Stage 1 Registered Report

Estimating the Effect of Reward on Sleep-Dependent Memory Consolidation – A Registered Report

David P. Morgan[1, 2, 3*], Juliane Nagel[1, 2, 3], N. Cagatay Gürsoy[1, 2, 3], Simon Kern[1, 2, 3], & Gordon B. Feld[1, 2, 3, 4*]

[1] Department of Clinical Psychology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2] Department of Addiction Behavior and Addiction Medicine, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[3] Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[4] Department of Psychology, Heidelberg University, Heidelberg, Germany

* Corresponding authors:

David P. Morgan (david.morgan@zi-mannheim.de)

Gordon B. Feld (gordon.feld@zi-mannheim.de)

Abstract

Rewards play an important role in guiding which memories are formed. Dopamine has been shown to be an important neuromodulator mediating the effect of rewards on memory. In rodents dopaminergic activity during learning has been shown to enhance reactivation of memory traces during sleep, the mechanism driving the benefits of sleep on consolidation. However, evidence that sleep consolidates high reward memories more strongly in humans is mixed and small samples sizes (among other factors) likely drive these inconsistencies. Therefore, we will compare memory for rewarded information between intervals of sleep and wake in a large representative online sample. Participants (*N* = 1750; stratified German sample) will study images associated with high and low rewards and complete a memory test directly afterwards as well as after retention. Our main prediction is that sleep will enhance the retention of high over low reward images compared to wake. In general, we also expect sleep to enhance retention (evident through a reduced decrease in performance compared to wake) and rewards to improve memory. This study will reveal whether sleep facilitates selective consolidation or whether processes at encoding and shortly thereafter suffice. Additionally, it will provide a benchmark effect size to evaluate sleep-based interventions for psychiatric disorders (e.g., addiction). It will also allow us to explore moderators of the effect, such as age and education level.

Keywords: Sleep, reward, memory consolidation

Word count: 13969

Introduction

An accumulation of evidence indicates that sleep actively supports the stabilization and transformation of long-term memory[1–3] and for the most part studies have demonstrated that sleep compared to wakefulness benefits memory across declarative and procedural tasks [e.g., 4,5–15,but see 16,17–19]. The preferred explanation for the benefits of sleep on long term memory are attributed to active systems consolidation, but alternative explanations for the impact of sleep on memory do exist (e.g., Passive Interference Reduction Hypothesis[20], Opportunistic Consolidation[21]). The active systems consolidation hypothesis posits that the associative connections between elements of new information are encoded by the hippocampus and over time these connections are redistributed to the neo-cortex via systems consolidation[22]. This redistribution of information is thought to preferentially occur during sleep, whereby memory traces that were encoded throughout prior wakefulness are replayed repeatedly and thereby strengthened, although it should be noted that replay also occurs during wakefullness[23,24]. During active systems consolidation, sleep specific brain activity and especially the activity of hallmark oscillations (slow oscillations, hippocampal ripples and sleep spindles) that putatively coordinate this replay are thought to drive greater memory performance in those tasks [see 25,26–27,28,29, but also see 30,31]. The limited availability of these reactivation opportunities during sleep[32,33] suggests the selective consolidation of only relevant information, e.g., rewarded information[2]. However, it has not yet conclusively been shown that memories associated with a reward are consolidated more strongly during sleep.

Reward plays an important role in memory [34–42,for a review see 43]. In the declarative domain, its role has been demonstrated in humans using the motivated learning task. In that task, stimuli associated with a high or low reward are presented to participants and corresponding rewards are paid out for subsequent successful retrieval[34]. Researchers have consistently shown that memory for items associated with higher reward is greater than for those associated with lower rewards in humans[34–39]. Often such studies include a period of sleep, implicating the role of sleep in consolidation of reward associated memories[35,41,42, 44] and consolidation of highly rewarded information has been linked to sleep spindle activity[38,42]. This link to spindle activity during sleep suggests that sleep and reward fundamentally interact to consolidate motivationally relevant information indicating that reward plays a crucial role even long after encoding has taken place. However, the precise mechanism and time-frame by which sleep benefits reward memories remains ambiguous.

At encoding, dopamine modulates memory performance by recruiting reward areas in a ventral-striatum-ventral-tegmental-area-hippocampus feedback loop.[45] Using the Motivated Learning Task (in humans) a landmark study demonstrated that high reward cues activated the nucleus accumbens (located in the ventral striatum), the ventral tegmental area and the hippocampus during encoding[34]. Hippocampus activity was functionally coupled with activity in the ventral striatum and this predicted subsequent memory performance for high reward items. Behaviorally, this effect manifested as greater memory performance for high vs. low rewards at high levels of confidence. Regarding sleep, there is no consensus whether sleep enhances rewarded memories through additional dopaminergic neuromodulation during reactivation[41,45,46] or rather dopamine sets a tag during learning that leads to enhanced reactivation without additional dopaminergic neuromodulation[40]. Before answering this, it is first necessary to establish behaviorally whether or not sleep preferentially consolidates highly rewarded memories over lowly rewarded memories. Only then can the underlying neuronal mechanisms be characterized.

Independent of the putative underlying neurophysiological mechanisms, in humans, evidence is inconclusive, overall, regarding sleep’s role for rewarded memories. Several studies did not find that rewards enhance sleep-dependent memory consolidation.[10,47,48] In one experiment, participants were asked to learn object locations associated with high or low reward and were tested on those locations after a nap or a period of wakeful rest.[47] No difference in the magnitude of memory for high and low rewards was found between the napping and wakeful conditions. The absence of this effect is not uncommon and even extends to comparisons of a full night of sleep with typical daytime wakefulness and across recognition memory and verbal free recall tasks[10,48]. Nevertheless, sleep was still found to benefit memory overall.

This conflicts with another study using a procedural finger sequence tapping task.[49] In that study sleep preferentially consolidated highly rewarded sequences relative to a period of wakefulness. That finding was corroborated by another a study using a recognition memory task where a retention interval including a nap yielded greater memory for highly rewarded items compared to lowly rewarded items and this difference was not present in an equivalent wake condition[50]. However, in the latter experiment, there was no significant interaction between those groups, which despite the authors' conclusions would be necessary to conclude that high vs low reward items are preferentially consolidated during sleep[51]. One study found that the benefits of sleep on reward compared to wake may only unfold after much longer periods, which could allow further consolidation processes to take place[44].

Mutually exclusive theoretical conclusions from these studies can be drawn by ignoring the respective evidence that is not in their favor. Either sleep selectively consolidates information associated with high rewards[2] or reward related processes during encoding together with sleep-independent consolidation processes initiated shortly after learning are sufficient to enhance reward memory.[21] A third possibility is that consolidation does not affect reward related differences in memory performance and the difference are only due to encoding processes. Like for other memories it is evident that sleep is involved in the consolidation of rewarded memories per se, yet it is unclear whether sleep specifically enhances differences in memory performance based on reward amplitude (e.g., high vs. low rewarded information).

On the one hand, one could attempt to explain the divergent findings by evaluating the large number of differences in experimental designs (e.g., recognition vs recall, images vs words, napping vs 12-hours of sleep etc.). Here, one would conclude that the enhancement of sleep's beneficial effect by reward is sensitive to a host of moderators, as has been discussed for other inconsistencies in the field. For instance, mode of retrieval (e.g., free recall vs. recognition), mode of learning (e.g., implicit vs. explicit), material learned (e.g., declarative vs. procedural) and the timing of sleep (e.g., delay between learning and sleep onset) are all thought to moderate the sleep effect[52]. Such views have recently been reiterated in an assessment of the robustness of the sleep effect on memory[53]. However, this explanation leads to the unsatisfactory conclusion that the enhancement of sleep's beneficial effect on memory by reward is sensitive to moderators that were not systematically controlled in many of these studies. On the other hand, there exists a striking similarity between all of these experiments: low statistical power (maximum n = 20 per group)[e.g., 10, 30, 40-42,44,47,49,50]. Small samples have been shown to reduce generalizability, increase false negatives as well as false positives and can overestimate effect sizes[54,555], which may be the source of divergence. An argument that was sympathetically pointed out in a recent systematic review of the sleep and reward memory literature[56].

Our study will address this divergence by performing a large-scale investigation of the influence of rewards on sleep-dependent memory consolidation in the general population and asks the question: do rewards affect the magnitude of sleep-dependent memory consolidation? It is highly relevant to understand the impact of sleep on rewarded information since it guides (mal-) adaptive behavior such as unhealthy eating, smoking or alcohol consumption. Reward related learning mechanisms and other dopamine related plastic changes in the brain have been proposed to play a crucial role for establishing addictive behavior[57]. However, it remains unclear whether sleep-dependent consolidation of drug taking experiences occurs. Showing that sleep has a unique and sizable role for preferentially consolidating rewarded memory in the general population may fuel systematic investigations and targeted sleep interventions to better understand and treat, e.g., substance abuse and anxiety disorders. One such intervention may make use of the targeted memory reactivation procedure[58], where cues are used to reactivate memories during sleep. In some scenarios cueing during sleep has been shown to extinguish conditioned fear responses[59] and therefore extinguishing addictive behavior during sleep by using appropriate cues may be promising.

We will implement the AM:PM-PM:AM design in an online testing environment to collect a large sample of representative participants effectively, a strategy that has been successfully used by us before in a previous large-scale registered report in sleep and memory research[17]. In the AM:PM PM:AM design, participants undergo a wake condition, where the learning phase occurs in the morning (AM) and the test phase occurs in the evening (PM) on the same day. Participants also undergo a sleep condition, where the learning phase occurs in the evening (PM) and the test phase occurs the following morning (AM). In recent years researchers investigating the impact of sleep on memory have begun using web-based alternatives by performing online sleep experiments[60,61]. It should be noted that generally such experiments do not appear to limit the capacity to detect the impact of sleep on memory. Reward memory will be measured using a paradigm adapted from earlier studies[34,40,41] and recently validated in our laboratory to yield positive effects of reward on memory performance (see supplementary material: https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg), where participants (N = 1750) will study images associated with high to low rewards and will retain them across sleep and wakefulness. This paradigm uses a recognition task to measure memory performance and although recognition tasks have been shown to be somewhat less sensitive to the effect of sleep on memory than free or cued recall procedures[62] our power analysis indicates that we have sufficient power.

We predict (see Figure 1 and Table 1), **H1) that sleep will yield greater retention compared to an equivalent period of wake (although we expect a general decline in performance across retention); H2) that items associated with high rewards will be better retained compared to those associated with low rewards; H3) the magnitude of the decline of high reward memories will be less in the sleep condition compared to the wake condition.** In addition, to these three main hypotheses our study will include several control variables to investigate known confounding factors (i.e., vigilance, sleepiness, general retrieval performance, memory strength and task difficulty) as well as variables that will allow us to explore moderating factors (i.e., age, education status, morningness-eveningness, mental health, shift work, travel and medication). Of note, our study will not be able to show how sleep parameters mechanistically affect reward memory, as the sleep vs. wake design cannot reveal such relationships (irrespective of whether sleep deprivation or as in our case day wakefulness is being used). However, our study will enable the planning of much more resource intensive mechanistic studies that manipulate sleep (e.g., by drugs) by delivering an effects size estimate with much less uncertainty than previously.

