative decrease. Error bars show the standard error.

ANOVAs. Compared to initial test, there was a significant reduction in the percentage of gist propositions (F (1, 55) = 15.53, p < .001, ηp^2 = .22) and omissions (F (1, 55) = 5.85, p = .019, ηp^2 = .10) at the delayed test. Conversely, there was a significant increase in the percentage of propositions scored as either distortion (F (1, 55) = 17.05, p < .001, ηp^2 = .24) or importation (F (1, 55) = 17.51, p < .001, ηp^2 = .24) at delayed test, compared to initial test. The effect of delay period (sleep or wake), and the interaction between session and delay period, were all non-significant (all p > .26).

When we examined the different distortion subtypes separately, we found significantly more inferences at the delayed test compared to the initial test (F (1, 55) = 11.29, p = .001, ηp^2 = .17). We also found a significant session * delay period interaction for the change in normalizations (*F* (1, 55) = 4.27, *p* = .044, ηp^2 = .07; **Figure S2**). This interaction was driven by a significant increase in normalizations from initial to delayed test in the wake delay group (*t* (29) = 2.16, *p* = .04, *d* = 0.39), with no change between sessions in the sleep delay group (*t* (26) = 0.95, *p* = .35, *d* = 0.18). No other effects were significant (all *p* > .06).

Discussion

Bartlett's War of the Ghosts paradigm is probably the most famous historical example of a scientific study of how memory can be transformed and distorted over time. Here, we studied how two key processes, namely overnight sleep and post learning retrieval practice, impacted memory for the War of the Ghosts story over a 12-hour delay. Our key findings were that, compared to a day of wake without retrieval practice, both sleep and retrieval practice independently resulted in more accurate propositions being produced at the delayed test, but at the same time, led significantly more inferences and significantly fewer importations. This finding adds further evidence that the testing effect can be as powerful as sleep (Abel et al., 2019; Antony & Paller, 2018; Bäuml et al., 2014), and extends the literature to show testing also induces distortions of memory, again in a manner similar to sleep. Although such transformations can be beneficial in terms of memory flexibility and generalization, there are also times that such distortions of the original event is unwanted (e.g. eyewitness testimony). This is also an important consideration for educators, when considering the impacts of testing and sleep on the retention and accurate recall of educational materials.

Our finding that sleep resulted in improved accuracy compared to wake fits with the large body of evidence that sleep strengthens memories for facts and events (Ellenbogen et al., 2006). For example, memory for word pairs is significantly better after a period of sleep compared to an equivalent period of wakefulness (e.g. Denis, Schapiro, et al., 2020; Feld et al., 2016; Payne et al., 2012; Plihal & Born, 1997). The present work shows that this benefit extends to a story-based task, similar to other work (Aly & Moscovitch, 2010; Tilley & Empson, 1978). This finding is important, as it provides evidence that sleep's benefit to memory translates to a more ecologically valid test of memory.

We also found that participants who slept made more inferences than participants who stayed awake. As well as being strengthened, memories undergo transformation processes during sleep. One of these transformations involves improvements in relational memory, the ability to generalize new information across existing stores of knowledge. Prior experimental work has suggested that making new inferences about the relationships between abstract visual stimuli benefits from time, and especially sleep (Alger & Payne, 2016; Ellenbogen et al., 2007; Huguet et al., 2019; Werchan & Gómez, 2013). In the present study, we defined an inference as a proposition that contained an event not explicitly stated in the story but that conceivably could have occurred between two explicitly described events. The fact that these occurred more frequently after sleep may reflect the sleeping brain's ability to generalize aspects of the story and utilize existing knowledge structures to form gist-consistent inferences regarding how events in the story likely linked together.

