Sleep, relative to wake, increases both veridical and false memory in the DRM paradigm: A registered report

Matthew H.C. Mak, Alice O'Hagan, Aidan J. Horner, M. Gareth Gaskell

Department of Psychology, University of York, UK

Author Note

We have no conflict of interest to disclose. Correspondence concerning this article should be addressed to Matthew Mak, Department of Psychology, University of York, Heslington, York, YO10 5DD, United Kingdom; Email: matthew.mak@york.ac.uk

Acknowledgements

This research was supported by a BA/Leverhulme Small Research Grant (Number: SRG21\210150) awarded to Matthew Mak and Gareth Gaskell, who were also supported by a grant from the Economic and Social Research Council (ESRC; ES/T008571/1). Aidan Horner was supported by an ESRC grant (ES/R007454/1). The funders have no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We would like to express our gratitude to members of the SLAM group for their valuable discussions regarding this work, as well as extend our thanks to the four reviewers for providing constructive feedback on this study.

Open Science Statement

All the materials, data, and analysis scripts are publicly available on Open Science Framework (https://osf.io/9pdyf/).

Abstract

Human memory is known to be supported by sleep. However, less is known about the effect of sleep on false memory, where people remember events that never occurred. In the laboratory, false memories are often induced via the Deese-Roediger-McDermott (DRM) paradigm where participants are presented with semantically related words such as nurse, hospital, and sick (studied words). Subsequently, participants are likely to falsely remember that a related lure word such as doctor was presented. Multiple studies have examined whether these false memories are influenced by sleep, with contradictory results. A recent meta-analysis suggests that sleep may increase DRM false memory when short lists are used. We tested this in a registered report (N=488) with a 2 (Interval: Immediate vs. 12-hr Delay) × 2 (Test Time: 9AM vs. 9PM) between-participant DRM experiment, using short DRM lists (N = 8 words/list) and free recall as the memory test. We found that the Sleep and Wake participants were well-matched on the number of total responses, but those in the Sleep group produced fewer intrusions (i.e., words that were neither studied nor lure words), and when this was statistically controlled for, they showed a greater tendency to recall more critical lures as well as more studied items. Our findings support the view that sleep may facilitate gist abstraction and/or spreading activation, alongside strengthening/protecting newly encoded declarative memories.

Keywords: Sleep, False Memory, DRM, Recall, Gist Abstraction, Spreading Activation

Experiment

1 Overview

It is possible to index DRM false memory via free recall or recognition (e.g., Stadler et al., 1999). In this experiment, we used free recall only, because recall tends to be more prone to sleep-related memory effects than recognition [1–4]. Our experiment comprised a study and a test phase. In the study phase, a participant encoded 20 short DRM wordlists, with each containing 8 words. Short lists were chosen because Newbury and Monaghan's [1] meta-analysis pinpointed a clear sleep effect in these lists. In the test phase, participants recalled the wordlists in a free recall procedure.

Participants were randomly assigned to one of the four groups: AM-control, PM-control, Sleep, or Wake. Those assigned to the control (aka Immediate) groups carried out the test phase immediately after the study phase, with those in the AM group starting at 9AM (± 1hr) and those in the PM group starting at 9PM (± 1hr). No difference in false or veridical recall was expected between these groups, as prior DRM studies (e.g., [5–7]) have consistently demonstrated that immediate recall was equivalent between morning and evening. The inclusion of these control groups helped rule out potential circadian effects on encoding and retrieval (and relatedly, monitoring in the Activation/Monitoring Framework). Finally, participants assigned to the Sleep and Wake groups (collectively referred to as the Delay groups) started the test phase approximately 12 hours after the study phase. Those in the Wake group studied the DRM wordlists in the morning (9AM ± 1hr) and completed the test phase in the evening (9PM ± 1hr) on the same day. Those in the Sleep group encoded the wordlists in the evening (9PM ± 1hr) and completed the test phase in the morning (9AM ± 1hr) the next day.

2 Research questions and corresponding predictions

This experiment set out to address a key question:

#1 Does overnight sleep (vs. daytime wakefulness) influence DRM false recall?

Our prediction was based on the meta-analysis by Newbury and Monaghan [1], who reported that when a study used short lists, sleep consistently increased DRM false memory. We, therefore, predicted a post-sleep (vs. post-wake) increase in DRM false recall, whereas there would be no such difference between the AM- and PM-control groups.

Our study also addressed a peripheral question:

#2 Does overnight sleep (vs. daytime wakefulness) increase veridical recall of the studied list words?

Again, our prediction is based on Newbury and Monaghan [1], who found that sleep benefits veridical memory in short lists. Our prediction to this question was, therefore that veridical recall would be greater post-sleep than post-wake, whereas there would be no such difference between the AM- and PM-control groups.

3 Design

#1 Does overnight sleep (vs. daytime wakefulness) influence DRM false recall?

For this question, the dependent variable was whether a critical lure is recalled or not (i.e., binary). There were two independent variables: Interval (Immediate vs. Delay) and Test Time (9AM vs. 9PM), both of which were manipulated between-participants. In other words, the four groups were coded as in Table 1:

Table 1: How the four groups were coded using Interval and Test Time.

Groups		Interval		Test Time
AM-control	=	Immediate	+	9AM
PM-control	=	Immediate	+	9PM
Sleep	=	Delay	+	9AM
Wake	=	Delay	+	9PM

To address Research Question #1, we first tested if any difference between the Sleep and Wake groups was significantly different from that between the AM- and PM-control groups (i.e., an interaction between Interval and Test Time). This is important because it allows us to rule out time-of-day effects. Then, we tested for the simple effect of Test Time (9AM vs. 9PM) within the Immediate and Delay groups. If there is (1) a significant Interval x Test Time interaction and (2) a significant Test Time effect within the Delay groups (Sleep > Wake), we will be able conclude that sleep (but not time-of-day) increases false recall.[1]

#2 Does overnight sleep (vs. daytime wakefulness) increase veridical recall of the studied list words?

For this research question, the dependent variable was whether a studied list word was recalled or not (i.e., binary). As per Question #1, there were two between-participant manipulations: Interval (Immediate vs. Delay) and Test Time (9AM vs. 9PM). We first tested if there was an interaction between Interval and Test Time. Then, we tested for the simple effect of Test Time within the Immediate and Delay groups. Note that this research question is secondary to the first.

4 Target sample size and stopping rules

Our target sample size was 120 participants/group (i.e., 480 participants in total), defined as those who remained in the sample after applying the exclusion criteria outlined in section 9. This sample size gives us >=90% power to detect all the desired effects for our Research Questions (See Appendix D for a detailed power analysis).

5 Recruitment

5.1 Online recruitment.

Participants were recruited online via Prolific (https://www.prolific.co/). All participants completed the experiment unsupervised and at a location of their own choosing. We chose online testing, as opposed to lab-based testing, for at least two reasons. First, given the unpredictability of the COVID-19 pandemic, we did not want to risk the possibility of data collection being disrupted. Second, given the time limit on the funding for this work, it would have been logistically difficult to reach the target sample size were the study conducted in person.

One key concern associated with online testing is data quality. This stems from the fact that researchers cannot monitor participants during an online experiment. However, it has been repeatedly demonstrated that as long as appropriate measures are taken (e.g., [8,9]), data quality from online experiments is no different from lab-based experiments (e.g., [10–13]). Furthermore, two recent online studies using the same experimental design [14,15] found clear evidence of a sleep benefit in the classic paired-associate learning paradigm, replicating well-established evidence from lab-based experiments (e.g., [16,17]). Importantly, the

[1]Prior studies in the 'Sleep × DRM' literature (e.g., Fenn et al., 2009; Payne et al., 2009) conducted two separate statistical tests, one comparing Sleep vs. Wake, another comparing AM- vs. PM- controls. They then concluded that Sleep had an effect on DRM false memory beyond time-of-day effects when Test Time was significant in the Sleep vs. Wake comparison ($p < 0.05$) but not in the AM vs. PM comparison ($p > 0.05$). Unfortunately, however, this is not sufficient (Nieuwenhuis et al., 2011), as "the difference between 'significant' and 'not significant' is not itself statistically significant" (Gelmam & Stern, 2006). Therefore, in order to rule out time-of-day effects, one needs to show that Sleep vs. Wake is significantly different from AM vs. PM-control. This can be captured by an Interval × Test Time interaction.

effect sizes for sleep from these online studies were roughly equivalent to those reported by lab-based studies. Together, these suggest that it is possible to detect sleep-related memory effects in online experiments, as long as the appropriate measures are put in place. These are detailed in the Procedures section (#8) below.

5.2 Recruitment method.

Following two previous sleep studies conducted via Prolific [15,18], we put a short survey on the platform to recruit a pool of participants (N = 2296). This survey is available in Appendix A and was hosted on Qualtrics. The first half of the survey asked for basic demographic information: gender identity, age, current country of residence, first language, ethnicity, highest education attainment, and history of developmental/sleep disorders (if any). The survey then provided a brief outline of the main study. It stated that if enrolled, participants would be randomly allocated to one of the four groups and that no preferences would be accommodated. Participants then indicated whether they would like to enrol in the main study. Of the 2296 respondents, 1940 expressed interest in taking part, who were then screened for their eligibility (see inclusion criteria below). Those who fitted our inclusion criteria were then randomly allocated to one of the four experimental groups. A private message was sent to each participant, notifying them of their group allocation. In the end, 534 participants completed both the study and test phases. These participants were reimbursed at a rate of ~£9.5/hr.

6 Inclusion criteria

We applied these inclusion criteria to ensure comparability with prior studies (e.g., [5,7,19,20]):

1. Aged 18-25
2. Speaks English as (one of) their first language(s)
3. No known history of any psychiatric (e.g., schizophrenia), developmental (e.g., dyslexia) or sleep (e.g., insomnia) disorders
4. Currently resides in the UK, indexed by their IP address (since this experiment requires participants to complete each phase at a certain time of day, it is necessary to restrict the location to prevent participants from taking the study in different time zones)
5. Normal vision or corrected-to-normal vision
6. Normal hearing
7. Able to complete the study using a laptop or a desktop PC
8. Able to complete both the study and test phases
9. Has an approval rate of >96% on Prolific. This helps ensure that a participant has a tendency to take online studies seriously.