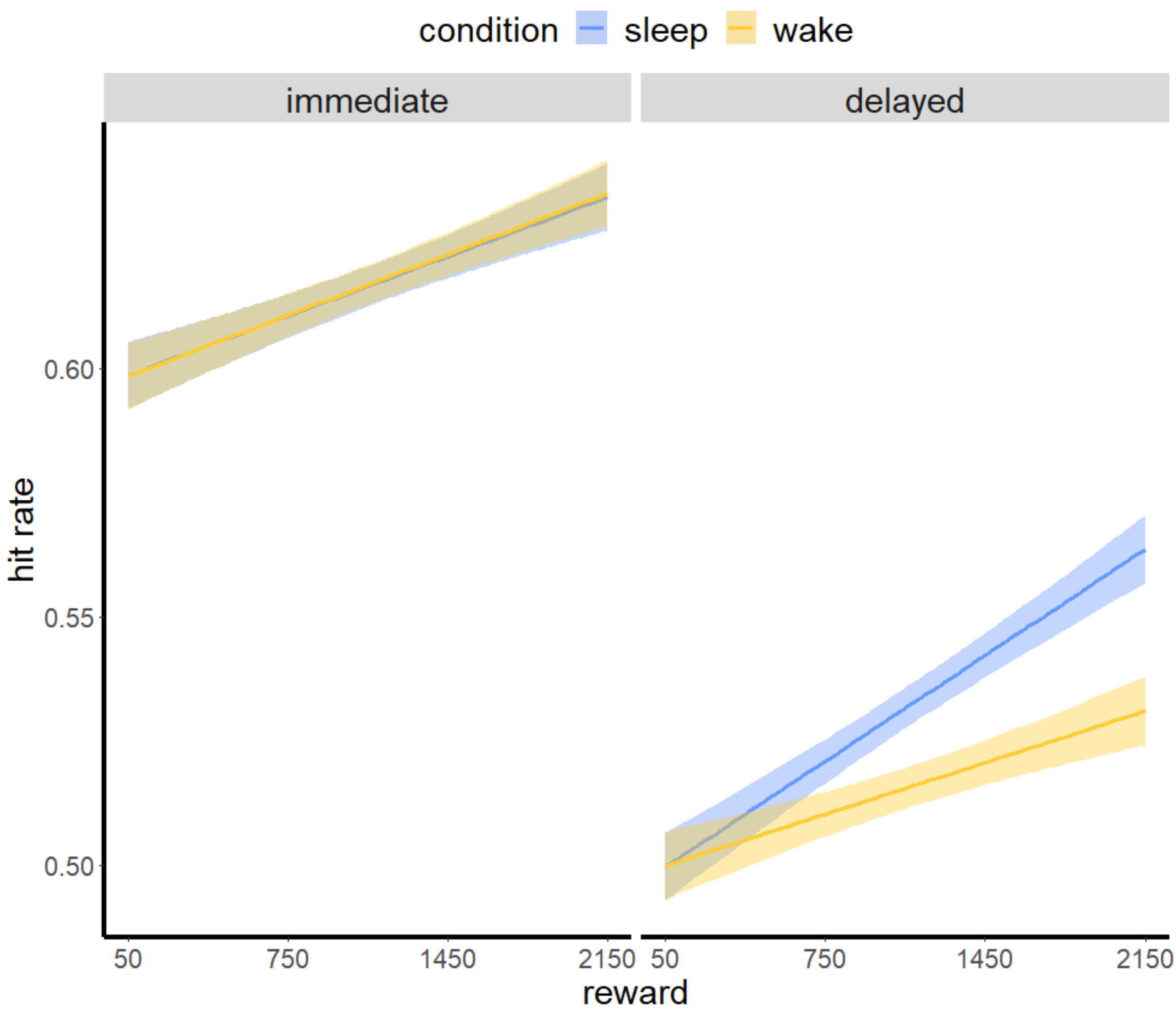


Figure 1. Visualization of the first simulated run of our predictions produced by our data generating model, with a sample size of $N = 1750$. The estimated memory performance for each reward category is represented by the thick lines and shaded areas represent standard error estimated using linear models. Note that the standard error is small due to the large sample size.

Table 1. Design table.

Sampling plan:	As per our data simulation described on page 29 a maximum of *N* = 1750 participants will be collected to test the hypotheses described below.				
Question	**Hypothesis**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**	**Interpretation given different outcomes**	**Theory that could be shown wrong by the outcomes**
Does sleep yield greater memory performance compared to an equivalent period of wake?	**H1:** In the sleep condition, hit rate for the Motivated Learning Task at delayed recognition will be greater when compared to the wake condition.	Linear mixed model: The interaction between retention x time point will be significant. Planned comparisons: the main effect of retention at delayed recognition will be significant. Equivalence: Our smallest effect size of interest for this comparison is in the same range as the reward effect or the reward X retention X timepoint interaction (both unstandardised estimates of .015; see Table 5). This is why we assume equivalence if the 90% confidence interval around the parameter estimated from the actual data does not include .015.	In our model the effect of retention is 0 because sleep only exerts its influence at delayed recognition for high rewards only. However, it is still plausible that an overall difference between sleep and wake emerges at delayed recognition (e.g., this depends on how motivating the low rewards are). For our smallest effect size of interest we chose a value that is comparable with our smallest effect size of interest for the reward effect and the reward X retention X timepoint interaction.	Significance: If the main effect of retention at delayed recognition is significant and the mean memory performance for the sleep condition is greater compared to the wake condition, sleep positively affects memory in comparison to an equivalent period of wake. This would confirm **H1**. Equivalence: If the difference between sleep and wake conditions is found to be statistically equivalent this would disconfirm **H1**. If the difference between sleep and wake conditions is found to be equivalent, H3 described below will still be tested as alternative interpretations may exist as described in our supplemental information, which can demonstrate the preferential impact of sleep on consolidation of high vs. low rewards.	If there is no difference in memory performance between the sleep and wake groups at delayed recognition this could demonstrate that the theory that sleep benefits memory generally over and above a period of wake is incorrect. Alternatively, if there is a difference in memory performance between the sleep and wake groups at delayed testing and the wake condition yields better memory performance than the sleep condition then this would point towards periods of wakefulness being more beneficial for memory retention as compared to sleep.”
Does information associated with high rewards yield greater memory performance compared to low rewards?	**H2:** The hit rate for high rewards will be greater than the hit rate for low rewards at delayed recognition	Linear mixed model: The interaction between reward x time point will be significant. Planned comparisons: the main effect of reward at immediate and delayed testing will be significant Equivalence: If the 90% confidence interval around the parameter estimated from the actual data does not include the parameter we set in our simulations (i.e. .015; see Table 5) we assume equivalence. This is because our simulations are based on our minimum effect sizes of interest.	The power analysis derived from our data simulation indicated that we have 95% power to detect at least an unstandardized effect size of .015 with an alpha of $p < .020$. This effect size was selected on the basis of our pilot data for the motivated learning task presented in the supplemental material.	Significance: If the main effect of reward at immediate and delayed recognition is significant and the mean memory performance for the high rewards is greater compared to low rewards, high rewards have a greater impact on positively affecting memory in comparison to low rewards. This would confirm **H2**. Equivalence: If the difference between high rewards and low rewards is found to be statistically equivalent this would disconfirm **H2**.	If there is no difference in memory performance between low rewards and high rewards this could demonstrate that the theory that high rewards are beneficial for memory is incorrect.
Does sleep yield greater recognition memory performance for high vs. low reward items?	**H3:** The magnitude of the positive effect of reward on the hit rate will be greater in the sleep condition compared to the wake condition at delayed recognition.	Linear mixed model: The interaction between retention x time point x reward will be significant. Planned comparisons: The interaction between retention x reward at delayed recognition will be significant. Equivalence: If the 90% confidence interval around the parameter estimated from the actual data does not include the parameter we set in our simulations (i.e. .015; see Table 5) we assume equivalence. This is because our simulations are based on our minimum effect sizes of interest.	The power analysis derived from our data simulation indicated that we have 95% power to detect at least an unstandardized effect size of .015 with an alpha of $p < .020$. This effect size was selected due to the unknown nature of the size of the interaction between retention and reward and resource constraints.	Significance: if the interaction between retention x time point x reward and the interaction between retention and reward at delayed recognition is significant such that the magnitude of the benefit of reward on memory is greater in the sleep condition compared to the wake condition at delayed testing this would confirm **H3**. Equivalence: If the difference in the magnitude of the effect of reward on memory between sleep and wake conditions at delayed recognition is found to be statistically equivalent this would disconfirm **H2**.	If there is not a greater positive effect of reward on memory performance in the sleep condition compared to the wake condition at delayed recognition then the theory that sleep preferentially consolidates reward information could be wrong. This would indicate that rewards exert their influence on memory predominantly during encoding.

Note: The factor retention refers to the retention manipulation and contains the two levels sleep and wake. Also, since we are using a declarative task, we cannot generalize our inferences to the procedural domain and declarative memory is meant whenever we write memory in this table

Methods

Participants

The size of our sample is guided by resource constraints as well as a data simulation based on the data shown in Figure 1 and previous literature [34,40,44,49,50]. Our predictions indicated that 1750 participants suffice to detect a very small effect size and a broad range of much larger effect sizes to achieve $1 - \beta = 0.95$. Of note, the simulation uses a data generating linear mixed model with specific input parameters shown in our analysis plan. Proportions of our representative sample stratified across sex (male and female), age (ages 20-29 to 50-59 years), highest professional qualification and highest school level qualification were calculated based on the German 2011 Census (See Figure 2). Sampling of strata will be ended individually as soon as they are full.

Participants will take part in this experiment online and will be recruited using targeted online advertisements on popular social media websites (e.g., Facebook, twitter) and media outlets (e.g., news websites). We will use Meta Advertisements, an advertisement service using Facebook and Instagram to target strata that we identify as currently under sampled. We will also use our contacts writing for national news outlets to further boost the visibility of the study. We will additionally implement a "refer a friend" strategy where participants can refer one or more friends. If at least one friend then goes on to complete the procedure the referrer will automatically receive a 5€ Amazon voucher. Participants will not receive compensation for their participation, but will have the chance to win a voucher dependent on their performance in the task. The voucher values will be 500 x 7.50€, 150 x 15€, 125 x 20€ and 100 x 25€ (adding up to 11000 € in vouchers). The average amount of the vouchers is thus 6.28 €, which is approximately the average bonus that we paid out in our pilot study which was the basis of our power calculation (see supplementary material; https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg). Participants will be informed that each point (gem) earned in the reward task will improve their chances of winning a high value voucher, but that this also depends on the performance of the other participants. To minimize attrition participants who do not complete the whole experiment will not have the opportunity to win a voucher. The German Psychological Society (DGPs) ethics committee approved this experiment. Written informed consent will be obtained from participants prior to participating in this experiment. Since we propose to collect a large stratified sample across multiple sessions, we estimate that data collection will be completed within 12 months.

Our inclusion and exclusion criteria are presented below in Table 2. Participants who meet the exclusion criteria will not be included in the data analysis and will be resampled until our desired sample size is achieved. We have chosen not to exclude participants with mental health conditions which can impact participants' memory consolidation. This is because based on previous experience conducting large-scale online sleep experiments, such exclusion criteria can cause severe limitations on the recruitment process, since mental health issues are quite wide spread (i.e., one in three women and about one in four men aged 18–79 in Germany meets diagnostic criteria of at least one mental disorder during the past 12 months[63]). Additionally, a main goal of this research is to yield a demographically diverse (representative) sample, which can be used to derive an effect size estimate of the impact of sleep on reward memory, to be used in therapeutic settings. Therefore, the effect size must be as generalizable as possible beyond the samples typically used in sleep and memory experiments which are largely performed with highly educated young students. Such samples create a translational gap between basic science and clinical research which limits the generalization of our findings to samples with mental health conditions.

Table 2. Inclusion criteria necessary to participate in the experiment and exclusion criteria to be included in the data analysis.

Inclusion Criteria	Exclusion criteria
Stratification: Sex, Male or Female Aged 20-59	Napping between study and test in the wake condition
Highest professional qualification Highest school leaving qualification	Sleeping less than 6 hours in the sleep condition
Resident in Germany	Consumption of alcohol between study and test within sessions 1 or 2
	Participants who respond too slowly on the Flankers task on 3 consecutive trials
	Participants who respond too slowly on 3 consecutive trials for the recognition memory test
	Participants who fail the validation questions on any occasion after their second attempt
	A d' score ± 3 SD away from the mean within each age category collapsed across timepoint (immediate vs. delayed), retention (sleep vs. wake), rewards and durations

To ensure completion of the sample, we will implement the following contingencies incrementally: 1) If after 7 months of data collection we have not achieved at least 50% of our desired sample size we will collapse the strata of the "highest professional qualification" and "highest school-leaving qualification" categories into three groups respectively; 2) If after 9 months of data collection we have not achieved at least 50% of our desired sample we will remove the aforementioned education strata; and 3) Finally, if after 11 months of data collection we have not achieved at least 50% of our desired sample we will open up data collection to the UK and USA (English versions of all materials already exist in the lab). In each scenario the stratification will be adjusted.