This increase in proportion of inferences after sleep coincided with a reduction in importations. A proposition was scored as an importation when the proposition inserted a new event that was *not* clearly related to the events in the story – a type of error akin to a flagrant false memory as opposed to a gist-based one. The increased inferences and decreased importations following sleep may indicate a role of sleep in adaptively differentiating different networks of interrelated memories, and reducing their overlap (Doxey et al., 2018; Hanert et al., 2017). Such a process would ensure that generalizations and integration occur within existing related knowledge schema, while potential interference from unrelated schemas can be reduced.

These effects were not unique to sleep. Post-learning retrieval practice, followed by a day of wake, produced the same effects on memory as a night of sleep, when compared to the wake, listen-only control group (i.e., retrieval practice increased accuracy and inferences, while reducing importations). There was no additive effect of sleep and retrieval practice, as the group that engaged in both retrieval practice and a night of sleep did not differ in delayed recall performance from the groups that received one or the other. One interpretation is that the changes in memory induced by retrieval practice altered the underlying memory representation to such a degree that sleep could not act on that representation further. A growing body of work has demonstrated other instances in which pre-sleep retrieval practice precludes any benefit of sleep, compared to a restudy condition (Abel et al., 2019; Antony & Paller, 2018; Bäuml et al., 2014). The findings of the present study illustrate that retrieval practice can induce similar memory transformations as a night of sleep. This fits with recent neuroimaging work suggesting that the neural changes induced by retrieval practice may be similar to those induced by sleep (Antony et al., 2017; Brodt et al., 2018).

Within the retrieval-practice groups, we were able to examine how memory changed between the initial and delayed test, when the delay period was filled with either a night of sleep or a day awake. In general, changes in memory between the initial and delayed test were similar across sleep and wake delays. Whether this effect of retrieval practice lowers interference that occurs across wake, reduces the need for sleep-associated consolidation mechanisms to further process the memory, or both, remains to be seen. The exception to this pattern was the change in normalizations made. Following retrieval practice, the wake delay group made significantly more normalizations at the delayed test, whereas there was no change across a night of sleep (with a trend towards fewer normalizations). This leaves open the possibility then that some types of memory distortion may be more likely occur across a wake delay compared to a sleep delay, and future work should seek to explore this further.

We did not collect any objective measurements regarding participants' sleep. As such, we do not know if any sleep stage or feature was specifically associated with any of the observed memory transformations. At a neural level, sleep promotes consolidation via large-scale changes to the neural networks associated with long-term memory (Klinzing et al., 2019; Payne & Kensinger, 2011; Takashima et al., 2009). It has been proposed that retrieval practice may initiate a more rapid consolidation process, leading to similar structural and functional changes as a night of sleep (Antony et al., 2017; Brodt et al., 2018). Future research using functional imaging and event-related electroencephalography will be able to shed further light on how sleep and retrieval practice reshape the neural underpinnings of memory. In addition, we contrasted retrieval practice with a no retrieval, listen-only condition, which differs from other studies of the testing effect that compare retrieval practice to restudy (Bäuml et al., 2014). This design feature makes it harder to directly compare our results to previous work investigating the interaction between sleep and retrieval practice.

Although our sample size is consistent with many other sleep and memory studies (e.g. (Denis et al., 2020; Klinzing et al., 2021; Noack et al., 2021; Payne, Stickgold, et al., 2008; Scullin & McDaniel, 2010), we did not perform an *a priori* power analysis due to the lack of a known effect size for this paradigm. Although this is a limitation of the present work, the reported data provides an important baseline for which future high-powered studies can estimate expected effect sizes from.

Bartlett's War of the Ghosts remains a classic demonstration of how memories are transformed and become distorted over time. Here, we investigated the specific roles of sleep and retrieval practice on these processes. We found that both sleep and retrieval practice strengthened accurate recall of the story, while also inducing an increase in story-related inferences. These findings illustrate the importance of both processes in memory strengthening and transformation, as well as highlighting the importance for future work to further examine when these two processes confer similar effects on memory, and under what conditions they may differ.