7 Materials

Prior studies in the DRM literature typically showed 8 to 15 words per list (e.g., [5,20,21]). Generally, within this range, showing fewer words reduces false recall rates ([21,22]; see also [23]). However, showing even fewer words per list (e.g., 3) results in floor or near-floor rates [22]. Given that sleep seems to have a larger effect on false memory when the gist trace or lure is encoded at a medium level during study [1], we opted for 8 words per list.

We made use of 20 DRM wordlists (see Table 2), taken from Roediger et al. [24]. Each list contained 8 semantically related words, and as per the standard DRM paradigm, they were arranged in a descending order of associative strength to the critical lures. A participant studied all 20 lists. We note that the original DRM lists by Roediger et al. [24] were tailored for American participants, and two words (e.g., trash, Mississippi)

were not immediately relatable to people in the UK. We, therefore, changed these words (e.g., trash → rubbish), as noted in Table 2.

We acknowledge that previous studies in the 'Sleep × DRM' literature typically showed participants 8 to 16 lists (e.g., [7,19]), so our participants studied more wordlists (i.e., 20). However, since we showed relatively few words per list, the total number of studied words was comparable to prior studies (i.e., 160 in the current vs. 96 to 225 in prior studies). Furthermore, an advantage of showing more wordlists is that more critical lures could be recalled (i.e., 20 lists = 20 lures), potentially increasing variability between participants and hence our ability to detect sleep-related effects.

8 Procedures

The procedure of the study is summarised in Figure 1. The study was hosted on Gorilla (www.gorilla.sc; [10]). A study phase took approximately 11 minutes. Here, participants first gave informed consent, completed a language/attention check, rated their level of sleepiness on the Stanford Sleepiness Scale (SSS; [25]), and viewed 20 DRM wordlists.

Immediately afterwards, participants in the AM/PM-control groups carried out the test phase. For those in the Delay groups, the test phase took place approximately 12 hours later. Here, both the Immediate and Delay participants rated their level of sleepiness on SSS and completed a short survey concerned with, for example, morningness/eveningness preference (rMEQ; [26]) and sleep duration/quality the night before (see Appendix B for the full survey). This survey helped determine whether the four groups were matched in terms of time-of-day preference and whether data from a participant needed to be discarded as a result of meeting the exclusion criteria described in section 9. Finally, the test phase concluded with a 10-minute free recall task where participants recalled as many of the words as they could from the previously seen wordlists.

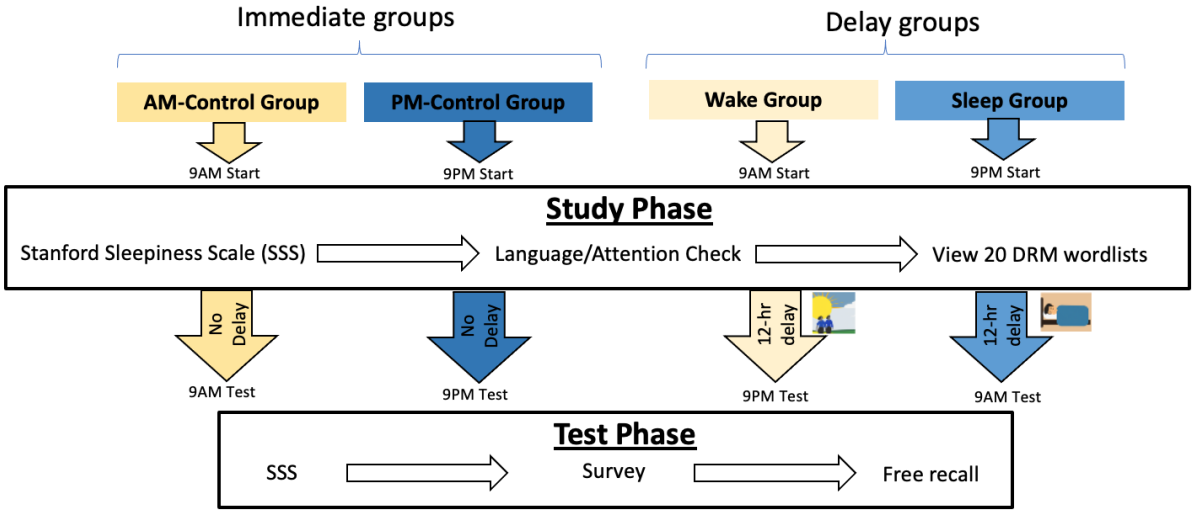

Figure 1: Experimental procedure.

8.1 Exposure to the DRM wordlists.

On the instruction page, participants were told that they would see some English words presented one after the other on the computer screen. They were asked to pay close attention to the words because they would be tested on them later on. No specific instruction was given regarding the subsequent test format.

During presentation, words in each DRM list were presented visually,[2] in a fixed order and arranged in descending associative strength to the unpresented critical lure (see Table 2 for order). Each list began with

[2]Newbury and Monaghan [@newbury2019a] found no evidence in their meta-analysis that the modality of presentation modulated the effect of sleep. However, in a set of three experiments, Fenn et al. [@fenn2009a] used auditory presentation in one

a fixation for 1 s, followed by the first word in a list. Each word was shown for 1 s, in a lowercase black font (Arial, size 26) on a white background, and separated by a 500-ms interstimulus interval. After presentation of the final word in a list was 5 s of blank screen. List order was randomised, and each list was seen once.

Table 2: The 20 DRM wordlists to be used.

Critical lure of each list	**False recall probability (Roediger et al. 2001)**	**List items (arranged in the order of presentation in study)**
Window	*65*	*door, glass, pane, shade, ledge, sill, house, open*
Sleep	*61*	*bed, rest, awake, tired, dream, wake, snooze, blanket*
Doctor	*60*	*nurse, sick, lawyer, medicine, health, hospital, dentist, physician*
Smell	*60*	*nose, breathe, sniff, aroma, hear, see, nostril, whiff*
Chair	*54*	*table, sit, legs, seat, couch, desk, recliner, sofa*
Smoke	*54*	*cigarette, puff, blaze, billows, pollution, ashes, cigar, chimney*
Sweet	*54*	*sour, candy, sugar, bitter, good, taste, tooth, nice*
Rough	*53*	*smooth, bumpy, road, tough, sandpaper, jagged, ready, coarse*
Needle	*52*	*thread, pin, eye, sewing, sharp, point, prick, thimble*
Rubbish	*49*	*garbage, waste, can, refuse, sewage, bag, junk, trash (Note 1)*
Anger	*49*	*mad, fear, hate, rage, temper, fury, ire, wrath*
Soft	*46*	*hard, light, pillow, plush, loud, cotton, fur, touch*
City	*46*	*town, crowded, state, capital, streets, subway, country, New York*
Cup	*45*	*mug, saucer, tea, measuring, coaster, lid, handle, coffee*
Cold	*44*	*hot, snow, warm, winter, ice, wet, frigid, chilly*
Mountain	*42*	*hill, valley, climb, summit, top, molehill, peak, plain*
Slow	*42*	*fast, lethargic, stop, listless, snail, cautious, delay, traffic*
River	*42*	*water, stream, lake, Thames (Note 2), boat, tide, swim, flow*
Spider	*37*	*web, insect, bug, fright, fly, arachnid, crawl, tarantula*
Foot	*35*	*shoe, hand, toe, kick, sandals, soccer, yard, walk*

Note 1. In Roediger et al. (2001), the critical lure for this list was trash, with rubbish being one of the list items. We used rubbish as the critical lure and trash as a list item because the former is the preferred term in British English.

Note 2. The original word in Roediger et al. was Mississippi. We replaced it with Thames.

There was a surprise attention check after the 4th, 9th, 13th, 18th lists, where participants saw an erroneous maths equation such as "3 + 3 = 11". It was presented for 1 s, in the same font and style as the list words [27]. Immediately afterwards, participants were asked to report what 3 + 3 was according to what was just shown.

8.2 Free Recall.

Participants had 10 mins to type out all the words they could remember from the study phase in a textbox. When there was 2 min left, a timer appeared. Participants could not proceed before the time was up. To maximise the likelihood that participants paid full attention instead of doing something else (e.g., playing with their phone) during recall, there was an attention check throughout: On the same page as the response textbox, there was a white square that turned red every 2 to 3 mins. The change in colour lasted for 10 s, during which a single digit was shown. Participants had to enter the digit into a separate textbox to show that they were paying attention. Throughout the 10-min recall task, the square turned red four times, so participants needed to enter four digits as they attempted the recall task.

of them and visual in the other two. As far as we are aware, this is the only study in the 'Sleep × DRM' literature that had used both modalities in the same set of experiments. Sleep appeared to have a larger effect on false memory when visual (vs. auditory) presentation was used. Given this, we opted for visual presentation in the current experiment.

8.3 Additional measures to ensure data quality.

At the start of the study phase, participants were encouraged to take the experiment seriously and were informed that their participation would contribute to science. After rating their level of sleepiness, participants must pass a language/attention check. This involved the auditory presentation of a short story. Replay and pausing were not permitted. Participants then answered two simple comprehension questions based on the story. Failure to answer both questions correctly led to their data being excluded from further analysis. These questions helped to ensure that participants could indeed understand English and were in a reasonably quiet environment. Next, to prevent participants from multitasking on the computer, both the study and test phases required participants to enter full-screen mode. Participants were told that exit from full-screen mode during the study may lead to no payment. This was made possible by Gorilla, which recorded the browser's and the monitor's sizes. At the end of the study phase, participants were asked to describe how they learnt the words in a sentence. Participants who said they wrote down or similarly recorded the words were excluded from further analysis.

9 Exclusion criteria

Exclusion was applied on the participant level. A participant's dataset was excluded from further analysis and replaced, if

1. they exited full-screen mode in any of the phases.
2. they failed the language/attention check at the start of the study.
3. they reported to have written down or recorded the wordlists during the study phase.
4. (Sleep and Wake groups only) they reported consuming any alcoholic drinks between study and test.
5. (Sleep group only) they reported to have had fewer than 6 hours of overnight sleep prior to test or rated their sleep quality as poor or extremely poor.
6. (Wake group only) they reported to have had a nap between study and test.
7. they failed more than one of the four attention checks (i.e., 3 + 3 = 11) at study.
8. they failed to report more than one of the four digits in the attention check of free recall.
9. they submitted a blank response in free recall.
10. (this criterion was removed; see Deviations from Stage-1 registered report on p. 1 for explanation)
11. their completion time for either the study or test phase is 3 standard deviations above or below their respective mean completion time of the first 480 participants who completed the study.