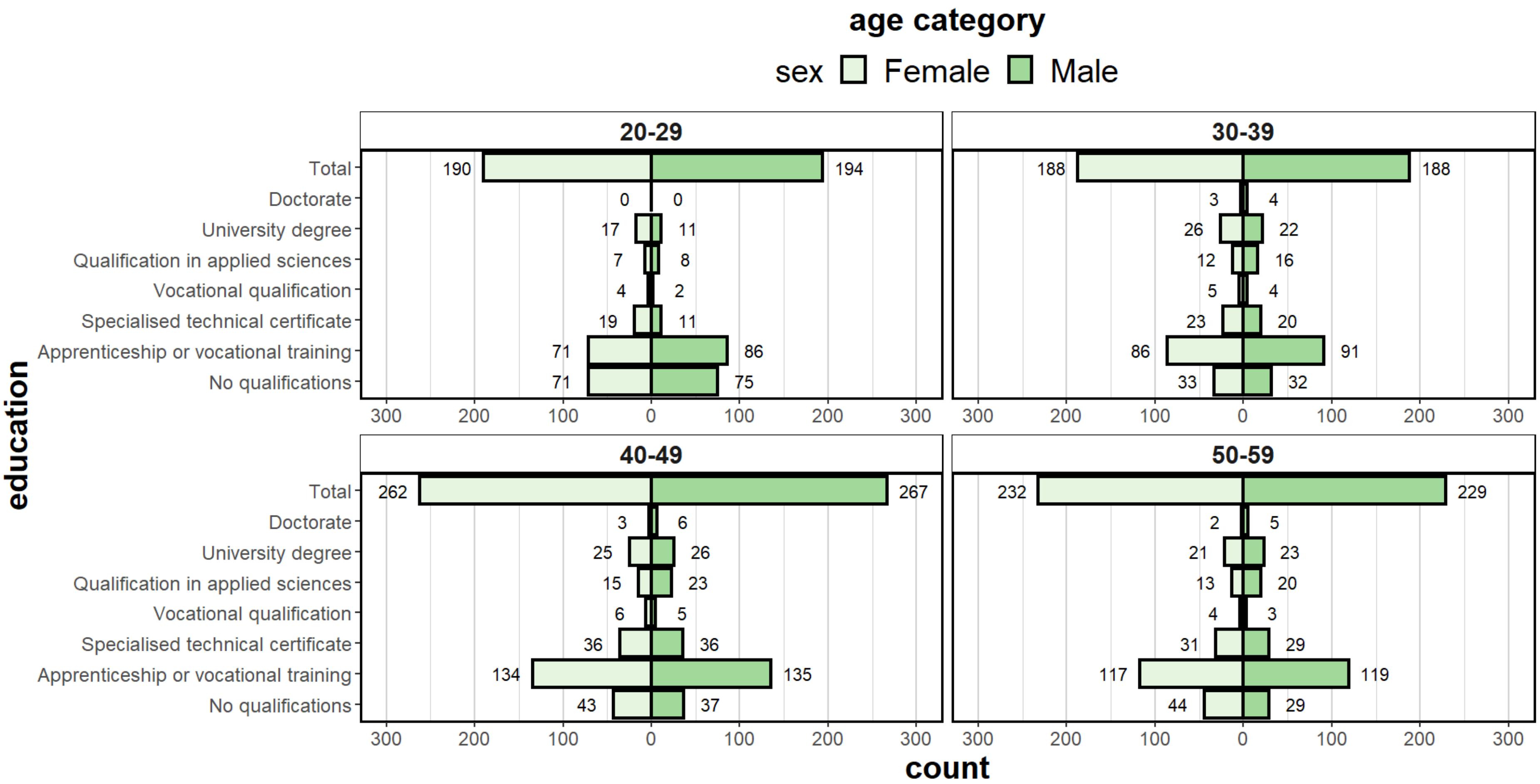
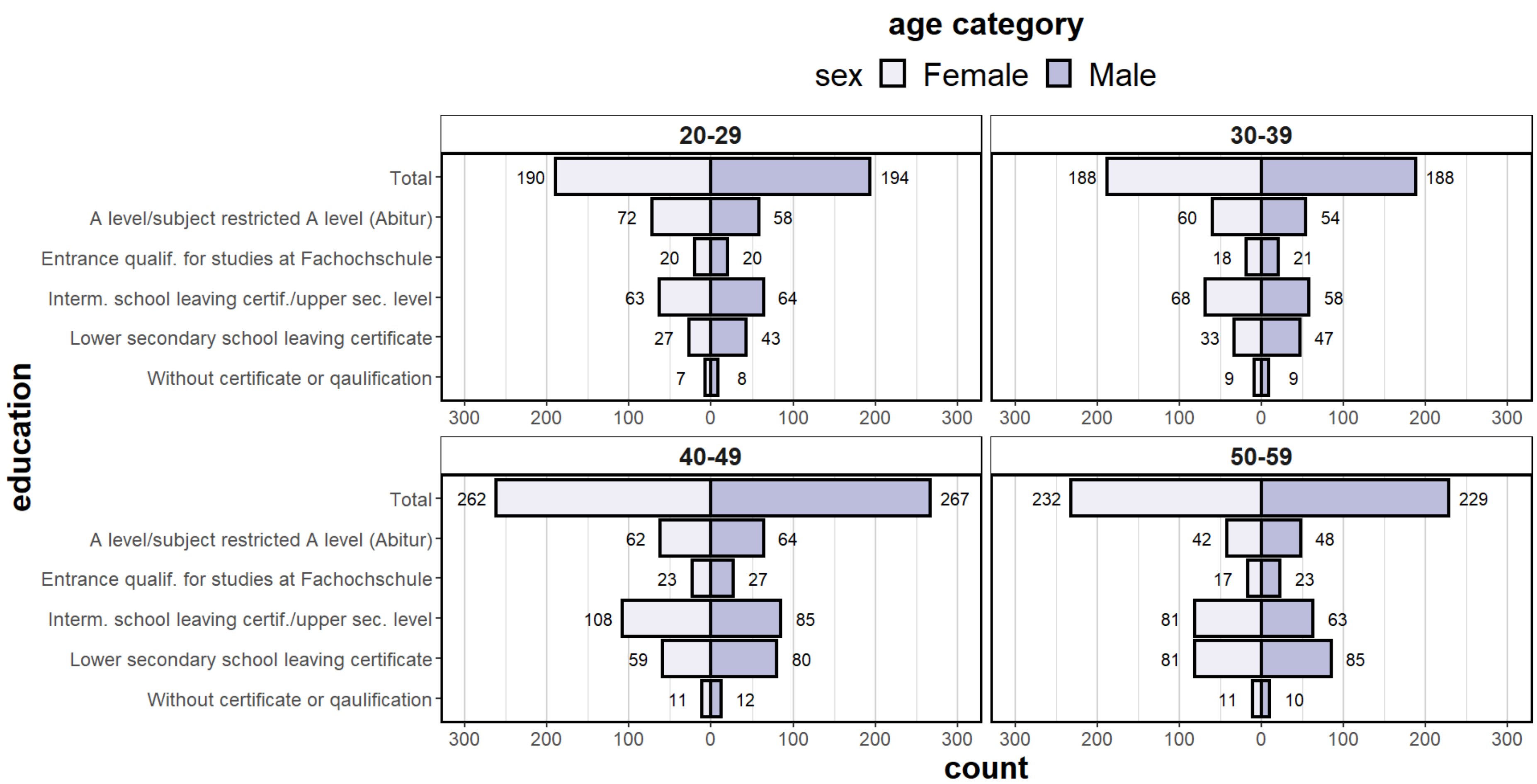


Figure 2. Demographic profile of the proposed stratified sample. For simplicity the strata are presented by highest professional qualification and highest school level qualification. Nevertheless the final strata will consist of each stratified combination of the relevant categories, see Supplemental Material. A) y-axes indicate the highest professional qualification attained and x-axes indicate the number of participants required within each sex, age and professional qualification category combination. The precise number of participants that are required to yield a representative sample are labeled against each bar for each sex, age and education category.



B) y-axes indicate the highest school level qualification attained and x-axes indicate the number of participants required within each sex, age and school level qualification category combination. Again, the precise number of participants that are required to yield a representative sample are labeled against each bar for each sex, age and education category.[1]

[1] The precise definitions of the education for highest professional qualification and highest school level qualification categories can be found at: https://shorturl.at/lpz58

Design

Participants will undergo the Motivated Learning task (see Figure 3) over two experimental sessions in a balanced AM-PM PM-AM cross-over design (see Figure 4), once with a retention interval of wake and a second time with a retention interval of sleep. From the introduction it is clear that a choice must be made to either assess memory using a procedural or a declarative task, which both have been shown to benefit from sleep in the retention interval[62,64,65] .We have chosen the former as in the literature there is no clear indication that a procedural task is better suited.

The Motivated Learning task is a recognition memory procedure and in our experiment memory is tested twice in both sleep and wake conditions, once after learning (to examine baseline performance) and again after sleep or wake. In the AM-PM PM-AM design when participants undergo the sleep condition they study images, complete an immediate recognition memory test in the evening (PM) and are tested once again the following morning (AM). Participants in the wake condition study images and complete an immediate recognition test in the morning and are subsequently tested again in the evening (PM) on the same day. Therefore, the experimental design has two within-subjects factors Retention (sleep; wake) and Time point (immediate; delayed) with two levels in each. The images that participants study are associated with rewards of four different magnitudes adding an additional within-subject factor reward with four levels to the design (50, 750, 1450, 2150). Our main analysis strategy for this design is based on linear mixed models (see Analysis Plan for details).