Open practices statement

The experiment reported in this article was not preregistered. Data and analysis code needed to reproduce the results reported in this manuscript are available on the Open Science Framework (https://osf.io/3b4hg/)

References

Abel, M., Haller, V., Köck, H., Pötschke, S., Heib, D., Schabus, M., & Bäuml, K.-H. T. (2019). Sleep reduces the testing effect-But not after corrective feedback and prolonged retention interval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(2), 272–287. https://doi.org/10.1037/xlm0000576

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Alger, S. E., & Payne, J. D. (2016). The differential effects of emotional salience on direct associative and relational memory during a nap. *Cognitive, Affective & Behavioral Neuroscience*, *16*(6), 1150–1163. https://doi.org/10.3758/s13415-016-0460-1

Aly, M., & Moscovitch, M. (2010). The effects of sleep on episodic memory in older and younger adults. *Memory*, *18*(3), 327–334. https://doi.org/10.1080/09658211003601548

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a Fast Route to Memory Consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Antony, J. W., & Paller, K. A. (2018). Retrieval and sleep both counteract the forgetting of spatial information. *Learning & Memory*, *25*(6), 258–263. https://doi.org/10.1101/lm.046268.117

Bartlett, F. C. (1932). *Remembering: A study in Experimental and Social Psychology*. Cambridge University Press.

Bäuml, K.-H. T., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect: It enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1568–1581. https://doi.org/10.1037/xlm0000025

Bergman, E. T., & Roediger, H. L. (1999). Can Bartlett's repeated reproduction experiments be replicated. *Memory & Cognition*, *27*(6), 937–947.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*, 41–100. https://doi.org/10.1016/s0065-2407(02)80062-3

Brodt, S., Gais, S., Beck, J., Erb, M., Scheffler, K., & Schönauer, M. (2018). Fast track to the neocortex: A memory engram in the posterior parietal cortex. *Science*, *362*(6418), 1045–1048. https://doi.org/10.1126/science.aau2528

Denis, D., Schapiro, A. C., Poskanzer, C., Bursal, V., Charon, L., Morgan, A., & Stickgold, R. (2020). The roles of item exposure and visualization success in the consolidation of memories across wake and sleep. *Learning & Memory*, *27*(11), 451–456. https://doi.org/10.1101/lm.051383.120

Doxey, C. R., Hodges, C. B., Bodily, T. A., Muncy, N. M., & Kirwan, C. B. (2018). The effects of sleep on the neural correlates of pattern separation. *Hippocampus*, *28*(2), 108–120. https://doi.org/10.1002/hipo.22814

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. https://doi.org/10.1073/pnas.0700094104

Ellenbogen, J. M., Payne, J. D., & Stickgold, R. (2006). The role of sleep in declarative memory consolidation: Passive, permissive, active or none? *Current Opinion in Neurobiology*, *16*(6), 716–722. https://doi.org/10.1016/j.conb.2006.10.006

Fenn, K. M., & Hambrick, D. Z. (2013). What drives sleep-dependent memory consolidation: Greater gain or less loss? *Psychonomic Bulletin & Review*, *20*(3), 501–506. https://doi.org/10.3758/s13423-012-0366-z

Fruzzetti, A. E., Toland, K., Teller, S. A., & Loftus, E. F. (1992). Memory and eyewitness testimony. In *Aspects of memory: The practical aspects, Vol. 1, 2nd ed* (pp. 18–50). Taylor & Frances/Routledge.

Hanert, A., Weber, F. D., Pedersen, A., Born, J., & Bartsch, T. (2017). Sleep in humans stabilizes pattern separation performance. *The Journal of Neuroscience*, *37*(50), 12238–12246. https://doi.org/10.1523/JNEUROSCI.1189-17.2017

Hoddes, E., Dement, W., & Zarcone, V. (1972). The development and use of the stanford sleepiness scale (SSS). *Psychophysiology*, *9*, 150.