10 Participants

Of the 534 participants who completed both the study and test phases, 46 were excluded for meeting one or more of the exclusion criteria. The full list of excluded participants and their respective reason for exclusion is available on OSF (see exclusion_OSF.csv). Our final sample size comprised 488 participants, with 124 in each of the Immediate groups, and 120 in each of the Delay groups. Group characteristics are summarised in Table 3.

Table 3: Group characteristics.

Characteristics	Immediate -AM	Immediate -PM	Sleep (aka Delay-AM)	Wake (aka Delay-PM)
N before exclusion	130	127	135	143
N after exclusion	124	124	120	120
Mean age (SD)	22.24 (2.18)	22.34 (2.03)	22.18 (1.93)	22.25 (1.93)
Gender (Female:Male:Other)	64 : 54 : 2	77 : 46 : 1	62 : 57 : 1	58 : 61 : 1
% participants identified as ethnically white	78.2%	73.4%	81.7%	80%
Mean SSS rating at study (SD)	2.58 (0.98)	2.64 (1.12)	2.66 (0.96)	2.58 (0.98)
Mean SSS rating at test (SD)	2.73 (1.04)	2.95 (1.29)	2.63 (1.21)	2.67 (1.18)
Mean rMEQ score (SD)	15.89 (1.67)	15.59 (1.91)	15.72 (1.83)	15.53 (1.99)
Mean N of intervening hr between study and test (SD)	NA	NA	12.22 (0.74)	12.14 (0.81)

Notes. (1) SSS stands for Stanford Sleepiness Scale and ranges from 1 to 6, with higher values indicating greater sleepiness. (2) rMEQ stands for reduced Morningness/Eveningness Questionnaire; it ranges from 1 to 25, with higher values indicating greater morningness preference.

Prior to the confirmatory analyses, we first checked if the four groups were matched on their morningness/eveningness preference and degree of sleepiness at study/test (as indexed by the Stanford Sleepiness Scale). These are summarised in Table 3. We compared the four groups on each of the measures using one-way ANOVAs, which showed no significant differences (SSS at study: $F = 0.21$, $p = .888$; SSS at test: $F = 1.63$, $p = .183$; rMEQ: $F = 0.97$, $p = .408$). We also compared the Sleep and Wake groups on the number of intervening hours between study and test using a between-participant t-test, which revealed no significant difference [$t(236.16) = 0.86$, $p = .389$]. In sum, our four groups were well-matched on these potentially confounding factors.

11 Data pre-processing

The free recall data were pre-processed. The first step was to remove any duplicate responses. The second was to correct all obvious spelling and typing errors to the nearest English words, defined as Levenstein distance ≤ 2 (e.g., *cigerette $\rightarrow$ cigarette).[3] Responses with added or dropped inflectional suffixes (i.e., -s, -ed, -ing, adjectival -er) were corrected. Responses with derivational changes were considered as intrusions. For instance, one of the studied words is pollution; if a participant recalled pollutions, the plural suffix was dropped; however, if a participant recalled pollutant, this was considered as an intrusion.

12 Results of Pre-registered Analyses

Following prior studies in the 'Sleep $\times$ DRM' literature, we adopted a frequentist approach for all our analyses. The alpha level was set at 0.05.

12.1 Positive control.

We checked if our paradigm consistently elicited the well-established DRM effect across participants. Given free recall, the chance level of a critical lure being produced is 0. We submitted the number of critical lures produced by all participants (Range: 0 to 20) to a one-sample t-test, with the chance level being 0. It showed that participants were susceptible to false recall [$t(487) = 26.96$, <0.001], providing evidence for the classic DRM false memory effect.

[3]If a misspelt response has more than one nearest word, the response was considered as an intrusion.

12.2 Control analysis.

Payne et al. [7] found that participants falsely recalled more critical lures post-sleep (vs. post-wake). However, it is possible that participants simply had a greater tendency to put down more unseen words after sleep, not because sleep increases DRM false recall per se. Therefore, before addressing our key research questions, we checked if participants across groups were comparable in terms of their bias in producing unseen items. In Payne et al. [7], this bias was indexed via the number of intrusions (i.e., neither the studied nor the lure items), which was roughly equivalent between their Sleep and Wake groups [M_{Sleep} = 5.6 vs. M_{Wake} = 6.2; p = .60] as well as between their AM and PM-control groups [MAM = 4.1 vs. MPM = 4.1; p = .99]. To check if this is the case in our data, we used a 2 (Interval: Immediate vs. Delay) × 2 (Test Time: AM vs. PM) Poisson regression. We chose Poisson regression, as opposed to ANOVA, because the intrusion data were count data, meaning that data distribution were right-skewed and hence unsuitable for ANOVA. Figure 2 summarises the number of intrusions in each group.

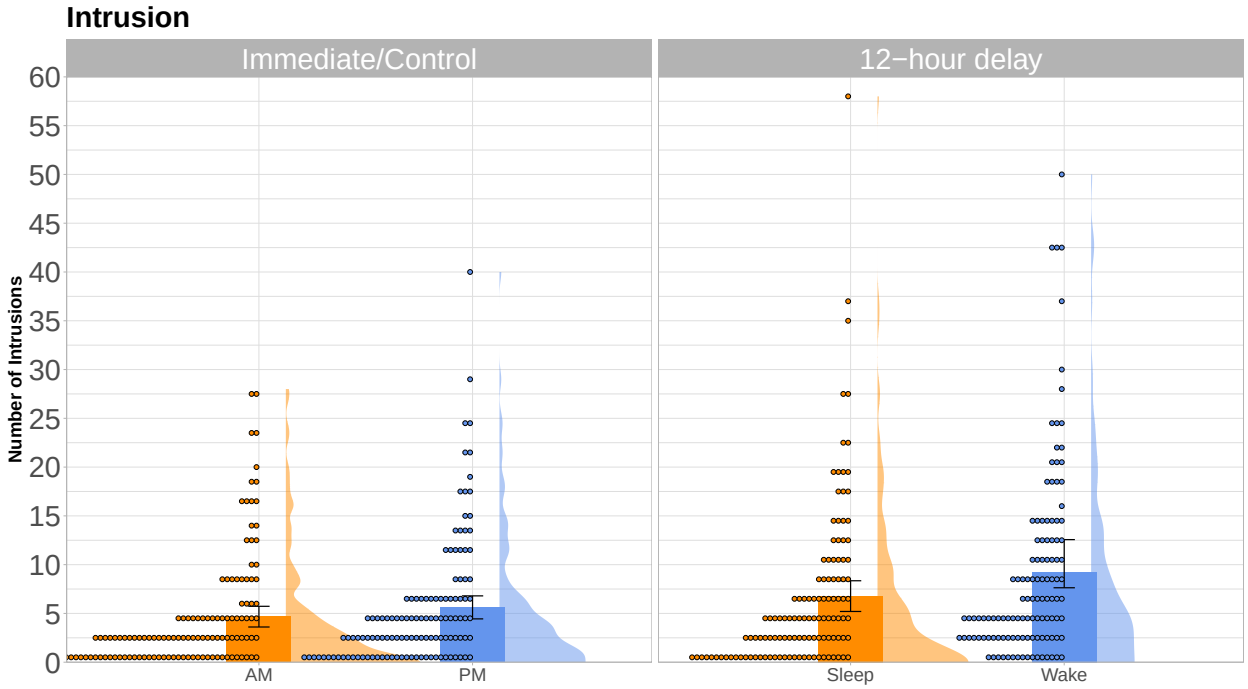

Figure 2: Number of intrusions produced, summarised across the four groups. Each dot represents an individual participant, and the error bars represent 95% confidence intervals.

A 2 × 2 Poisson regression revealed significant effects of Interval (β = 0.372, SE = 0.054, z = 6.845, p < <0.001) and Test Time (β = 0.185, SE = 0.056, z = 3.299, p < <0.001), which were qualified by a significant interaction (β = 0.214, SE = 0.072, z = 2.96, p = 0.003). Given this, we tested the simple effects of Test Time within the Immediate and Delay groups using the emmeans package [28]. Within the Immediate groups, the evening participants (M = 5.62; SD = 6.63) produced more intrusions than the morning participants (M = 4.67; SD = 5.97) (z = -3.299, p = <0.001). Likewise, in the Delay groups, the Wake participants, who completed free recall in the evening (M = 10.1; SD = 13.65), produced more intrusions than the Sleep participants, who completed recall in the morning (M = 6.78; SD = 8.71) (z = -8.808, p = <0.001). Together, our data indicate that participants who attempted free recall in the evening (vs. morning) were more prone to intrusions, and this effect was greater in the Delay than in the Immediate groups. These unexpected findings prompted us to explore whether the number of total responses (i.e., studied + lures + intrusions) differed

between morning and evening test time. Interestingly, this exploratory analysis (see section 13.1) showed no effect of Test Time. Together, these suggest that attempting free recall in the evening led to a selective increase in intrusions, but not necessarily a global increase in output bias. Finally, given that Test Time had a significant effect on intrusions, we followed our pre-registered analysis plan by adding However, neither Test Time nor their interaction is expected to be significant. If in the unlikely event that these are revealed to be significant, the number of intrusions as a numeric covariate in the 2 × 2 mixed-effects models below.

12.3 Confirmatory analysis 1.

This analysis addresses our key research question:

#1 Does overnight sleep (vs. daytime wakefulness) influence DRM false recall?

The number of critical lures falsely recalled is summarised across groups in Figure 3.