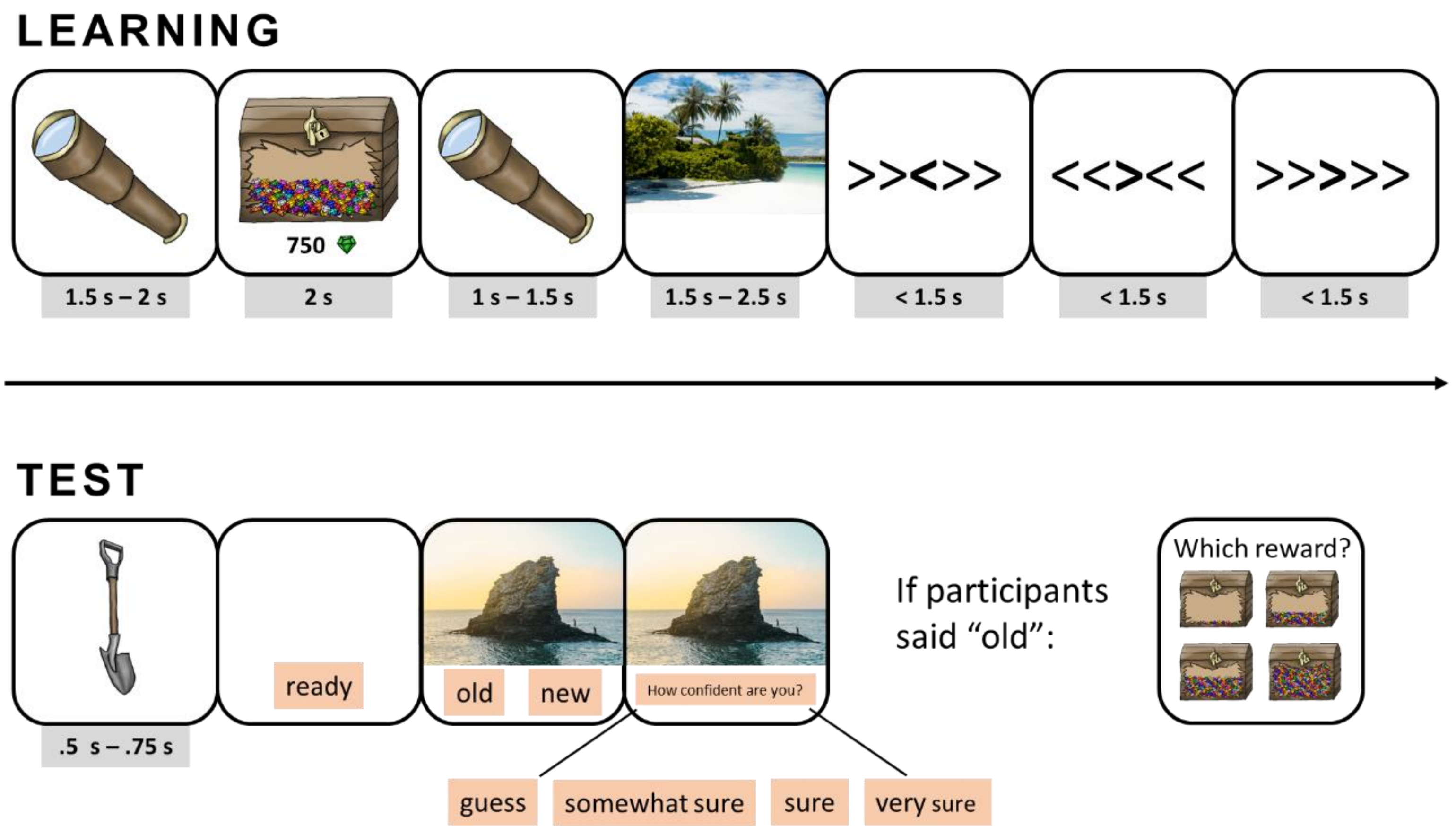


Figure 3. Motivated Learning Task. Example trials for the learning and recognition tasks. During learning, participants are required to memorise landscape images. Each image is associated with a different reward shown as gems in a treasure chest before each image. During test participants' memory for those images is tested. For each landscape image, participants decide whether an image is old (i.e., the image was shown during learning) or new (i.e., the image was not shown during learning) and rate their confidence in their decision using a 4-point Likert scale (guess, somewhat sure, sure, very sure). If a participant decides that an image is old, they will be asked to indicate the reward amount that image was associated with. If a participant makes a correct old/new decision they are rewarded the amount that was presented alongside the image during learning and if the participant makes an incorrect decision, they lose the mean value of all possible rewards (i.e. 1100 gems).

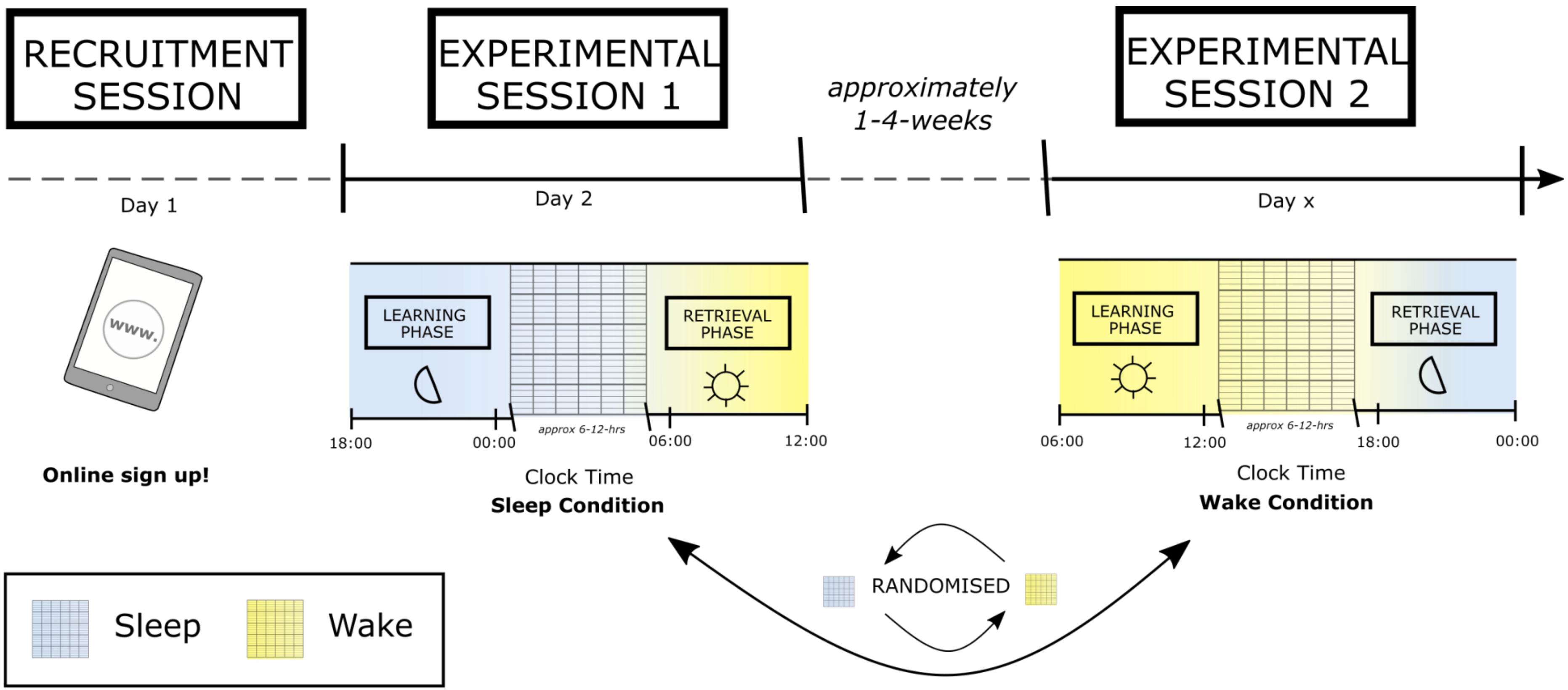


Figure 4. Experimental procedure for the proposed experiment. Before starting the experimental sessions, participants complete a recruitment session where their demographic information is collected and a number of questionnaires are completed. If participants are eligible to participate they undergo two experimental sessions, once with a retention interval of sleep and again with a retention interval of wake (in a counter-balanced order). In both sessions the procedure is otherwise identical. Both sessions are separated by at least 1 week and a maximum of 4 weeks.

Procedure

Participants are pseudo-randomly assigned to complete the sleep or wake condition first, such that the order of the sleep and wake conditions is balanced across the sample. This is also accounted for in our stratification where half of the participants within each stratum will complete either the sleep or wake condition first. Participants will complete the sessions separated by a minimum of 1 week and a maximum of 4 weeks. In the sleep condition, participants complete the learning phase (i.e. learning task and immediate recognition task) in the evening (between 18:00 – 00:00) and the retrieval phase (i.e. delayed recognition task) in the morning (between 06:00 – 12:00). In the wake condition, participants complete the learning phase in the morning (between 06:00 – 12:00) and the retrieval phase in the evening (between 18:00 – 00:00). In both cases participants must select a two-hour window separated by 12 hours in which the learning and test phases will be completed (i.e., 06:00 – 08:00, 08:00 – 10:00 or 10:00 – 12:00 and 18:00 – 20:00, 20:00 – 22:00 or 00:00). For example, if the participant completes the learning phase between 08:00 – 10:00 and the test phase between 20:00 – 22:00 in the wake condition they must also participate in both phases between 20:00 – 22:00 and 08:00 – 10:00 in the sleep condition. This will help to constrain differences in the retention interval between the sleep and wake conditions.

Recruitment session. The recruitment session can take place at any time prior to the experiment and participants will be asked to use the same device they used to sign up for all sessions. A captcha will be used on all session to avoid including bots. All data will be collected using the participant's chosen device, limited to either a computer, laptop or tablet. Therefore, participants completing the experiment on a computer or laptop will respond using their mouse and keyboard whereas those using a tablet will be able to respond using touch screen buttons. Next, they will receive information about the study and digitally sign the consent form. After that, they will answer the demographic questions and depending on strata vacancies they will be allowed to participate. Then participants will be introduced to the cover story of the experiment.

During the recruitment session participants will also complete the screening questions, St Mary's Hospital Sleep (SMHS) Questionnaire[66], Epworth Sleepiness Scale (ESS)[67], reduced Morningness-Eeveningness Questionaire (rMEQ)[68,69], Pittsburgh Sleep Quality Index (PSQI)[70], Alcohol Use Disorders Identification Test (AUDIT)[71] ,the Caffeine Consumption Questionnaire[72], the Behavioral Inhibition System/Behavioral Activation System scale (BIS/BAS)[73], the Becks Depression Inventory – Short Form (BDI - SF)[74], and finally the shortened version of the Hagen Matrices Test (HMT-S)[75]. For the Learning phase in Experimental Session 1 (see Figure 3), participants will receive an email with a participation link and times when the experiment can be started depending on which retention condition (sleep or wake) they are assigned to first. Participants will receive automated emails shortly before each part of the experimental procedure to remind them to participate. Participants will be asked to refrain from drinking alcohol (i.e. participants should not drink alcohol 24-hours prior to and during the first or second experimental sessions) and consume no more or less than their usual caffeine intake whilst they are actively participating in this experiment.

Experimental sessions. The first experimental session will take place the earliest within 24-hours of participants completing the recruitment session. At the beginning of the experimental session participants first indicate when they last consumed caffeine or alcohol and how much they consumed. Then they indicate their subjective sleepiness (SSS)[76] and complete a vigilance task (PVT)[77]. Next, they are presented with instructions describing the Motivated Learning task and how they should perform the first and second parts of the learning phase, the learning task (duration approximately 19 minutes) and the immediate recognition task (duration approximately 14 minutes). In those instructions participants are explicitly informed about the reward contingencies described on p. 24. Participants are then asked the validation questions to ensure that they understand the task and then undergo the learning phase (consisting of the learning task and the immediate recognition task). At this point, participants in the sleep condition will be instructed to go to sleep at their usual bedtime and wake up at their usual waking time and participants in the wake condition will be asked not to nap, since even ultra-short naps may allow for sleep-dependent consolidation[78].After at least twelve hours, participants will return to the experiment. Participants completing the sleep condition will first answer questions about their sleep quality (SMHS)[66] and will answer the sleep related questions. Participants completing the wake condition will also be asked the sleep related questions (pertaining to the night before participating) and will be asked "have you taken a nap today?" and if so "How long did the nap last, in minutes?". All participants are once again asked if they have consumed any alcohol or caffeine, how much they consumed, rate their subjective sleepiness and vigilance is assessed a second time. They then receive instructions on completing the retrieval phase, answer the validation questions a second time, complete the delayed recognition task (duration approximately 14 minutes) and complete a verbal fluency task[79]. This concludes one session of the experiment and participants will then receive further instructions about the second session. Participants repeat the experimental session, known as Experimental Session 2 (see Figure 4), but at different times depending on which retention condition they completed first. At the end of the second session, participants will be debriefed and receive a profile of the questionnaire data they have provided relative to population. This feedback on their questionnaire data is used as a further incentive for participants to complete the study.

Follow-up. Participants will complete a long-term retrieval phase of the Motivated Learning Task 3-months after completing the second experimental session. Data collected on this part of the task will be used for exploratory purposes only.

Materials

Motivated Learning Task. This task was adapted from a previous study and the appearance of the task has been adapted to map onto the cover story below[34].