Huguet, M., Payne, J. D., Kim, S. Y., & Alger, S. E. (2019). Overnight sleep benefits both neutral and negative direct associative and relational memory. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(6), 1391–1403. https://doi.org/10.3758/s13415-019-00746-8

Klinzing, J. G., Nienborg, H., & Rauss, K. (2021). Sleep does not aid the generalisation of binocular disparity-based learning to the other visual hemifield. *Journal of Sleep Research*, e13335. https://doi.org/10.1111/jsr.13335

Klinzing, J. G., Niethard, N., & Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, *22*(10), 1598–1610. https://doi.org/10.1038/s41593-019-0467-3

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J., & Rauss, K. (2017). Sleep Supports the Slow Abstraction of Gist from Visual Perceptual Memories. *Scientific Reports*, *7*(1), 42950. https://doi.org/10.1038/srep42950

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, *9*, 111–151.

Muehlroth, B. E., Sander, M. C., Fandakova, Y., Grandy, T. H., Rasch, B., Lee Shing, Y., & Werkle-Bergner, M. (2020). Memory quality modulates the effect of aging on memory consolidation during sleep: Reduced maintenance but intact gain. *NeuroImage*, *209*, 116490. https://doi.org/10.1016/j.neuroimage.2019.116490

Noack, H., Doeller, C. F., & Born, J. (2021). Sleep strengthens integration of spatial memory systems. *Learning & Memory*, *28*(5), 162–170. https://doi.org/10.1101/lm.053249.120

Payne, J. D. (2011). Sleep on it!: Stabilizing and transforming memories during sleep. *Nature Neuroscience*, *14*(3), 272–274. https://doi.org/10.1038/nn0311-272

Payne, J. D., Ellenbogen, J. M., Walker, M. P., & Stickgold, R. (2008). The role of sleep in memory consolidation. In *Learning and memory: A comprehensive reference: Vol. 2. Cognitive psychology of memory* (pp. 663–685). Elsevier.

Payne, J. D., & Kensinger, E. A. (2011). Sleep leads to changes in the emotional memory trace: Evidence from FMRI. *Journal of Cognitive Neuroscience*, *23*(6), 1285–1297. https://doi.org/10.1162/jocn.2010.21526

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science*, *19*(8), 781–788. https://doi.org/10.1111/j.1467-9280.2008.02157.x

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*(2), 681–766. https://doi.org/10.1152/physrev.00032.2012

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & McDermott, K. B. (2000). Distortions of memory. In *The Oxford handbook of memory* (pp. 149–162). Oxford University Press.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sanders, K. E. G., Osburn, S., Paller, K. A., & Beeman, M. (2019). Targeted Memory Reactivation During Sleep Improves Next-Day Problem Solving. *Psychological Science*, *30*(11), 1616–1624. https://doi.org/10.1177/0956797619873344

Schoch, S. F., Cordi, M. J., & Rasch, B. (2017). Modulating influences of memory strength and sensitivity of the retrieval test on the detectability of the sleep consolidation effect. *Neurobiology of Learning and Memory*, *145*, 181–189. https://doi.org/10.1016/j.nlm.2017.10.009

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Scullin, M. K., & McDaniel, M. A. (2010). Remembering to Execute a Goal: Sleep on It! *Psychological Science*, *21*(7), 1028–1035. https://doi.org/10.1177/0956797610373373

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernández, G. (2009). Shift from Hippocampal to Neocortical Centered Retrieval Network with Consolidation. *Journal of Neuroscience*, *29*(32), 10087–10093. https://doi.org/10.1523/JNEUROSCI.0799-09.2009

Talamini, L. M., Nieuwenhuis, I. L. C., Takashima, A., & Jensen, O. (2008). Sleep directly following learning benefits consolidation of spatial associative memory. *Learning & Memory*, *15*(4), 233–237. https://doi.org/10.1101/lm.771608

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *The Journal of Neuroscience*, *30*(43), 14356–14360. https://doi.org/10.1523/JNEUROSCI.3028-10.2010