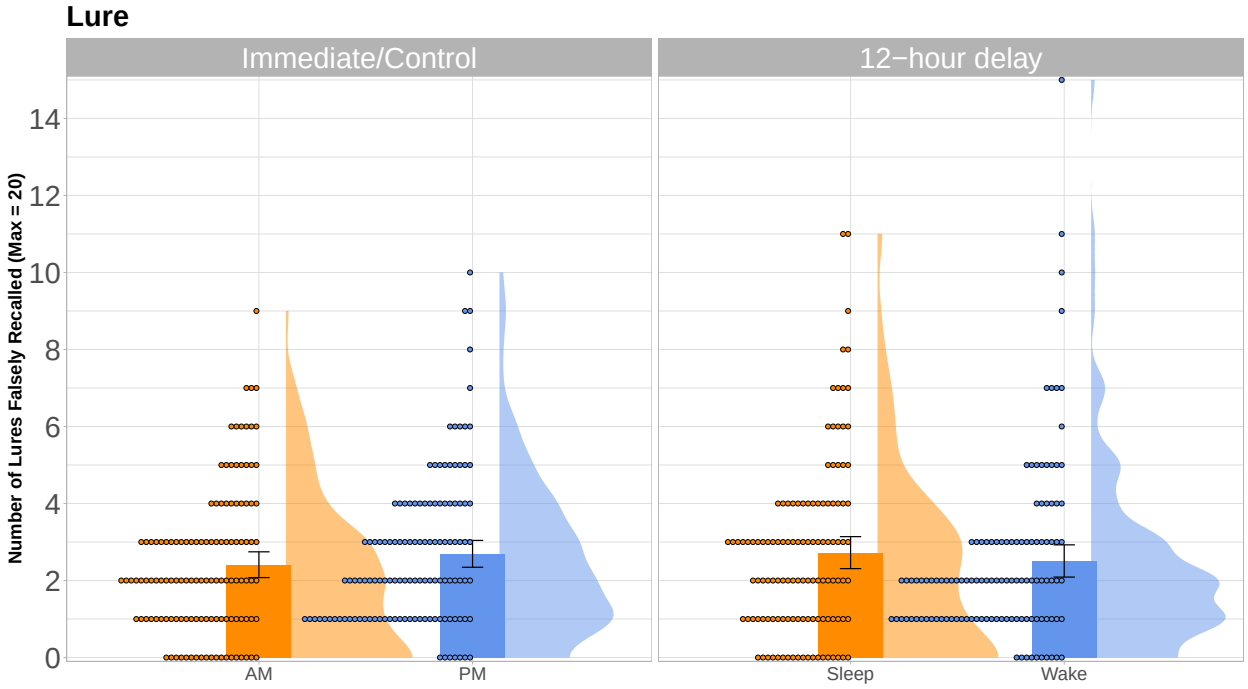

Figure 3: Number of critical lures falsely recalled, summarised across the four groups. Each dot represents an individual participant, and the error bars represent 95% confidence intervals.

A generalised linear mixed-effect model (GLMM) was fitted to the critical lure data on the item level (N of observations = 488‘ participants × 20 critical lures).[4] The dependent variable was binary: whether a critical lure was recalled or not (1 vs. 0). The fixed effects were the number of intrusions a participant produced, Interval (Immediate vs. Delay), Test Time (AM vs. PM), and an Interval by Test Time interaction. Interval and Test Time were coded using sum contrasts [29]. The random-effect structure was determined by the “buildmer” package [30], which automatically found the maximal model that was capable of converging

[4]The use of GLMM is a clear departure from prior ‘Sleep × DRM’ studies, the majority of which addressed the same research question using an independent t-test or ANOVA (e.g., Diekelmann et al., 2010; Monaghan et al., 2017; Payne et al., 2009). We explained in Appendix C why these statistical tests are usually not appropriate in the context of DRM recall and why GLMM are more advantageous.

using backward elimination (with the "bobyqa" optimiser). This means that model selection started from the maximal model, as justified by the experimental design [31]. The model we reported and based our interpretation on was the most maximal model that was capable of converging (see the upper half of Table 4 for the final random-effect structure and model output).

Table 4: Outputs from confirmatory generalised mixed-effect models examining the effects of Intrusions, Interval, and Test Time in false and veridical recall.

False (lure) recall				
Random-effect structure: (Intrusions | Participant.ID) + (1 | Lure)				
	Estimate	*SE*	*z*	*p*
Intercept	-2.328	0.11	-21.254	<.001*
Intrusions	0.034	0.006	5.487	<.001*
Interval (Immediate vs. Delay)	0.039	0.042	0.943	.346
Test Time (AM vs. PM)	0.035	0.041	0.85	.395
Interval x Test Time	-0.084	0.041	-2.046	.041*

Veridical (studied word) recall				
Random-effect structure: (Intrusions | Participant.ID) + (Interval | Studied.Item)				
	Estimate	*SE*	*z*	*p*
Intercept	-2.373	0.075	-31.705	<.001*
Intrusions	-0.007	0.005	-1.372	.17
Interval (Immediate vs. Delay)	0.307	0.04	7.721	<.001*
Test Time (AM vs. PM)	0.026	0.037	0.691	.489
Interval x Test Time	-0.124	0.037	-3.325	<.001*

The number of intrusions had a significant effect on lure recall, such that participants who produced more intrusions tended to recall more critical lures ($z = 5.487$, $p < <.001$). There were no main effects of Interval or Test Time ($zs < 0.95$, $ps > .34$), but there was a significant Interval by Test Time interaction ($z = -2.046$, $p < .041$). Following our pre-registered analysis plan, we proceeded to test the simple effects of Test Time within the Immediate and Delay groups, using the emmeans package [28] in R. Among the Immediate groups, there was no significant difference in lure recall between the AM-control and PM-control participants ($\beta = -0.098$, SE = 0.114, $z = -0.859$, $p = .39$). However, among the Delay groups, there was a significant difference ($\beta = 0.239$, SE = 0.119, $z = 2$, $p = .045$) such that the Sleep participants (M = 2.72, SD = 2.3) reported more critical lures than the Wake participants (M = 2.51, SD = 2.32).

Box 1

R codes for Confirmatory Analysis 1

```
contrasts(FalseRecall$Interval) <- contr.sum(2)  #sum contrast for interval
contrasts(FalseRecall$Test_Time) <- contr.sum(2) #sum contrast for test time
library(lme4)
library(buildmer)
# 2 x 2 GLMM
FalseRecallModel <- buildmer(Recalled ~ Interval * Test_Time + (1 | Participant) +
                    (Interval * Test_Time | Item), data = FalseRecall, family =
                    "binomial",buildmerControl = buildmerControl(direction='backward',
                    args = list(control=glmerControl(optimizer="bobyqa"))))
# Obtain the simple-effects of Test Test within the Immediate and Delay groups
```

```
library(emmeans)
emmeans(FalseRecallModel, pairwise ~ Test_Time | Interval)
```

12.4 Confirmatory analysis 2.

This analysis addresses the secondary question:

#2 Does overnight sleep (vs. daytime wakefulness) increase veridical recall of the studied list words?

Figure 4 summarises the number of studied list words correctly recalled across groups.

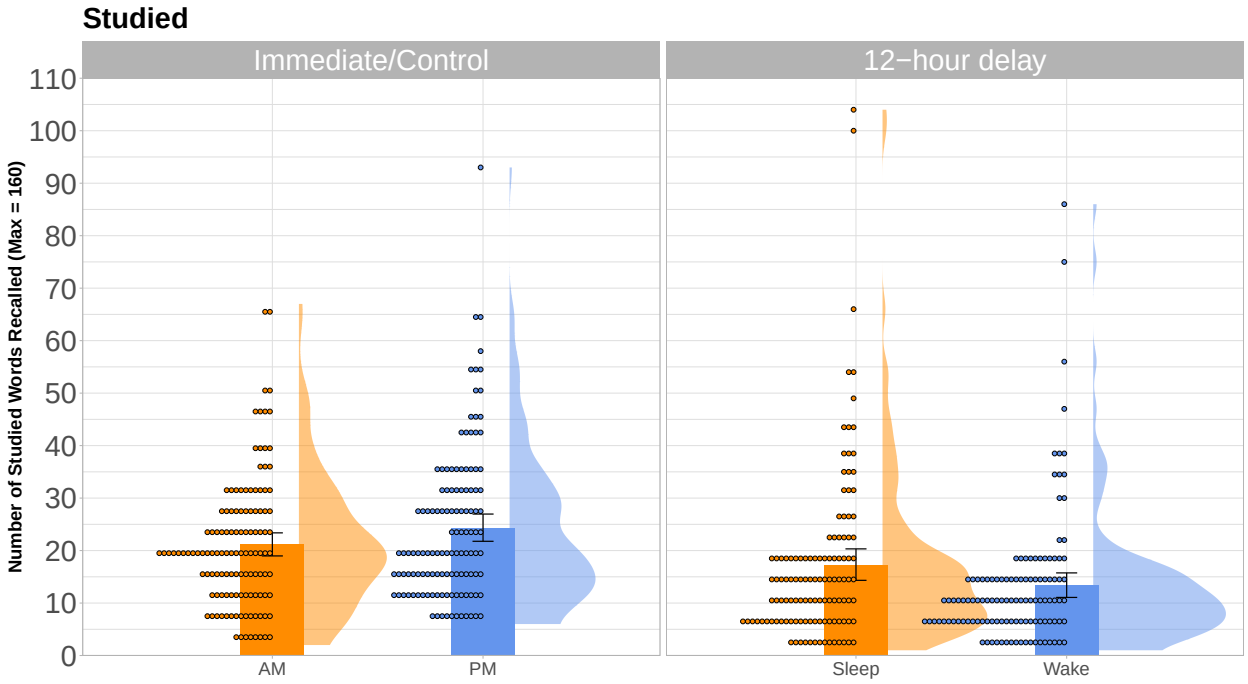

Figure 4: Number of studied list words correctly recalled, summarised across the four groups. Each dot represents an individual participant, and the error bars represent 95% confidence intervals.

We fitted a GLMM to the veridical recall dataset (N of observations = 488 participants × 160 studied words). The dependent variable was whether a studied word was recalled or not. The fixed effects were the the number of intrusions a participant produced, Interval (Immediate vs. Delay), Test Time (AM vs. PM), and an Interval by Test Time interaction. The coding scheme and computation procedure were the same as in the previous analysis. The model output and its random-effect structure are available in the lower half of Table 4. There were no significant effects of intrusions (z = -1.372, p = .17) or Test Time (z = 0.691, p = .489). However, there was a main effect of Interval (z = 7.721, p <.001) such that participants in the Delay groups recalled significantly fewer studied words (M = 15.37, SD = 14.97) than those in the Immediate groups (M = 22.77, SD = 13.59), indicating time-dependent memory decay. Importantly, there was a significant Interval by Test Time interaction (z = -3.325, p <.001*), so we broke it down with the emmeans package as pre-registered. Within the Immediate groups, the evening participants (M = 24.36, SD = 14.6) recalled more studied words than the morning participants (M = 21.18, SD = 12.36), although this was not statistically significant (β = -0.196, SE = 0.104, z = -1.88, p = .06). Within the Delay groups, there was a main effect of Test Time (β = 0.299, SE = 0.107, z = 2.797, p = .0052), such that the Sleep participants (M = 17.32,

SD = 16.59) outperformed their Wake counterparts (M = 13.41, SD = 12.93). Together, these support the well-established findings that sleep is typically beneficial to the retention of newly encoded declarative memories.