Cover Story. To enhance motivation in the Motivated Learning Task we gamified it with a cover story, where the participants are part of a pirate ship crew. As a crewmember the aim of the participant is to scout treasure hidden in different locations (landscape images) occupied by their allies, corresponding to the first part of the learning phase of the experiment. Participants scout at those locations alone so they cannot take the treasure with them. Participants navigate between the locations in the form of the Flankers Task embedded amongst a treasure map. Thus, they must remember the locations and scavenge the treasure when they return with the crew, which corresponds to the second part of the learning phase (immediate recognition) and the retrieval phase (delayed recognition). When participants return for the immediate or delayed recognition along with their crew, they revisit "old" locations (the locations shown during the first part of the learning phase) and "new" locations (locations that were not shown during the first part of the learning phase and that are known to be inhabited by rival pirate clans). The goal of the participant is to "dig" at "old" locations as that is a hit and they will be rewarded treasure. They should avoid choosing to "not dig" at "old" locations as that is a miss and the crew captain will punish the crew with a loss of treasure. Since the "new" locations are occupied by rival pirate crews "digging" at those locations, a false alarm, is costly, as the rival pirates will take treasure from them. However, if participants choose to "not dig" at the "new" locations, a correct rejection, the crew captain will reward them with treasure as digging there could have risked the crew's safety.

Task description. First, a fixation spyglass is shown during a jittered inter trial interval (ITI, 1500 – 2000 ms) indicating that it is time for the participant to begin scouting for treasure at a new location (see Figure 3). Then a treasure chest is shown (2000 ms) indicating how much treasure can be gained for correct recognition of this image during immediate or delayed recognition using one of four reward magnitudes (50, 750, 1450, 2150). This is followed by an additional fixation spyglass. Next, the image of the location is presented. Each image is only shown once during the learning task. After viewing each image, participants complete three trials of the flanker task to prevent rehearsal[70]. Participants are informed that their chances of winning a monetary bonus increase the more gems they collect. We will use four different image exposure durations (in ms 1500, 1833, 2167, 2500) to control for encoding strength. The durations and rewards associated with each image are counterbalanced so that all reward magnitudes are presented with each duration. Each of the sixteen reward × duration combinations are implemented eight times (using different images) therefore participants are shown 128 images during the first part of the learning phase. The images are pseudo randomly presented to ensure that the same reward or duration do not occur consecutively. The learning task is split into eight blocks with sixteen images presented per block and in each block at least six and at most 10 images will be associated with high reward (either 750 or 1100 gems) and at least six and at most 10 images will be associated with a long duration (either 1500 or 2000 ms). At the beginning and the end of the learning task participants will complete 4 additional trials, each with pseudo-random rewards and durations occurring only once, to buffer primacy and recency effects.

In the flanker task arrows will be presented to the participant and the direction that the middle arrow faces will correspond to the directional button which the participant must press[69] whereas the arrows adjacent to the middle arrow must be ignored. There are congruent (i.e. flanking arrows face the same direction, >>>>>) and incongruent (i.e. the flanking arrows face the opposite direction, >><>>) trials that will be split across all trials of the learning phase. If participants respond too slowly (i.e. >1.5s) they will be asked to speed up, participants who respond too slowly after three consecutive trials of the learning task (i.e., on nine consecutive flankers) will be excluded from the data analysis. The flankers trials are pseudo randomized such that a maximum of six trials can be of the same congruency and orientation in a row.

Participants’ memory for half of the learned images (i.e., 64 of 128) is tested in the immediate recognition task directly after the learning task and the other half of the images is tested in the delayed recognition task in the retrieval phase. One trial of the test phase is shown in Figure 3. A recognition trial begins with a shovel during a jittered ITI (500 - 1000ms), indicating to participants that they will begin collecting treasure. Next participants must click a ‘continue’ button to ensure that the mouse pointer or finger (when using a tablet) is in approximately the same position for all trials. Then participants are shown the image of the location and make three decisions.

First, participants must indicate if the image is “old” or “new” to measure memory performance. If the image is “old” and the participant decides the image is “old”, then that is a hit and participants are rewarded the number of gems that the image is associated with. If the image is “new” and the participant decides that the image is “new” then that is a correct rejection and they are rewarded the average reward (1100 gems). If the image is “old” and the participant decides that the image is “new” then that is a miss and the participant loses the average reward. If the image is “new” and the participant decides that the image is “old” then that is a false alarm and they lose the average reward amount. The second question participants are asked is “how certain are you?” using a four-point Likert scale (“guess”, “somewhat sure”, “sure”, “very sure”). Confidence is routinely measured in recognition memory tasks and we have decided to keep this assessment, as in some cases reward effects have been reported to be more pronounced for high confidence items[34]. Finally, if the participant decided that the image is “old” they are asked “which treasure do you think can be found here?” and must select one of the four reward options that they believe the current image is associated with. This question will measure source memory for the reward categories. Participants are asked to decide if the image is “old” or “new”, rate their confidence and select the associated reward as fast as possible. Each decision must be made within 5000 ms. If participants do not respond within that time to either of the questions they will receive a warning message. After three warning messages participants will be reminded that they will be excluded from the experiment if they do not respond fast enough. When participants are shown a warning message, participants are still able to respond. Response times starting from the presentation of the location to the time at which an old/new decision is made and from that decision to the time at which participants rate their confidence and from that time until a reward is selected will be recorded for exploratory analyses.

For the immediate recognition task there will be eight blocks with 16 trials each, equaling 128 trials. Sixty-four of the 128 trials are old landscape images (i.e., half of the learned images) and the remaining 64 are new landscape images. Old and new trials are pseudo randomly presented such that no more than four target or lure trials can occur in a row and the same reward and duration can also not appear in a row. Between each block, participants will be shown an animation of the number of gems they have collected so far. However, this mock feedback is not influenced by true performance but rather corresponds to the slightly jittered mean number of gems that could be earned $\pm$ 1 SD with 50% accuracy during the task. This is done to keep motivation high for all participants irrespective of their true performance. In our pilot experiments (see supplementary material; https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg), none of the participants noticed this was mock feedback.

In the delayed recognition task, participants complete the same procedure as in the immediate recognition task, except that participants are presented with a different set of images, i.e., the remaining 64 old images and 64 completely new images. In the follow-up, participants are shown all 256-target images they were shown during the first and second recognition phases in sessions 1 and 2 and will be shown 256 completely new images as lures.

Landscape images. The images are allocated in a way which means that each image is balanced across the combinations of reward and duration as well as the different time points and old/new assignments. The landscape images were collected from the creative commons online repository (https://search.creativecommons.org/). A pilot study conducted on Prolific (https://www.prolific.co/; *N* = 152, see supplementary material; https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg) assessed those images in terms of aesthetics, composition, memorability, familiarity, whether or not the exact images have been seen before and memory accuracy. Participants rated images on those factors and subsequently completed a recognition memory test. This pilot allowed us to balance out differences on those factors between the images across the conditions as well as reward and duration categories in the Motivated Learning Task and eliminate images that are extremely recognizable.

Demographic information. All participants will be asked the following questions in a custom online questionnaire: What is your age?; What is your biological sex?; Which gender do you identify as?; What is your ethnicity?; What is your highest level of school education?; What is you highest professional qualification?; What is your aspired level of education?; Which type of school did you go to?; What is your current occupation? What is your relationship status?; Do you have children and if so how old are they?; Are you currently living in Germany? If yes, what are the first two numbers of your postcode? Do you live in an urban or rural area?

Participant questionnaire. Participants will be asked the following yes/no questions: Do you currently smoke cigarettes? If “yes” how long have you smoked them for?; Do you currently take any recreational drugs? If “yes” which drugs do you take and how long have you taken them for?; Do you currently suffer from a diagnosed sleep disorder?; Do you currently suffer from a diagnosed neurological disorder?; Are you currently taking any prescribed medication?; Do you currently suffer from a diagnosed addiction disorder?; Do you currently suffer from a diagnosed mental health disorder?; Participants will be asked to state, which disorder they suffer from and which medication they are taking if they answer “yes” to the questions regarding sleep, neurological, addiction or mental health disorders or those who answer “yes” to taking medication will be asked to indicate which medication they are taking; Have you traveled across time zones within the past three weeks? If “yes”, where did you travel to?; Do you currently work as a shift worker? Or have you ever worked as a shift worker?. If “yes”, how long have/did you work(ed) as a shift worker? and have you worked night shifts within the past 6 months?. The data collected from this questionnaire will be used for exploratory purposes only. Of note, we will not use these questions top exclude participants even though this is done in similar research. We do this to enable exploring moderators.

Sleep questionnaires.

The data collected from the following sleep related questionnaires will be used for exploratory analyses only, examples of those analyses are provided underneath each questionnaire.

Caffeine Consumption Questionnaire. Participants will indicate which caffeinated products they have consumed throughout the day before participating (including coffee, decaffeinated coffee, espresso, black, green, white, or mate tea, cocoa drink, iced tea, drinks with tea extract, cola and mixed cola beverages, energy drink, energy shot, alcopops with energy drink, cola or coffee and chocolate) and will indicate when they consumed those products (breakfast, between breakfast and lunch, lunch, between lunch and dinner, dinner and after dinner) [72]. The caffeine consumption questionnaire will be translated directly from English into German. The amount of caffeine which participants have consumed will be used in exploratory analyses to determine whether or not memory performance in the sleep and wake conditions for high and low rewards is moderated by caffeine consumption.

Sleep Related Questions. All participants will indicate their bedtime, rising time and the length of time that they spent asleep. Participants will also indicate if they had any awakenings and if so, how many they had.

Epworth Sleepiness Scale (ESS). The ESS asks participants to rate their general sleepiness in eight everyday scenarios using a four-point scale (0 = would never doze, 1 = slight chance of dozing, 2 = moderate chance of dozing and 3 = high chance of dozing)[67]. Scores on the ESS range from 0-24, a low ESS score indicates low levels of general sleepiness and a high score indicates high levels of general sleepiness. The German version of the ESS will be used in this experiment[80]. The Epworth sleepiness scale will be used to determine whether higher levels of sleepiness cause detrimental effects to the relationship between sleep and memory consolidation for rewarded information.

Stanford Sleepiness Scale (SSS). The SSS asks participants to rate their current level of subjective sleepiness on a seven-point scale (1 = feeling active, vital, alert or wide awake; 7 = no longer fighting sleep, sleep onset soon, having dream-like thoughts)[76]. A low score on the SSS indicates a low level of state sleepiness and a high score on the SSS indicates a high state level of sleepiness. SSS scores will be used to determine whether differences in memory performance between the sleep and wake condition may be attributed to differences in subjective sleepiness. The SSS will be directly translated from English to German[as in 41].

Reduced Morningness-Eveneningness Questionnaire (MEQr). The MEQr is a reduced version of the full Morningness-Eveneningness Questionaire (MEQ), which uses only 5-items from the MEQ (i.e. items 1, 7, 10, 18 and 19: e.g., "During the first half hour after having woken in the morning, how tired do you feel?")[69,81]. The MEQr measures an individual's chronotype (i.e. the time of day that an individual feels most alert)[69] and scores on the MEQr range from 4-26. Scores below 12 are indicative of a morning type whereas scores greater than 17 are indicative of an evening type. Scores between 12 and 17 are indicative of neither type[69]. The German version of the MEQr will be used in this experiment[81]. The MEQr will be used to determine whether chronotype synchrony (i.e., whether you are participating at a time that matches your chronotype) impacts the relationship between sleep and memory consolidation for reward.

St Mary's Hospital Sleep (SMHS) Questionnaire. The SMHS is a subjective measure of sleep quality over the last 24 hours[66]. Participants will answer items 6 ("How many times did you wake up?; using a 7-point Likert scale ranging from 1 = "Not at all" to 7 = "More than six times") and 9 ("How well did you sleep last night?"; using a 6-point Likert scale ranging from 1 = "very badly" to 6 = "Very well"). Lower scores on item six indicate high sleep quality and higher scores indicate poor sleep quality. Higher scores on item nine indicate high sleep quality and lower scores indicate poor sleep quality. The selected SMHS items will be directly translated from English to German. Ratings for both items will be used to see if memory performance for high to low reward items is correlated with the level of sleep quality experienced between the learning and testing phases of the sleep condition.