Tilley, A., & Empson, J. (1978). REM sleep and memory consolidation. *Biological Psychiatry*, *6*, 293–300.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*(6972), 352–355. https://doi.org/10.1038/nature02223

Wagoner, B., & Gillespie, A. (2014). Sociocultural mediators of remembering: An extension of Bartlett's method of repeated reproduction. *British Journal of Social Psychology*, *53*(4), 622–639. https://doi.org/10.1111/bjso.12059

Webb, C. E., Turney, I. C., & Dennis, N. A. (2016). What's the gist? The influence of schemas on the neural correlates underlying true and false memories. *Neuropsychologia*, *93*, 61–75. https://doi.org/10.1016/j.neuropsychologia.2016.09.023

Werchan, D. M., & Gómez, R. L. (2013). Generalizing memories over time: Sleep and reinforcement facilitate transitive inference. *Neurobiology of Learning and Memory*, *100*, 70–76. https://doi.org/10.1016/j.nlm.2012.12.006

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*(1), 8–14. https://doi.org/10.1016/j.jarmac.2014.07.003

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *BioRxiv*, 2021.01.05.425378. https://doi.org/10.1101/2021.01.05.425378

Supplementary Materials

Table S1. Percentage of propositions exhibiting each memory category at the initial memory test

	Wake	Sleep
	M (SD)	M (SD)
Accuracy	11.57 (8.54)	10.30 (7.12)
Gist	36.92 (10.39)	36.40 (8.67)
Omission	19.35 (7.27)	18.60 (7.61)
Distortion (inference)	9.66 (6.73)	9.08 (6.49)
Distortion (normalization)	4.93 (4.30)	5.19 (4.77)
Distortion (incorrect placement)	1.28 (2.33)	2.53 (2.75)
Importation	16.30 (11.70)	17.90 (8.28)