12.4.1 Complementary Bayesian analysis

Although our inference was based on a frequentist approach, we pre-registered to use a Bayesian analysis to complement and test the strength of our results.

Bayes Factors were computed for (1) the Interval × Test Time interaction in the false and veridical mixed-effect models above, and for the simple effects of Test Time within the (2) Immediate and (3) Delay groups. Following the procedures in Gilbert et al. [32], a Bayes Factor was computed using the Bayesian Information Criterion (BIC) approximation from two competing GLMMs. For instance, in computing the Bayes Factor for the Interval × Test Time interaction, two models were needed: An alternative model containing the full fixed-effects structure (Intrusions + Interval + Test Time + Interval × Test Time), and a null model lacking the interaction.[5] To estimate the Bayes Factor, we used the formula e $e\Delta BIC_{10}/2$, where ΔBIC_{10} is the BIC for the null model minus the BIC for the alternative model [33–35]. This produces a Bayes Factor_{10}, which was interpreted with reference to Lee and Wagenmakers' [36] heuristics. The current BIC approximation method has the advantage of being a straightforward solution for mixed-effects models; however, its usage remains controversial as it is known to favour the simpler model (i.e., the null hypothesis; [34,37,38]). Table 5 summarises the Bayes Factors derived from our mixed-effects models.

Table 5: Bayes Factors for the Interval x Test Time interactions and the simple effects of Test Time in the lure and veridical recall data.

Effects	Bayes Factor 10
False (lure) recall	
Interval x Test Time	2e-05
Test Time in Immediate groups	0.5
Test Time in Delay groups	0.5
Veridical (studied word) recall	
Interval x Test Time	29122042
Test Time in Immediate groups	0.00127
Test Time in Delay groups	0.01763

Surprisingly, all the Bayes Factors, except for the Interval × Test Time interaction in the studied word model, were below 0.1. These, according to Lee and Wagenmakers [36], can be taken as extreme evidence for the null hypotheses. In other words, there is a discrepancy between our frequentist and Bayesian analyses. We stress that this Bayesian analysis is complementary in nature and our primary test of significance remains the frequentist test.

13 Results of Exploratory Analyses

In this section, we present the results of four exploratory analyses, which explored [1] the number of total responses (i.e., studied + lure + intrusions) across groups, [2] whether the effect of sleep on lure recall is modulated by veridical recall, as suggested by Diekelmann et al. [39], [3] the extent to which a lure being produced is predicted by its corresponding list items being recalled, and [4] the semantic distance between intrusions and critical lures.

[5]To obtain the Bayes Factor for the simple effects of Test Time, the alternative model will contain Test Time as the sole fixed effect while the null model will contain no fixed effects.

13.1 Number of total responses

In light of the finding that participants who completed free recall in the evening (vs. morning) produced more intrusions, we asked whether this was driven by these participants having a greater tendency to put down more responses generally. To test this, we calculated the number of total responses by each participant (i.e., studied + lures + intrusions), which is summarised across groups in Figure 5.

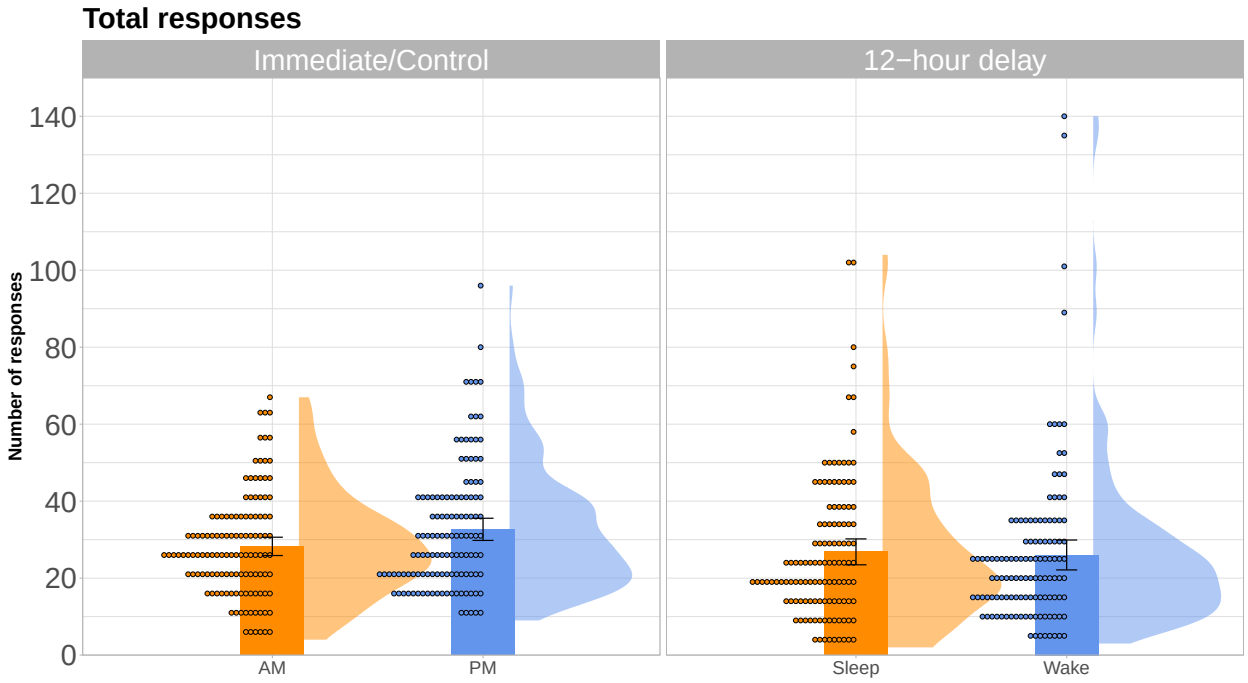

Figure 5: Number of total responses (i.e., studied + lure + intrusions), summarised across the four groups. Each dot represents an individual participant, and the error bars represent 95% confidence intervals.

Unlike the intrusion data which had an overall mean of 6.8 and a minimum of 0, the number of total responses had a mean of 28.5 and a minimum of 2, suggesting that it is better to consider total responses as continuous, as opposed to count, data. As such, we used a 2 × 2 between-participant ANOVA to test for the effects of Interval and Test Time on the number of total responses, which was log-transformed to give a more normal distribution. The ANOVA revealed a main effect of Interval [$F(1, 484) = 21.35$, $p < .001$], such that participants in the Immediate groups (M = NaN, SD = NA) gave more responses than those in the Delay groups (M = NaN, SD = NA). This pattern is consistent with the confirmatory analysis on the studied list words. However, importantly, there was no significant effect of Test Time [$F(1, 484) = 1.68$, $p = .196$], and the Interval by Test Time interaction was also non-significant [$F(1, 484) = 2.39$, $p = .123$]. Together with the intrusion data, this exploratory analysis suggests that attempting free recall in the evening led to a selective increase in intrusions but not necessarily an increase in general response bias.

13.2 Is the effect of sleep on false recall modulated by veridical memory?

In 36 participants (18 Sleep and 18 Wake), Diekelmann et al. [39] found that the effect of sleep on DRM false recall was modulated by veridical recall. In their analysis, they performed a post-hoc median split on adjusted veridical recall (i.e., correct recall minus intrusions), separating their participants as either high or low performers. They reported that participants in the Sleep (vs. Wake) group falsely recalled more lures, but

only if they were low performers. This finding, as far as we can see, has yet to be replicated, so we explored whether it could be observed in our data, which is more well-suited to test for individual differences given our large sample size.

In this exploratory analysis, we followed Diekelmann et al. [39] by focusing on the Delay groups[6] and by performing a median split on our sample's adjusted veridical recall (Median = 5), classifying our participants as either high (>5; N = 117) or low ($\leq$ 5; N = 123) performers. We followed the analysis approach of Diekelmann et al. [39], but since the number of critical lures a participant produced is count data, we used a 2 (Group: Sleep vs. Wake) $\times$ 2 (Adjusted veridical recall: High vs. Low Performers) Poisson regression, instead of an ANOVA. This analysis showed no main effect of Group (β = 0.052, z = 0.494, p = .621) but a main effect of Adjusted veridical recall (β = -0.438, z = -3.766, p < .001), such that high performers (M = 3.17, SD = 2.71) tended to falsely recall more critical lures than low performers (M = 2.09, SD = 1.68) ([40,41], but see [24,42,43]). Furthermore, importantly, the interaction between Group and Adjusted veridical recall was not significant (β = 0.049, z = 0.30, p = .763). Despite this, we followed Diekelmann et al. [39] by comparing low performers in the Sleep and Wake groups. Contrary to their findings, emmeans showed that our low performers in both groups recalled essentially the same number of critical lures (M_{Sleep} = 2.09, SD_{Sleep} = 1.63 vs. M_{Wake} = 2.09, SD_{Wake} = 1.74; β = 0.133, z = 0.021, p = .982). In other words, this exploratory analysis found no evidence that the effect of sleep on false recall was modulated by veridical memory, casting doubt over the robustness and reliability of Diekelmann et al.'s [39] findings.

13.3 The relationship between lure and veridical recall on a list level

Here, we asked whether recall probability of a lure (e.g., doctor) is predicted by the number of corresponding list items being recalled (e.g., nurse, hospital, sick), and if it does, whether it differs between the Sleep and Wake groups. These questions help shed light on the degree to which sleep increases lure recall via processes such as retrieval-induced generalisation[7] or gist abstraction[8], as these processes may predict a different degree of interdependence between lure and veridical recall. If sleep (vs. wake) promoted retrieval-induced generalisation, lure and veridical recall should become more strongly correlated with each other after sleep, because better veridical recall for a set of studied words may generalise to the corresponding critical lure (or vice versa). On the other hand, if sleep (vs. wake) promoted gist abstraction, lure recall may become less related to memories for the corresponding list items, because theories of gist abstraction such as iOtA may predict that sleep would selectively boost the overlapping gist memory (i.e., the lure) but not necessarily the specific studied words.