The Pittsburgh Sleep Quality Index (PSQI). The PSQI is another subjective measure of sleep quality, except participants are asked about their sleep habits and over the past month (e.g., "During the past month, how often have you had trouble sleeping because you cannot get to sleep within 30 minutes?")[70]. The PSQI consists of 18 items, which are clustered into seven component scores, which each range from 0-3 and are summed. Thus, PSQI scores can range from 0-21, where lower scores indicate poor sleep quality and higher scores indicate good sleep quality. The German version of the PSQI will be used in this experiment[82]. Like the SMHS scores on this scale will be used to see if memory performance for high to low reward items is correlated with participants general level of sleep quality experienced over the past month.

Psychomotor Vigilance Task (PVT). The PVT is a sustained attention task used to measure participants' objective vigilance[77]. We will use a 3-minute version of the Psychomotor Vigilance Task adapted from a 5-minute version of the task[83]. In this reaction time task, participants have to press the space bar as soon as a millisecond clock appears on the screen. The following measures will be analysed: median reaction speed (1/reaction time in ms) and percentage of lapses (number of lapses divided by the number of valid stimuli, excluding false starts; lapse = reaction time ≥ 500 ms). Reaction times shorter than 100 ms will be regarded as anticipated responses and treated as errors of commission. Participants will be instructed to respond as soon as a stimulus is shown on the screen. That is, they should shorten their reaction times as best they can but should not press the response button too early - this is a false start. Alongside the SSS, data collected from this task will be used to determine whether differences identified between sleep and wake conditions are due to that manipulation alone and not differences in subjective and objective vigilance, respectively.

Regensburger Wortflüssigkeits-Test. The Regensburger Wortflüssigkeits-Test[79], is a measure of verbal fluency. Participants are asked to type as many words as possible within a two-minute time window. They do this twice: once for words beginning with the letter p or m and once for words belonging to the category professions or hobbies. The order of the letter and category version will be balanced across participants. The different versions are used for the two experimental sessions respectively. The data will be used to determine differences in general retrieval performance between sleep and wake groups. The order of the cued letters and categories will be randomised for each participant between the retrieval sessions.

BIS/BAS scale. The BIS/BAS scale[73] is a measure of both the behavioral inhibition system (BIS) and behavioral activation system (BAS), both of which are related to motivation towards moving away from aversive outcomes and moving towards goal-oriented outcomes respectively. The scale consists of a total of 24 self-report items, 7 of the items are associated with the BIS component (e.g., "criticism or scolding hurts me quite a bit") and 13 of the items are associated with the BAS component (e.g., "I go out of my way to get things I want"). The other 4 items are fillers. For all of the item's participants respond using a 4-point Likert scale (1 = "very true for me" – 4 = "very false for me"). Higher scores on the BIS component of this scale indicate that an individual is more likely to experience negative feelings when pursuing a goal. The BAS component can be broken down further into three categories, the BAS responsiveness score, drive score and fun-seeking score. Generally higher scores on the BAS components indicated that an individual is more likely to seek out a goal because it is rewarding. This scale will be used to perform exploratory analyses to determine if there are any relationships between BIS and BAS scores on memory performance. The German version of this scale will be used in this experiment[84].

Becks Depression Inventory – Short Form (BDI - SF). The BDI – SF is a shortened version of the original BDI[85] containing only 13 items instead of 21. The BDI – SF is a measure of depressive symptoms, which are indicative of depression. For each item on this scale participants respond using a 4-point Likert scale (e.g., 1 = "I do not feel sad" - 4 = "I am so sad or unhappy I can't stand it"). Scores on the BDI-SF range from 0 – 39, where lower scores are indicative of fewer depressive symptoms. This scale will be directly translated from English into German. This scale will be used to determine whether there is a reduced effect of high rewards on memory after sleep for participants who report higher levels of depressive symptoms.

Hagen Matrices Test Short Version (HMT-S). The HMT-S[75] is an adapted shorter measure of the HMT which measures intelligence, specifically induction and fluid reasoning. In this task participants are required to identify patterns and rules in a series of puzzles. They are shown six 3 X 3 matrices of patterns that are incomplete and are missing one part. Participants must select the correct solution to the matrix from 8 potential options. In order to successfully complete the task participants must be able to identify the rules that govern the matrices they are shown. Each item is given a score of 1 if it is answered correctly and a score of 0 if the response is incorrect or missing. Thus, participants scores can range from 0 to a total of 6. Higher scores in this task indicate that an individual has greater induction and fluid reasoning abilities. Given that intelligence has been broadly related to the benefits of sleep on memory consolidation and the occurrence of neurophysiological activity that occurs during sleep[86] it is plausible that performance on this task might influence memory performance on the MLT. Therefore, we will conduct exploratory analyses to determine whether or not the benefits of sleep on reward related memory performance is associated with performance on the HMT-S.

Validation questions. To ensure that participants understand the Motivated Learning Task they are provided with information about how they will be rewarded gems or how they might lose gems as described previously (see Table 3). To test their understanding, they are asked the following questions: You see a landscape image that you already saw when you were scouting for treasures. The treasure chest belonging to it contains 750 gems. You correctly identify the picture as “old” - what happens?; You see a landscape image that you already saw when you were scouting for treasures. The treasure chest belonging to it contains 2150 gems. You make a mistake and identify the picture as “new” - what happens?; You see a landscape image that you did not see when you were scouting for treasures. The treasure chest belonging to it contains 2150 gems. You correctly identify the picture as "new" - what happens?; and finally you see a landscape image that you did not see when you were scouting for treasures. The treasure chest belonging to it contains 1450 gems. You make a mistake and identify the picture as “old” - what happens? The chances of getting all four validation questions that have four options to answer each correct by chance is $0.25^4 = 0.004$. On the occasions which participants will answer the validation questions they will be given two opportunities to correctly answer all of them. If they incorrectly answer at least one of the validation questions on their first attempt they will be given a second opportunity to answer them. If they incorrectly answer at least one of the questions on the second try they will be excluded from the experiment. If the participants answer all of the validation questions correctly on their first or second attempt they will be able to continue the experiment. On the second attempt participants are also shown the instructions for the motivated learning task a second time. Given that participants will complete the validation questions four times, the gems referred to in the questions will be adjusted each time.

Table 3. Reward contingencies for the Motivated Learning Task.

		Trial Type	
	Reward Contingencies	Target	Lure
Response	“Yes”	Hit (win n gems)*	False Alarm (lose 1100 gems)
	“No”	Miss (lose 1100 gems)	Correct Rejection (win 1100 gems)

*n refers to the number of gems which are associated with a given target image.

Wake experience. Participants will be asked to document their wake experience in the wakefulness condition. For this, they will be asked, “please provide a short description of your activity during each hour of the retention period”. To document their wake experience participants will be asked to recall and approximate their activity of each hour during the day by typing it into the relevant fields. Based on the received answers we will categorize the data and perform exploratory analyses.

Seriousness check. To ensure that participants performed the experiment seriously and did not engage in nefarious activities, such as repeat participations or masking of their true location via VPN we will perform a seriousness check at the end of the study. Participants will be asked, “It would be very helpful if you could tell us at this point whether you have taken part seriously, so that we can use your answers for our scientific analysis, or whether you were just clicking through to take a look at the survey? Please note that any answer that you provide to this question will not impact your chances of winning in the prize draw or prevent you from being added to the prize draw” and can respond with “I have taken part seriously” or “I have just clicked through, please throw my data away”. This approach has been shown to improve data quality in online studies[87]. This information will be used in exploratory analyses to reveal whether the seriousness of participant’s responses impacts the confirmatory analyses described below in the analysis plan.

Analysis Plan

The data simulations presented below and in Figure 1 were carried out using R (version 4.2.0) running in RStudio[88]. All analyses will be performed in Rstudio after the data have been collected.

Data Pre-processing.

Hit and False Alarm Rates. To compute the hit rate, the number of hits will be divided by the corresponding number of target trials. Comparatively, to compute the false alarm rate the corresponding number of false alarms will be divided by the number of lure trials. Hit and false alarm rates will be computed for all combinations of retention (sleep vs. wake), timepoint (immediate vs. delayed) and reward (50 vs. 750 vs. 1450 vs. 2150). This means that hit and false alarm rates are computed for each participant are collapsed across all durations for all levels of interest and duration conditions will be used to perform exploratory analyses. The main focus of analyses of the duration conditions will be to confirm that low memory strength items (those that were shown for the shortest time) benefit most from sleep-dependent consolidation, as has been reported before[7,89]. The duration conditions will also allow us to perform exploratory analyses that take into account differences in memory performance due to age or other demographic variance. Following the original paper which developed the MLT[30], the hit rate will be used as our main outcome variable. In this task, the hit rate is the most conceptually relevant outcome measure. This is because only targets are associated with a reward and the lures are not, since they are only shown in the test phase. Therefore, only the hit rate should be modulated by the reward, not the false alarm rate.

Discriminability. d′ will be computed from the hit and false alarm rates (ignoring the reward categories) for each participant as follows:

$$d' = z(HR) - z(FAR)$$

The measure d′ is an operationalization of discriminability (i.e., participants ability to discriminate between old and new images)[90]. d′ will be used to calculate participants general memory performance and exclude outliers as described above.

Model Specification and Hypothesis Testing

Our hypotheses (shown in Table 1 and Figure 1) will be tested and formalised using the R-package lmerTest[91] in the following linear mixed effects model with a maximal random effects structure as is recommended in the literature[92]:

```
hit rate ~ timepoint * retention * reward + ((timepoint + retention + rewa
rd) ^ 2 | subject)
```

This maximal linear mixed effects model includes all interactions and main effects as well as random intercepts and slopes for each participant for all parameters, with the exception of the three-way interaction where only one data point per participant exists, as the slope for that interaction and the random residual error would be indistinguishable. Deviation coding will be used for all categorical predictors in this model (See Table 4). Reward will be scaled such that a change in reward values reflects an increase of 1000 gems collapsed across duration categories. p-values produced by the lmerTest package using Satterthwaites degrees of freedom will be used to evaluate relevant parameters in this model (see below).

Table 4: Coding scheme.

predictor	-0.5	0.5
timepoint	immediate	delayed
retention	wake	sleep

Note. Deviation coding of predictors for the main analysis.

We will use the maximal model to give us an indication of whether our prediction that the magnitude of decline in memory for high vs. low rewarded images will be greater after a period of wake compared to a period of sleep at delayed recognition. This is represented in the timepoint × retention × reward parameter. If the timepoint × retention × reward is non-significant and an equivalence test suggests equivalence, we will conclude that there is no effect of reward on sleep-dependent memory consolidation. If the timepoint x reward parameter is non-significant and an equivalence test suggests equivalence, we will conclude that reward does not affect consolidation and reward effects are due to processes during encoding alone. If the timepoint x retention x reward is significant, we will followed it up with additional tests since the interaction could be taking place in any combination of those variables (for example at both immediate and delayed recognition). Therefore, we will reduce the model by the timepoint parameter and examine two linear mixed effects model for both immediate and delayed recognition with the following maximal effects structure:

```
hit rate ~ retention * reward + ((retention + reward) ^ 2 | subject)
```

First, we will examine the reduced model for immediate recognition to determine whether or not the retention × reward parameter is significant, which it would be if an interaction was present at immediate recognition. We do not expect that this will be a significant interaction since sleep is not expected to exert any impact on reward memory here as it has not yet occurred. However, we do expect that the reward parameter will be significant, where memory for high rewards will be greater than memory for low rewards at immediate recognition. If counter to this expectation the reward parameter at immediate recognition is not significant, this in combination with the timepoint × retention × reward of the full model would mean that rewards at encoding do not suffice to explain the memory enhancing effects of rewards.