Note. M = mean, SD = standard deviation

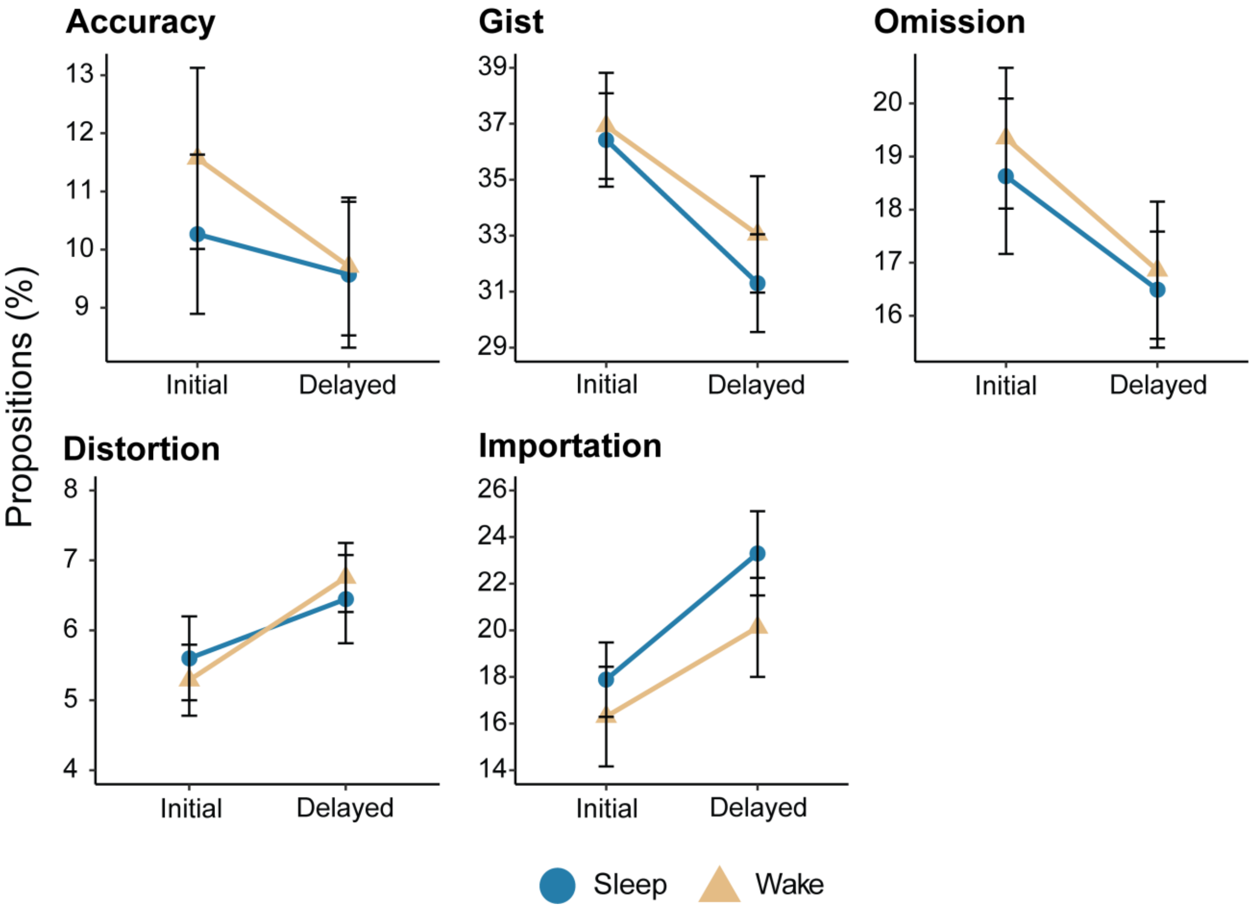

Figure S1. Change in memory between the initial and delayed tests within the two retrieval practice groups. Error bars display the standard error.

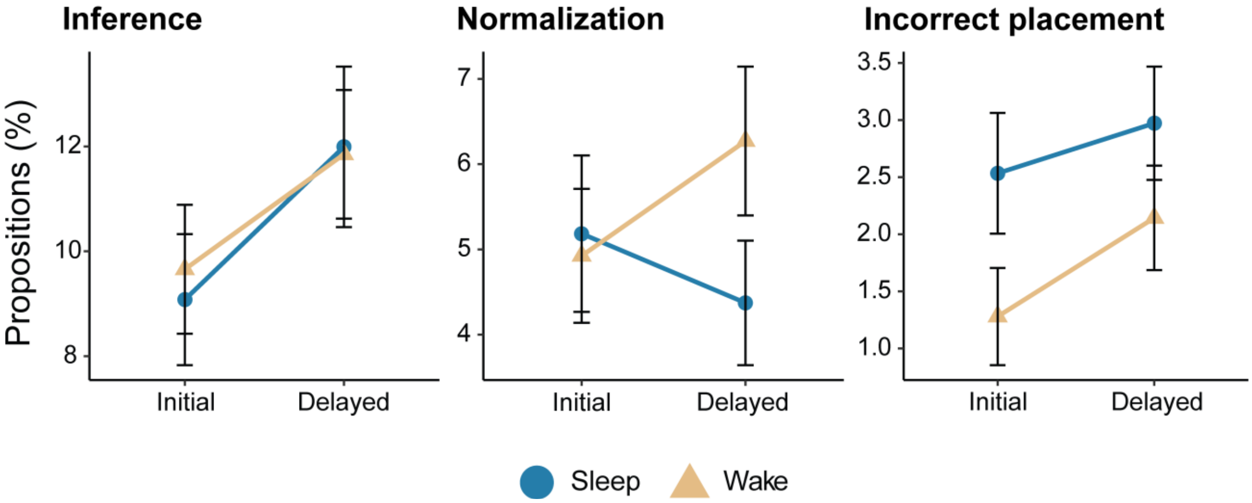

Figure S2. Change in distortion subtypes between the initial and delayed tests within the two retrieval practice groups. Errors bars display the standard error.