In this exploratory analysis, we first calculated a participant's number of correct recalls per DRM wordlist (Range = 0 - 8) and used this to predict recall of the corresponding critical lure in a generalised mixed-effect model, which had Number of intrusions, Number of correct recall per list, Interval (Immediate vs. Delay), Test Time (AM vs. PM), and an interaction of the latter three as the fixed effects. Interval and Test Time were effect coded, and the random-effect structures contained a by-participant intercept only, as prescribed by the buildmer package. The model output is summarised in Table 6.

[6]Note that Diekelmann et al. (2010) did not have Immediate/Control groups.
[7]Retrieval-induced generalisation: Retrieval of one word cueing retrieval of a related word (e.g., Berens & Bird, 2017)
[8]Gist abstraction: Extraction of the central or essential meaning of learned information

Table 6: Outputs from the exploratory generalised mixed-effect model examining the effects of intrusions, correct recall per list, Interval (Immediate vs. Delay), and Test Time (Sleep vs. Wake) in false recall.

Fixed effects	**Estimate**	**SE**	**z**	**p**
Intercept	-3.543	0.092	-38.6	<.001
Intrusions	0.034	0.006	6.011	<.001
Correct recall/list	0.799	0.025	31.848	<.001
Test Time	0.092	0.07	1.319	.187
Interval	0.039	0.07	0.551	.582
Correct recall/list x Test Time	-0.062	0.022	-2.778	.00548
Correct recall/list x Interval	-0.127	0.022	-5.702	<.001
Interval x Test Time	-0.095	0.069	-1.372	.17
Correct recall/list x Interval x Test Time	0.062	0.022	2.798	.00514

Given the significant three-way interaction, we broke it down by computing two additional GLMMs, one within the Immediate and another within the Delay group. These models had Number of intrusions, Number of correct recall per list, Test Time (AM vs. PM), and an interaction of the latter two as the fixed effects. Table 7 summarises the model outputs.

Table 7: Outputs from the exploratory generalised mixed effect models examining the effects of Intrusions, Correct recall per list, and Test Time (Sleep vs. Wake) in false recall.

	Immediate				**Delay**			
Fixed effects	**Estimate**	**SE**	**z**	**p**	**Estimate**	**SE**	**z**	**p**
Intercept	-3.681	0.153	-24.109	<.001*	-3.544	0.136	-26.104	<.001*
Intrusions	0.07	0.01	6.854	<.001*	0.023	0.007	3.243	.001*
Correct recall/list	0.656	0.03	22.033	<.001*	0.95	0.041	23.121	<.001*
Test Time	0.022	0.09	0.245	.807	0.163	0.104	1.571	.116
Correct recall/list x Test Time	-0.001	0.026	-0.037	.971	-0.125	0.036	-3.471	<.001*

Across the Immediate and Delay groups, the number of intrusions and correct recall per list both had a main effect (ps $\leq$ .001) such that they were positively correlated with lure recall. However, the effect of Test Time was not significant in either group (ps $>$ .11). Finally, for the Correct recall per list $\times$ Test Time interaction, it was significant in the Delay (z = -3.471, p <.001) but not the Immediate groups (z = -0.037, p = .971). To interpret the former, we used the R package, effects [44], to visualise it (see Figure 6A) and plotted lure recall probability against veridical recall on a participant level (see Figure 6B).

In line with the previous exploratory analysis on adjusted veridical recall, we found that when a participant recalled more studied items from a DRM wordlist, they were also more likely to recall the corresponding critical lure, suggesting some kind of retrieval-induced generalisation; importantly, however, this effect was weaker in the Sleep than the Wake participants. This finding is striking, especially in light of our confirmatory findings that overall, the Sleep participants produced more critical lures (with intrusions controlled for) and more studied list items than the Wake participants. What this exploratory analysis suggests is that after sleep (vs. wakefulness), whether a lure was recalled may be less reliant on the retrieval of its corresponding studied items, potentially hinting that sleep may affect DRM false memory primarily via some kinds of gist abstraction process.

13.4 Semantic distance between intrusions and critical lures

As per the DRM literature, responses that were neither the studied list words nor the critical lures were classified as (non-critical) intrusions. For instance, our participants studied nurse, sick, lawyer, medicine,

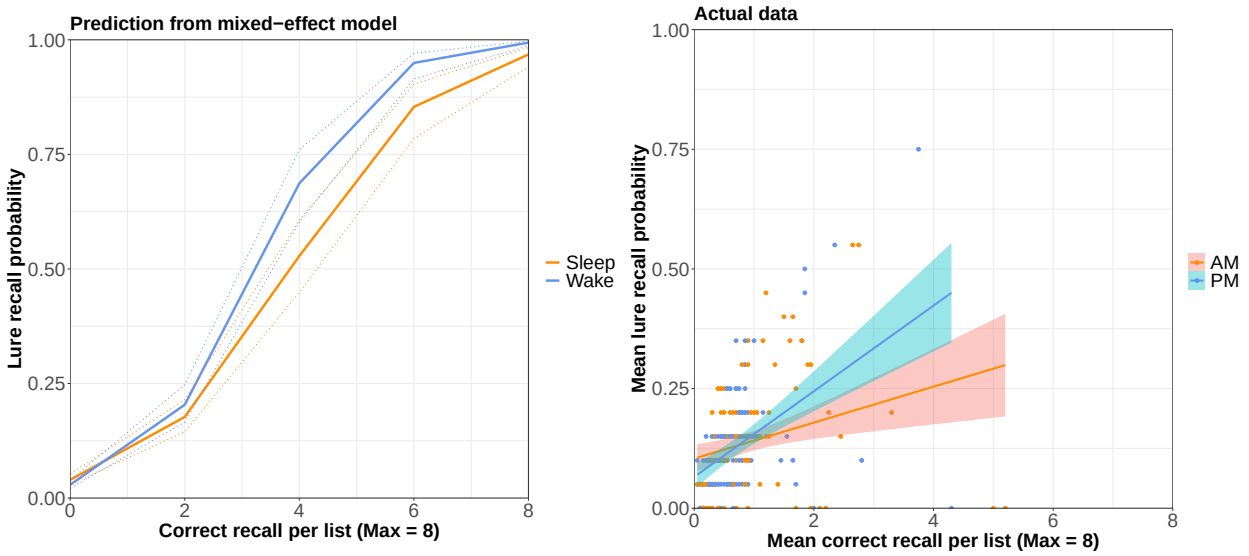

Figure 6: (A) Prediction from generalised mixed-effect model on the combined effects of correct recall/list and group (sleep vs. wake) on false recall (B) Correlation between mean lure and studied word recall in the Sleep and Wake groups. Each dot represents an individual participant. Dotted lines/shaded areas represent 95% confidence intervals.

health, hospital, dentist, and physician, with doctor as the critical lure. Responses such as clinic and coconut would both be considered intrusions, but clearly, clinic is semantically more related to the studied list words than coconut. In other words, there is much diversity within the intrusion data. Here, we explored whether our four groups differed in terms of the semantic distance between intrusions and critical lures, as indexed by pre-trained semantic spaces (ukWaC; [45]) derived from word2vec [46]. We reasoned that since lure recall was greater in our Sleep (vs. Wake) participants, the intrusions produced by these participants could be more related to the lures in semantic space (e.g., [15]). To test this, we computed the cosine similarities between each intrusion and each of the 20 critical lures (see Table 7 for an illustration). The intrusion-lure pair with the highest cosine similarity (i.e., the nearest neighbour) was used for this analysis.

Since the number of intrusions produced varies greatly between participants, we averaged the lure-intrusion cosine on a participant level and used this as the dependent variable. A 2 (Interval) × 2 (Test Time) between-participant ANOVA revealed no effects of Interval ($z = 2.541$, $p = .112$) or Test Time ($z = 0.085$, $p = .771$), and their interaction was also non-significant ($z = 0.886$, $p = .347$). We explored further by comparing the cosine in the Sleep and Wake groups. While this comparison is in the predicted direction ($M_{Sleep} = 0.381$, $SD_{Sleep} = 0.072$ vs. $M_{Wake} = 0.37$, $SD_{Wake} = 0.085$), it was not statistically significant according to an emmeans pairwise comparison ($z = 0.874$, $p = .383$). We interpret these null findings as suggesting that the effect of sleep in the DRM paradigm may be fairly restricted to the lures/studied list words.

13.5 Results summary

In light of the extensive dataset and the comprehensive analyses conducted, we decided to summarise our key findings in Table 8 to help readers gain a better understanding of the overall picture.

14 Time-of-day effects on intrusions

Rather unexpectedly, and contrary to the null findings from prior 'Sleep × DRM' studies (e.g., [7,19]), participants who completed free recall in the evening (i.e., the PM-control and Wake groups) produced more

intrusions than those in the morning (i.e., the AM-control and Sleep groups). Notably, Test Time did not have a significant effect on the number of total responses, suggesting that completing free recall in the evening (vs. morning) led to a selective increase in intrusions but not a global output bias. Also worth noting is that our four groups were well-matched on their circadian preference (see rMEQ scores in Table 3), suggesting that the effect of Test Time on intrusions is unlikely to be attributable to free recall being completed at optimal or non-optimal times of day, which are known to affect performance on some cognitive tasks (e.g., [47–49]). As to why evening (vs. morning) testing led to a selective increase in intrusions, we propose that it might be due to accumulation of information throughout the day.