Moreover, we will also examine whether the retention condition parameter is significant. If so, this would indicate that there is a time of day effect at immediate recognition between the sleep and wake conditions, which may occur due to learning either in the morning (in the wake condition) or the evening (in the sleep condition). For instance, one might expect that participants in the wake condition that learn in the morning perform better as they are well rested, whereas those in the sleep condition learn in the evening and are tired and as a result they perform worse. If that is the case then baseline hit rates will be added as a covariate to the reduced model of delayed recognition to determine whether they have an impact on our interpretation of the findings as they pertain to the impact of retention on reward.

To test our main hypotheses we examine the delayed recognition model and predict that the retention and reward parameters will be significant, where the sleep group collapsed across reward categories will have greater memory compared to the wake group and high reward images will be better remembered than low reward images collapsed across retention conditions, respectively **(H1 and H2; see Table 1).** Consequentially we will examine the retention × reward parameter to assess our final hypothesis, the interaction between retention and reward at delayed recognition **(H3: see Table 1)**. We will follow up this interaction by performing linear mixed models on each unit of reward between sleep and wake groups as follows:

```
hit rate ~ retention + (retention | subject)
```

We expect in this analysis that significant differences between sleep and wake groups at delayed recognition will emerge in the highest reward categories (1450 and 2150) and that performance at low reward categories (50 and 750) will be statistically equivalent. Simulated data demonstrating this pattern of results are shown in Figure 1. If memory performance at immediate recognition is included as a covariate and is significant and the interaction is not significant then it will be concluded that the covariate explains more variation in our data than the interaction between sleep and reward, and the data will be interpreted as such and explanations will be explored. However if the covariate is included and is significant or non-significant and the interaction term remains significant then we will conclude that after accounting for the variation explained by baseline scores, the retention × reward interaction persists. This pattern of results would replicate previous research indicating that sleep benefits memory performance over a period of wakefulness and that sleep exerts its influence on memory for high rewards only[44,49,50].

Evidence indicating that the parameters (including covariates) described above are unlikely under the null hypothesis will be determined via p-values < .020 and all follow-up tests will use the same threshold.

Resolving Model Convergence Issues. It is possible that our maximal models will not converge due to “overparamitization” within models containing all possible parameters, such as random intercepts and slopes[93]. Yet, it makes sense to start with the maximal model, since not including those parameters can yield an increased risk of Type I error[94]. We will perform 30,000 iterations of the maximal model. If the maximal model still fails to converge after 30,000 iterations or the estimated correlation parameters lie at 0 or +1, the data will be fitted using a zero correlation model. If after 30,000 iterations a model is not identified, random slopes per participant starting with the highest order components will be excluded until a model is identified. It is also possible that a model is identifiable, but overparamatization is indicated in a random-effects Principle Component Analysis (implemented using the rePCA() function in the lme4 R-package) [95].

Table 5: Simulation model parameters

parameter name	parameter value
fixed effects	
intercept	0.60
timepoint	-0.10
reward	0.015
timepoint:retention:reward	0.015
random effects	
participant intercept	0.12
participant timepoint sd	0.25
participant reward sd	0.02
error	0.12

Note. Model parameters used in the data-generating model. The parameter value for reward reflects the change in reward per 1000 gems.

If this occurs, components of the same order which have the smallest variance will be removed from the model. In this scenario, the fixed effects parameters will be evaluated using p-values calculated using lmerTest. To prevent p-hacking, p-values will only be calculated once a model with good convergence is identified. If either of the following scenarios occur it will be concluded that our model derived from the lmer package does not have good convergence: 1) the package is unable to converge on a final model and no output is produced; and 2) a model is produced but a singular fit is identified indicating that the model has been overfitted to the data.

Main effects estimated to be 0 or close to 0 will not be removed, thus ensuring that p-values derived from the identified model can be meaningfully interpreted and confidence intervals can be used in equivalence tests. Lastly, models that fail to converge will be documented and presented in the Supplement upon Stage 2 submission.

Data Simulation. Model convergence aside, linear mixed effects models also present a unique challenge for estimating effect sizes and analyses, since variance is shared and partitioned amongst all parameters. Therefore, there is no consensus as to how one should compute effect size estimates for main effects and interactions[96]. Our solution was to create a data generating model including parameters that reflect our hypotheses (see Table 5) and simulate data to calculate an appropriate sample size[97,98]. In this data generating model, the data produced are aggregated over duration so the hit rate is computed across each timepoint × reward × retention combination. The reward parameter was scaled such that a change in reward values reflects an increase of 1000 gems collapsed across duration categories. Thereby, we could set the hypothetical reward of zero, immediate recognition and the wake condition to baseline, which could then be compared to an increase of 1000 gems (reward effect), delayed recognition and the sleep condition, respectively. In other words, the main effect of reward reflects an increase in hit rate for every 1000 gems (per reward category). The Main effect of timepoint reflects the change in hit rate between immediate and delayed testing. Finally, the main effect of retention reflects the change in hit rate between the sleep and wake condition. Additionally, unlike the data analysis model shown above, this data generating model uses dummy coding and not deviation coding for each of the predictors (See Table 5).

To make our data generating model as realistic as possible we included by-subject random intercepts and slopes for main effects and interactions that are expected to be non-zero in addition to the residual error that is normally distributed. The residual error in this model was based upon pilot data which were collected in an online environment and can therefore approximate the error that we may encounter in this experiment using the same task. Moreover, for memory measures it is known that the measurements between immediate and delayed recognition will be correlated, so a high correlation between the time points was included (simulated correlation from the first run of the data simulation: $r = 0.87$). With that in mind we also assumed that memory would decay for each participant between immediate and delayed recognition so a reduction in memory performance is also included in the data simulation.

Table 5: Coding scheme.

predictor	0	1
timepoint	immediate	delayed
retention	wake	sleep

Note. Dummy coding of predictors for the data-generating model.

The parameters shown in Table 5 were also derived from our predictions described previously and the existing literature and data. For example, the parameter reflecting the impact of reward on memory in Table 5 was derived from the pilot study which was conducted to validate our task (see supplemental material; https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg). The estimated size of that effect was decreased slightly from that data to account for the fact that we will be using a more heterogenous sample. However, deriving parameters for the precise impact of sleep on memory for high rewards at delayed recognition proved challenging. A recent meta-analysis across different tasks indicated that the impact of sleep on memory at delayed recognition ranges between d = - .252 - 1.14 [8] in young and older adults. A meta-analysis on emotional memory found that the effect lies at d = .470[99] indicating that there is much variability in the size of the sleep effect on memory. This variability is likely increased by the heterogeneous sample we are collecting. The challenge is further complicated by effect size inflation in meta-analyses due to possible publication bias. However, in the face of uncertain effect sizes the goal of our simulation was not to determine a precise a priori effect size. Instead, our approach was to find a compromise between resource constraints and achieving 95% power to detect a broad range of effect sizes.

Therefore, under the reasonable assumption that the impact of sleep on memory and more specifically its impact on high reward at delayed recognition could also be smaller than is reported in the literature we chose parameters for the effect of sleep on reward at delayed recognition which yielded an overall small effect. This effect emerges at high rewards and is 0 at low rewards, because we modeled the data in such a way that there is no difference between wake and sleep for the lowest reward categories (as is shown in Figure 1). The simulation thus replicates previous research in line with our predictions,[44,49,50] but at a much smaller effect size than has previously been reported for both the impact of sleep on memory at delayed recognition and the timepoint × reward × retention interaction. Note that we do not include a parameter for simulating the effect of retention since we assume that the impact of sleep exerts itself on delayed recognition modulated by rewards (see Table 1 for further information). This is supported by the finding that cues bias reactivation for cued memories at the expense of non-cued memories[101]. Overall, the decisions used to create the model led to the chosen parameters shown in Table 5 and the data generating model below:

```
hit rate ~ timepoint + reward + timepoint:retention:reward + (timepoint + r
eward + timepoint:retention:reward | subject)
```

Power Analysis and Sample Size. To calibrate our power analysis to achieve at least β = .95 to detect our interactions of interest, we first simulated 1000 data sets using the parameters presented in Table 5 and the data generating model described above starting with a sample size of 1500 participants. However, since the maximal model does not always converge, for ease of simulation a simpler data analyzing model was used:

```
hit rate ~ timepoint:reward:retention + (1|sub_id)
```

Once the first 1000 simulations were complete the proportion of data sets, which yielded our significant interactions of interest at α > .020 was calculated. If the number of data sets which yielded our interaction of interest at α > .020 was lower than 95% the simulation was repeated and the number of participants included was increased by 50. We repeated that process until at least 95% power was achieved, which was the case at 1750 participants.

To establish the plausibility of our power analysis, in terms of identifying the timepoint × reward × retention interaction parameter shown in table 5, we calculated the mean 90% confidence intervals across all 1000 simulated data sets for our last simulated scenario of 1750 participants. This was carried out to rule out the possibility that the confidence intervals of that effect include zero as this would indicate that it is possible that no effect would be identified and would mean our predictions are implausible. This was not the case, the lower and upper 90% confidence intervals were .009 and .021 respectively. We also investigated the vulnerability of our analysis to Type I errors. This was achieved by simulating 1000 data sets using the identical parameters above shown in Table 5 with the exception of the timepoint × reward × retention interaction, which was set to a value of 0. In those simulated data sets 2.10% incorrectly identified an effect at $p < .020$.

Positive Controls. To ensure that the data we have collected are of sufficient quality for testing our hypotheses presented in Table 1, we will perform the following positive controls: 1) we will use a repeated measures t-test to confirm that memory between the lowest and highest reward categories is significantly different such that hit rate is greater for high rewards compared to low rewards in the delayed recognition test; 2) we will confirm that a retention interval of 12 hours yields a significant decline in memory performance between immediate and delayed testing by comparing d' (collapsed over the other conditions) between immediate and delayed testing using a repeated measures t-test; and, finally, 3) we will confirm that participants memory performance as measured using d' is significantly different from zero at delayed testing collapsed across all other conditions, which it should be if participants are capable of discriminating between targets and lures. In the event that one of these tests yields a statistically non-significant result (as determined using an alpha of p > .020) then equivalence tests[100] will be used and carried out against an equivalence bound of Cohens d = -0.10 - 0.10.

If any of the above analyses are found to be equivalent then it will be concluded that our data cannot be used to test our hypotheses.

Control Analyses. Equivalence tests[101] will be carried out to determine whether control variables across sleep and wake conditions are statistically equivalent, and therefore can be ruled out as variables that may otherwise explain differences in memory performance between those conditions. These tests will be carried out against an equivalence bound of Cohens d = -0.10 - 0.10, we consider effects within this range to be unlikely to influence the main analyses proposed in our analysis plan. Variables that are not equivalent will be considered in any interpretation of differences in memory performance between sleep and wake conditions and will be added as covariates to the model specified above to determine whether our initial interpretation of the model changes. Therefore, after evaluating our model without any covariates, the covariates will be added sequentially to determine the relative impact of each of them individually on our interpretation of the data. For example, if a given covariate explains a significant amount of variability in our data such that the remaining variance explained by our predictions is no longer significant, then it will be concluded that in our design the predicted effect is not detectable.

Equivalence tests will be used to compare SSS scores as well as median reaction speed and the percentage of lapses in the PVT during the learning phase and the retrieval phase and the number of words generated in the Regensburger Wortfluessigkeitstest during the retrieval phase between sleep and wake conditions. It is possible that the performance on the motivated learning task improves in the second relative to the first session. We will perform an equivalence test to determine whether there are differences between Experimental Sessions 1 and 2 in immediate recognition and if so session number will be included as a predictor in our model to determine whether our conclusions change. The remaining equivalence test that will be performed will be used to examine whether or not the hit rate for low reward items at delayed testing is statistically equivalent.