Participants tested in the evening would have engaged in various daytime activities in the preceding 10-12 hours, while those tested in the morning would have been sleeping for the majority of those hours. In other words, participants tested in the evening would have accumulated a large amount of sensory and linguistic information, which might interfere or even compete with memories for the studied list words at retrieval, increasing the likelihood of intrusions being produced. In contrast, for participants in the AM-control and Sleep groups, not only would they have less accumulated sensory input from the preceding hours, but they may also benefit from one of the proposed functions of sleep, which is to "reset" the brain by pruning (relatively unimportant) information accumulated prior to sleep [50,51]. As such, it seems reasonable to infer that participants tested in the evening (vs. morning) may have experienced more interference from information accumulated throughout the course of the day, which may have, in turn, led to an increase in intrusions at retrieval. This interpretation also fits with the finding that Test Time had a greater effect in the Delay (Wake > Sleep) than in the Immediate groups (PM > AM). Participants in the Wake groups attempted recall after 12 hours of daytime wakefulness, so interference from accumulated sensory input is likely to impede not only retrieval, but also memory storage throughout the 12-hour interval. In contrast, participants in the PM-control group attempted recall shortly after study, so interference from accumulated information is likely to be mostly restricted to memory retrieval. In other words, interference from accumulated sensory input is expected to be greater in the Wake than in the PM-control group, which was in fact the case.

Finally, having considered the potential reason why evening testing may increase intrusions, we turn to why no previous 'Sleep × DRM' studies had reported the same. We argue that this is because prior studies were underpowered. In our intrusion data, the effect size of Test Time was small: An exploratory comparison of the AM vs. PM groups using a Mann-Whitney U test revealed an effect size of Pearson's r = -0.096 (which roughly corresponds to Cohen's d = 0.193), while the same test on the Sleep vs. Wake groups revealed an effect size of r = 0.175 (d = 0.355). In order to detect the latter at 80% power (assuming alpha = 0.05) in a two-tailed between-participant Mann-Whitney U-test, a total sample size of 264 participants is required (G*Power; [52]), which is substantially greater than the sample sizes of the vast majority of existing memory/sleep studies (e.g., [7,53]). As such, it is not surprising that few prior sleep studies had reported time-of-day effects in declarative memories.

15 Some reflections

Mixed evidence in the existing literature. Our well-powered registered report had a total sample size of 488, with 120 in each of the Sleep and Wake groups. This far surpasses the sample sizes of prior 'Sleep × DRM' studies, where the median N was 27.6 per group. In our data, the effect of sleep on lure recall was only significant when controlling for differences in intrusion rates, and we checked the size of this sleep effect in an exploratory analysis, putting the estimate at Cohen's d = 0.274.[9] This estimate is substantially lower than that from Newbury and Monaghan's [1] meta-analysis, which reported to be Cohen's d = +0.92 (95% CI: 0.54, 1.30). The larger sample size and well-powered nature of our registered report contribute to a more precise estimation of the effect size, highlighting the need for cautious interpretation of prior evidence and the possibility that prior effect sizes may have been inflated due to, for example, publication bias and small sample sizes. Furthermore, given sleep had such a small effect on DRM false memory, it is perhaps not surprising that previous studies had produced mixed evidence. In light of these, we echo the view of a recent article [54] that future sleep research must prioritise robust methodologies (e.g., registered report

[9]We used the eff_size function in the emmeans package to estimate the effect size of Test Time (sleep vs. wake) in our confirmatory GLM model, with sigma = sigma(lure_model) and edf = infinite.

and pre-registration) and larger sample sizes to enhance the reliability and generalisability of sleep-related memory effects [see also 55].

References

1. Newbury CR, Monaghan P. 2019 When does sleep affect veridical and false memory consolidation? A meta-analysis. *Psychonomic Bulletin and Review* **26**, 387–400. (doi:10.3758/s13423-018-1528-4)
2. Berres S, Erdfelder E. 2021 The sleep benefit in episodic memory: An integrative review and a meta-analysis. *Psychol Bull* **147**, 1309–1353. (doi:10.1037/bul0000350)
3. Diekelmann S, Wilhelm I, Born J. 2009 The whats and whens of sleep-dependent memory consolidation. *Sleep medicine reviews* **13**, 309–321. (doi:10.1016/j.smrv.2008.08.002)
4. Lipinska G, Stuart B, Thomas KGF, Baldwin DS, Bolinger E. 2019 Preferential consolidation of emotional memory during sleep: A meta-analysis. *Front. Psychol* **10**. (doi:10.3389/fpsyg.2019.01014)
5. Fenn KM, Gallo DA, Margoliash D. 2009 Reduced false memory after sleep. *Learning & Memory* **16**, 509–513. (doi:10.1101/lm.1500808)
6. Monaghan P, Shaw JJ, Ashworth-Lord A, Newbury CR. 2017 Hemispheric processing of memory is affected by sleep. *Brain and Language* **167**, 36–43. (doi:10.1016/j.bandl.2016.05.003)
7. Payne JD, Schacter DL, Propper RE, Huang L, Walmsley EJ, Tucker MA, Walker MP, Stickgold R. 2009 The role of sleep in false memory formation. *Neurobiology of Learning and Memory* **92**, 327–334. (doi:10.1016/j.nlm.2009.03.007)
8. Rodd J. 2019 How to maintain data quality when you can't see your participants.
9. Curtis AJ, Mak MHC, Chen S, Rodd JM, Gaskell MG. 2022 Word-meaning priming extends beyond homonyms. *Cognition* **226**, 105175. (doi:10.1016/j.cognition.2022.105175)
10. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. 2020 Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* **52**, 388–407. (doi:10.3758/s13428-019-01237-x)
11. Barnhoorn JS, Haasnoot E, Bocanegra BR, Steenergen H. 2015 QRTEngine: An easy solution for running online reaction time experiments using qualtrics. *Behavior Research Methods* **47**, 918–929. (doi:10.3758/s13428-014-0530-7)
12. Mak MHC, Twitchell H. 2020 Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin and Review* **27**, 1059–1069. (doi:10.3758/s13423-020-01773-0)
13. Mak MHC, Hsiao Y, Nation K. 2021 Lexical connectivity effects in immediate serial recall of words. *Journal of Experimental Psychology: Learning, Memory and Cognition* **47**, 1971–1997. (doi:10.1037/xlm0001089)
14. Ashton JE, Cairney SA. 2021 Future-relevant memories are not selectively strengthened during sleep. *PLoS ONE* **16**, 0258110. (doi:10.1371/journal.pone.0258110)
15. Mak MHC, Curtis AJ, Rodd JM, Gaskell MG. 2023 Recall and recognition of discourse memory across sleep and wake. (doi:10.31234/osf.io/6vqh9)
16. Lo JC, Dijk DJ, Groeger JA. 2014 Comparing the effects of nocturnal sleep and daytime napping on declarative memory consolidation. *PLoS ONE* **9**. (doi:10.1371/journal.pone.0108100)
17. Plihal W, Born J. 1997 Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience* **9**, 534–547. (doi:10.1162/jocn.1997.9.4.534)
18. Mak MHC, Curtis AJ, Rodd JM, Gaskell MG. 2023 Episodic memory and sleep are involved in the maintenance of context-specific lexical information. *Journal of Experimental Psychology: General* (doi:10.1037/xge0001435)
19. McKeon S, Pace-Schott EF, Spencer RMC. 2012 Interaction of sleep and emotional content on the production of false memories. *PLoS ONE* **7**, 1–7. (doi:10.1371/journal.pone.0049353)

20. Shaw JJ, Monaghan P. 2017 Lateralised sleep spindles relate to false memory generation. *Neuropsychologia* **107**, 60–67. (doi:10.1016/j.neuropsychologia.2017.11.002)

21. Swannell ER, Dewhurst SA. 2013 Effects of presentation format and list length on children's false memories. *Journal of cognition and development* **14**, 332–342. (doi:10.1080/15248372.2011.638689)

22. 1997 Associative processes in false recall and false recognition. *Psychological Science* **8**, 231–237. (doi:10.1111/j.1467-9280.1997.tb00417.x)

23. Alakbarova D, Hicks JL, Ball BH. 2021 The influence of semantic context on false memories. *Memory & Cognition* **49**, 1555–1567. (doi:10.3758/s13421-021-01182-1)

24. Roediger HL, Watson JM, McDermott KB, Gallo DA. 2001 Factors that determine false recall: A multiple regression analysis. *Psychonomic bulletin & review* **8**, 385–407. (doi:10.3758/bf03196177)

25. Hoddes E, Dement WC, Zarcone V. 1973 The development and use of the stanford sleepiness scale (SSS. *Psychophysiology* **10**, 421–436.

26. Adan A, Almirall H. 1991 Horne & ostberg morningness-eveningness questionnaire: A reduced scale. *Personality and Individual Differences* **12**, 241–253. (doi:10.1016/0191-8869(91)90110-W)

27. Thomas KA, Clifford S. 2017 Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* **77**, 184–197. (doi:10.1016/j.chb.2017.08.038)

28. Lenth RV. 2021 Emmeans: Estimated marginal means, aka least-squares means.

29. Barr D. 2019 Coding categorical predictor variables in factorial designs.

30. Voeten C. 2021 Buildmer: Stepwise elimination and term reordering for mixed-effects regression.

31. Barr DJ, Levy R, Scheepers C, Tily HJ. 2013 Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278. (doi:10.1016/j.jml.2012.11.001)

32. Gilbert RA, Davis MH, Gaskell MG, Rodd JM. 2018 Listeners and readers generalize their experience with word meanings across modalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**, 1533–1561. (doi:10.1037/xlm0000532)

33. Masson MEJ. 2011 A tutorial on a practical bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods* **43**, 679–690. (doi:10.3758/s13428-010-0049-5)

34. Lindeløv JK. 2018 How to compute bayes factors using lm, lmer, BayesFactor, brms, and JAGS/stan/pymc3. (doi:https://rpubs.com/lindeloev/358672)

35. Wagenmakers E-J. 2007 A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* **14**, 779–804.

36. Lee MD, Wagenmakers EJ. 2014 *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

37. Vandekerckhove J, Matzke D, Wagenmakers E-J. 2014 Model comparison and the principle of parsimony. In *Oxford handbook ofcomputational and mathematical psychology*, Oxford, UK: Oxford University Press.

38. Weakliem DL. 1999 A critique of the bayesian information criterion for model selection. *Sociological Methods & Research* **27**, 359–397. (doi:10.1177/0049124199027003002)

39. Diekelmann S, Born J, Wagner U. 2010 Sleep enhances false memories depending on general memory performance. *Behavioural Brain Research* **208**, 425–429. (doi:10.1016/j.bbr.2009.12.021)

40. Thapar A, McDermott KB. 2001 False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition* **29**, 424–432. (doi:10.3758/BF03196393)

41. Toglia MP. 1999 Recall accuracy and illusory memories: When more is less. *Memory* **7**, 233–256. (doi:10.1080/741944069)

42. Cann DR, Mcrae K, Katz AN. 2011 False recall in the deese-roediger-McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology* **64**, 1515–1542. (doi:10.1080/17470218.2011.560272)

43. Stadler MA, Roediger HL, McDermott KB. 1999 Norms for word lists that create false memories. *Memory & Cognition* **27**, 494–500. (doi:10.3758/bf03211543)

44. Fox J, Weisberg S. 2019 An r companion to applied regression.

45. Baroni M, Bernardini S, Ferraresi A, Zanchetta E. 2009 The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* **43**, 209–226.

46. Günther F, Dudschig C, Kaup B. 2015 LSAfun - an r package for computations based on latent semantic analysis. *Behavior Research Methods* **47**, 930–944. (doi:10.3758/s13428-014-0529-0)

47. Hasher L, Chung C, May CP, Foong N. 2002 Age, time of testing, and proactive interference. *Canadian Journal of Experimental Psychology* **56**, 200–207. (doi:10.1037/h0087397)

48. Krishnan HC, Lyons LC. 2015 Synchrony and desynchrony in circadian clocks: Impacts on learning and memory. *Learning and Memory* **22**, 426–437. (doi:10.1101/lm.038877.115)

49. May CP, Hasher L, Foong N. 2005 Implicit memory, age, and time of day: Paradoxical priming effects. *Psychological Science* **16**, 96–100. (doi:10.1111/j.0956-7976.2005.00788.x)

50. Tononi G, Cirelli C. 2006 Sleep function and synaptic homeostasis. *Sleep Medicine Reviews* **10**, 49–62. (doi:10.1016/j.smrv.2005.05.002)

51. Tononi G, Cirelli C. 2014 Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* **81**, 12–34. (doi:10.1016/j.neuron.2013.12.025)

52. Faul F, Erdfelder E, Buchner A, Lang AG. 2009 Statistical power analyses using GPower 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods* **41**, 1149–1160. (doi:10.3758/BRM.41.4.1149)

53. Yaremenko S, Sauerland M, Hope L. 2021 Circadian rhythm and memory performance: No time-of-day effect on face recognition. *Collabra: Psychology* **7**, 1–13. (doi:10.1525/collabra.21939)

54. Cordi MJ, Rasch B. 2021 How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology* **67**, 1–7. (doi:10.1016/j.conb.2020.06.002)

55. Nemeth D *et al.* 2019 Pitfalls in sleep and memory research and how to avoid them: A consensus paper. (doi:10.20944/preprints201908.0208.v2)

56. Herbison P. In press. Analysing count data.

57. Lo S, Andrews S. 2015 To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology* **6**, 1171. (doi:10.3389/fpsyg.2015.01171)

58. Brysbaert M, Stevens M. 2018 Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition* **1**, 1–20. (doi:10.5334/joc.10)

59. Schäfer T, Schwarz MA. 2019 The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology* **10**, 1–13. (doi:10.3389/fpsyg.2019.00813)

60. Albers C, Lakens D. 2018 When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of experimental social psychology* **74**, 187–195. (doi:10.1016/j.jesp.2017.09.004)

61. Cortex. 2013 Guidelines for authors.

62. Kumle L, Võ MLH, Draschkow D. 2021 Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior research methods* **53**, 2528–2543. (doi:10.3758/s13428-021-01546-0)

63. Bates D, Kliegl R, Vasishth S, Baayen H. 2015 Parsimonious mixed models.

64. Green P, Macleod CJ. 2016 SIMR: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* **7**, 493–498. (doi:10.1111/2041-210X.12504)

Appendix A: Screening Survey

Appendix B: Survey in the test phase

Appendix C: Justifications for using GLMM for false recall (as opposed to t-test/ANOVA)

The number of critical lures falsely recalled by a participant is count data, ranging from 0 to anything from 8 to 20 (depending on how many DRM lists were shown). In Payne et al. [7] for instance, the mean number of lure recalls was ~3.25 (out of 8). Count data with a low mean almost never approximates a normal distribution, because it is truncated at 0 (i.e., negative scores are impossible) and is skewed to the right [56]. Parametric tests like t-test and ANOVA assume a normal distribution, so they are unlikely to be suitable for false recall data. GLMM, on the other hand, does not assume normal distribution [57]. In addition, GLMM has numerous advantages over a t-test or an ANOVA; for instance, it can take by-participant and by-item variance into account, giving researchers the ability to test whether the effect of an independent variable generalises across participants and items [e.g., 58].

Appendix D: Power analysis

'According to Newbury and Monaghan's [1] meta-analysis, the effect size for sleep in DRM false memory is Hedge's g = +0.92 (95% CI: 0.54, 1.30; p < .001) in short lists (10 words per list).[10] However, due to certain biases in the psychology literature (e.g., publication bias), it has been argued that published effect sizes are generally inflated [e.g., 59] and that power analysis should be based on the lowest meaningful estimate [60,61]. Therefore, we opted for a more conservative (yet contextualised) effect size estimate and went for the lower-bound of the 95% confidence interval reported by Newbury and Monaghan [1], which is g = +0.54.

On estimating power in GLMM, a recent guideline [62] recommends using well-powered data from previous experiments. However, as far as we are aware, no prior 'Sleep × DRM' studies have made their data publicly available. We, therefore, simulated a dataset for our power calculation (available on OSF).

The first step is to determine a sample size. We have the financial resources to reach up to 160 participants/group (i.e., 640 in total), so we began by fabricating a dataset containing 120 participants/group. We made up 20 false recall observations for each participant. We then split the dataset by Interval, so one dataset for the Delay (Sleep + Wake) groups, another for the Immediate (AM + PM-control) groups, with each containing 4800 observations from 240 participants. In the first dataset, we simulated the false recall data for the Delay groups such that they approximated the data distribution [M_{Sleep} = 45.9% (SD = 20.6%) vs M_{Wake} = 36.3% (SD = 21.2%), p = .005] from a prior study (Payne et al., [7]; Experiment 1). Then, the data were manipulated to fit with our effect size assumption, such that the effect size for sleep is d = +0.54 [M_{Sleep} = 44.7% (SD = 12.6%) vs. M_{Wake} = 38.0% (SD = 12.0%); t(237.4) = 4.2, p = <.001]. Afterwards, we simulated the false recall data in the second dataset for the AM and PM-control groups such that they also approximated the data distribution in Payne et al. [7] (Experiment 1; MAM = 42.5% (SD = 19.6%) vs. MPM = 46.3% (SD = 23.5%), p = .57) and that they did not differ significantly from each other [MAM = 42.9% (SD = 13.0%) vs. MPM= 43.5% (SD = 14%); t(233.3) = -0.37, p = .709, d = -0.05].[11]

[10] Hedge's g and Cohen's d are interchangeable when the sample size is larger than 30 (Kline, 2004; Lakens, 2013).

[11] Notably, the standard deviations (SDs) from Payne et al. (2009; Experiment 1) are larger than those in our fabricated datasets. This is because Payne et al. showed only 8 wordlists (i.e., maximum lure recall = 8) while ours will show 20 (i.e., maximum lure recall = 20). To explain why this matters, a more concrete example here is useful: In Payne et al, a participant falsely recalling 3 lures would have a false recall rate of 37.5% while another recalling 4 lures would have a rate of 50%. So there is a 12.5% difference between each successive number. In our fabricated data, recalling 3 lures has a false recall rate of 15% while 4 lures has a rate of 20%, so there is a 5% difference. Therefore, understandably, the SDs in our fabricated datasets are

We then merged the two fabricated datasets together and fitted a GLMM to it, using the lme4 package [63]. The dependent variable, fixed effects structure, coding scheme, and computation procedures were identical to those described in section 12.3. Table D1 shows the fixed-effects estimates from the converged model, which has a by-participant intercept only.

Table D1

Based on the fixed-effects estimates, we estimated the power a sample size of 480 (i.e., 120/group) has for detecting an Interval and Test Time interaction. We conducted Monte Carlo simulation using the "simr" package [64] in R (see Box 2 for R codes). After 500 simulations, it was estimated that a sample of this size gives 90% power (95% CI: 87.03, 92.49) to detect a significant interaction. Then, we estimated the power we have for detecting a simple effect of Test Time within the Delay groups (i.e., Sleep vs. Wake). Following the simulation procedures above, it was estimated that 120 participants/group (i.e., 240 in the Delay groups) will give about 90% power (95% CI: 97.41, 99.56). In sum, our power calculation showed that having 120 participants/groups will give ample power (>90%) to detect both an Interval x Test Time interaction and a simple effect of Sleep vs. Wake. Finally, we also estimated that we will have at least 80% power as long as we have >99 participants/group. Therefore, in case we fail to reach our target sample size of 120 participants/group before funding expires but manage to recruit >99/group, our proposed experiment will still have satisfactory power.

We note that the focus of our proposed experiment is Research Question #1 [Does sleep (vs. wakefulness) influence DRM false recall?], so we based our power analysis on this question. Despite this, the estimate of 120 participants/group will also give over 90% power to detect the desired effects for Research Question #2 [Does sleep (vs. wake) increase veridical recall?], assuming the effect size for sleep is similar between Questions #1 and #2. Furthermore, since there are more studied list words than critical lures (160 vs. 20), the GLMM for addressing Question #2 will have substantially more observations than that for Question #1 (~76800 vs. ~9600), boosting power on the item level. In short, our target sample size of 120 participants/group will give us sufficient power for both Research Questions.

Box 2

R codes for Monte Carlo simulation

```r
library(simr)
fixef(fabricated_model)["Interval1:Test_Time1"] <- 0.076
set.seed(99)
powerSim(fabricated_model, fixed("Interval1:Test_Time1"), nsim=500)
```

```
## Power for predictor 'Interval1:Test_Time1', (95% confidence interval):
##       90.00% (87.03, 92.49)
##
## Test: z-test
##       Effect size for Interval1:Test_Time1 is 0.076
##
## Based on 500 simulations, (0 warnings, 0 errors)
## alpha = 0.05, nrow = 9600
##
## Time elapsed: 0 h 3 m 34 s
```

necessarily lower than theirs (by roughly a half).