Since our sample might include individuals who have a sleep disorder (including insomnia, somnambulism, sleep apnea, REM sleep behavior disorder, narcolepsy or restless legs syndrome) or neurological disorder known to impact memory (specifically dementia, Alzheimer's disease and amnesia), we will include the presence or absence of these afflictions as predictors in an additional model to determine whether our conclusions change. If so, we will explore reasons why this may be.

Moreover, it is also unclear whether sleep benefits memories generally across low and high rewards, or more highly rewarded items benefit more at the cost of no sleep benefit for the lowest rewarded items. Therefore, we will conduct a repeated-measures t-test to compare the hit rate for the lowest reward category between the sleep and wake conditions at delayed testing. If that t-test is significant at $p < .020$, it will be concluded that sleep actively consolidates information which individuals are not highly motivated to learn even in competition to more highly rewarded information. If it is not significant, then equivalence tests will be conducted against an equivalence bound of Cohens $d = -0.10$ - 0.10 to conclude that the sleep effect for lowest reward category is not meaningfully higher than 0.

In addition, if the main analysis for H3 is not significant (i.e., we do not find a timepoint x retention x reward effect), we will conduct an analysis on a restricted sample. In this sample we will exclude all participants with mental disorders (including sleep disorders) and limit the age range to 20-39 years old. This will allow us to control whether the effect of rewards on sleep-dependent consolidation is only evident for young healthy adults.

Additionally, given that the original experiment which used the Motivated Learning Task found that the benefits of reward on memory were present at high levels of confidence, we will also perform a similar analysis on our data. This will enable us to determine whether the impact of sleep-based memory consolidation on reward information is modulated by confidence, a known proxy of memory strength[102]. Therefore, we will add the participants confidence ratings for correct responses on ‘old’ trials (i.e. hits) as a parameter to our original analysis and evaluate whether the reward by confidence interaction at delayed recognition is significant. Data at delayed recognition will be used since the original experiment identified the effect after a 24-hour retention interval. It is expected that, if the original study’s findings replicate, the magnitude of the benefit of reward on memory will be larger for high confidence responses compared to lower confidence responses. If it is the case that the interaction is significant then further exploratory analyses will be conducted to consider other variables of interest such as retention and timepoint, although we have no strong predictions at this point in time.

Finally, we will also perform an analysis using the sensitivity measure d’ to enable comparisons between our analyses and previous research on recognition memory (that did not manipulate reward values) and sleep dependent memory consolidation. To do this we will perform t-tests on the delayed recognition data as well as the difference score.

References

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

2. Feld, G. B. & Born, J. Neurochemical mechanisms for memory processing during sleep: Basic findings in humans and neuropsychiatric implications. *Neuropsychopharmacology* **45**, 31–44 (2020).

3. Stickgold, R. Sleep-dependent memory consolidation. *Nature* **437**, 1272–1278 (2005).

4. Barret, T. & Ekstrand, B. Effect of sleep on memory. III. Controlling for time of day effect. *Journal of Experimental Psychology* **96**, 321–327 (1972).

5. Brawn, T. P., Fenn, K. M., Nusbaum, H. C. & Margoliash, D. Consolidating the Effects of Waking and Sleep on Motor-Sequence Learning. *Journal of Neuroscience* **30**, 13977–13982 (2010).

6. Drosopoulos, S., Wagner, U. & Born, J. Sleep enhances explicit recollection in recognition memory. *Learning & Memory* **12**, 44–51 (2005).

7. Drosopoulos, S., Schulze, C., Fischer, S. & Born, J. Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General* **136**, 169 (2007).

8. Gui, W.-J. *et al.* Age-related differences in sleep-based memory consolidation: A meta-analysis. *Neuropsychologia* **97**, 46–55 (2017).

9. Jenkins, J. G. & Dallenbach, K. M. Obliviscence during Sleep and Waking. *The American Journal of Psychology* **35**, 605–612 (1924).

10. Tucker, M. A., Tang, S. X., Uzoh, A., Morgan, A. & Stickgold, R. To Sleep, to Strive, or Both: How Best to Optimize Memory. *PLOS ONE* **6**, e21737 (2011).

11. Van Ormer, E. B. Retention after intervals of sleep and waking. *Archives of Psychology* **137**, 49 (1932).

12. Wagner, U., Kashyap, N., Diekelmann, S. & Born, J. The impact of post-learning sleep vs. Wakefulness on recognition memory for faces with different facial expressions. *Neurobiology of Learning and Memory* **87**, 679–687 (2007).

13. Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A. & Stickgold, R. Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron* **35**, 205–211 (2002).

14. Walker, M. P., Stickgold, R., Alsop, D., Gaab, N. & Schlaug, G. Sleep-dependent motor memory plasticity in the human brain. *Neuroscience* **133**, 911–917 (2005).

15. Wilson, J. K., Baran, B., Pace-Schott, E. F., Ivry, R. B. & Spencer, R. M. Sleep modulates word-pair learning but not motor sequence learning in healthy older adults. *Neurobiology of aging* **33**, 991–1000 (2012).

16. Bailes, C., Caldwell, M., Wamsley, E. J. & Tucker, M. A. Does sleep protect memories against interference? A failure to replicate. *Plos one* **15**, e0220419 (2020).

17. Morgan, D. P., Tamminen, J., Seale-Carlisle, T. M. & Mickes, L. The impact of sleep on eyewitness identifications. *Royal Society Open Science* **6**, 170501 (2019).

18. Pöhlchen, D., Pawlizki, A., Gais, S. & Schönauer, M. Evidence against a large effect of sleep in protecting verbal memories from interference. *Journal of Sleep Research* e13042 (2020).

19. Rickard, T. C., Cai, D. J., Rieth, C. A., Jones, J. & Ard, M. C. Sleep does not enhance motor sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **34**, 834–842 (2008).

20. Jenkins, J. G., Dallenbach, K. M. Oblivescence during sleep and waking. *American Journal of Psychology* **32**, 605–612 (1924)

21. Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., Wixted, J. T. An opportunistic theory of cellular and systems consolidation. *Trends in Neurosciences* **34**, 504-14 (2011)

22. Dudai, Y. Karni, A., Born, J. The consolidation and transformation of memory. *Neuron* **88**, 20-32 (2015)

23. O'Neill, J., Pleydell-Bouverie, B., Dupret, D. & Csicsvari, J. Play it again: Reactivation of waking experience and memory. *Trends in Neurosciences* **33**, 220–229 (2010).

24. Carr, M. F., Jadhav, S. P., Frank, L. M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience* **2**, 147-53 (2011)

25. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).

26. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience* **12**, 1222–1223 (2009).

27. Marshall, L., Helgadóttir, H., Mölle, M. & Born, J. Boosting slow oscillations during sleep potentiates memory. *Nature* **444**, 610–613 (2006).

28. Ngo, H.-V. V., Martinetz, T., Born, J. & Mölle, M. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron* **78**, 545–553 (2013).

29. Ven, G. M. van de, Trouche, S., McNamara, C. G., Allen, K. & Dupret, D. Hippocampal Offline Reactivation Consolidates Recently Formed Cell Assembly Patterns during Sharp Wave-Ripples. *Neuron* **92**, 968–974 (2016).

30. Ackermann, S., Hartmann, F., Papassotiropoulos, A., Quervain, D. J. F. de & Rasch, B. No Associations between Interindividual Differences in Sleep Parameters and Episodic Memory Consolidation. *Sleep* **38**, 951–959 (2015).

31. Mantua, J. Sleep Physiology Correlations and Human Memory Consolidation: Where Do We Go From Here? *Sleep* **41**, (2018).

32. Feld, G. B., Weis, P. P. & Born, J. The limited capacity of sleep-dependent memory consolidation. *Frontiers in Psychology* **7**, 1368 (2016).

33. Feld, G. B. & Born, J. Sculpting memory during sleep: Concurrent consolidation and forgetting. *Current Opinion in Neurobiology* **44**, 20–27 (2017).

34. Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B. & Gabrieli, J. D. E. Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron* **50**, 507–517 (2006).

35. Braun, E. K., Wimmer, G. E. & Shohamy, D. Retroactive and graded prioritization of memory by reward. *Nature Communications* **9**, 4886 (2018).

36. Loftus, G. R. & Wickens, T. D. Effect of incentive on storage and retrieval processes. *Journal of Experimental Psychology* **85**, 141–147 (1970).

37. Madan, C. R., Fujiwara, E., Gerson, B. C. & Caplan, J. B. High reward makes items easier to remember, but harder to bind to a new temporal context. *Frontiers in Integrative Neuroscience* **6**, (2012).

38. Gruber, M. J., Ritchey, M., Wang, S.-F., Doss, M. K. & Ranganath, C. Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events. *Neuron* **89**, 1110–1120 (2016).

39. Weiner, B. & Walker, E. L. Motivational factors in short-term retention. *Journal of Experimental Psychology* **71**, 190–193 (1966).

40. Asfestani, M. A. *et al.* Consolidation of reward memory during sleep does not require dopaminergic activation. *Journal of Cognitive Neuroscience* 1–16 (2020).

41. Feld, G. B., Besedovsky, L., Kaida, K., Münte, T. F. & Born, J. Dopamine D2-like receptor activation wipes out preferential consolidation of high over low reward memories during human sleep. *Journal of Cognitive Neuroscience* **26**, 2310–2320 (2014).

42. Studte, S., Bridger, E. & Mecklinger, A. Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain and Language* **167**, 28–35 (2017).

43. Miendlarzewska, E. A., Bavelier, D. & Schwartz, S. Influence of reward motivation on human declarative memory. *Neuroscience & Biobehavioral Reviews* **61**, 156–176 (2016).

44. Igloi, K., Gaggioni, G., Sterpenich, V. & Schwartz, S. A nap to recap or how reward regulates hippocampal-prefrontal memory networks during daytime sleep in humans. *eLife* **4**, e07903 (2015).

45. Lisman, J. E. & Grace, A. A. The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron* **46**, 703–713 (2005).

46. Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS biology* **7**, (2009).

47. Oudiette, D., Antony, J. W., Creery, J. D. & Paller, K. A. The Role of Memory Reactivation during Wakefulness and Sleep in Determining Which Memories Endure. *Journal of Neuroscience* **33**, 6672–6678 (2013).

48. Baran, B., Daniels, D. & Spencer, R. M. C. Sleep-Dependent Consolidation of Value-Based Learning. *PLoS ONE* **8**, (2013).

49. Fischer, S. & Born, J. Anticipated reward enhances offline learning during sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **35**, 1586–1593 (2009).

50. Bennion, K. A., Payne, J. D. & Kensinger, E. A. The impact of napping on memory for future-relevant stimuli: Prioritization among multiple salience cues. *Behavioral neuroscience* **130**, 281 (2016).

51. Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E.-J. Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature neuroscience* **14**, 1105–1107 (2011).

52. Diekelmann, S., Wilhelm, I. & Born, J. The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews* **13**, 309–321 (2009).

53. Cordi, M. J. & Rasch, B. How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology* **67**, 1–7 (2020).

54. Button, K. S. *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376 (2013).

55. Lakens, D. & Evers, E. R. K. Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspectives on Psychological Science* **9**, (2014).

56. Davidson, P., Jönsson, P., Carlsson, I., & Pace-Schott, E. Does sleep selectively strengthen certain memories over others based on emotion and perceived future relevance?. *Nature and Science of Sleep* **13**, 1257-1306 (2021).

57. Berridge, K. C., Robinson, T. E. Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist* **71**, 670-679 (2016).

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

59. Hauner, K. K., Howard, J. D., Zelano, C., Gottfried, J. A. Stimulus-specific enhancement of fear extinction during slow-wave sleep. *Nature neuroscience*, **16**, 1553-1555 (2013).

60. Ashton, J. E., Cairney, S. A. Future-relevant memories are not selectively strengthened during sleep. *Plos One* **16**, e0258110 (2021).

61. Kroneisen, M., Kuepper-Tetzel, C. E. Using day and night–Scheduling retrieval practice and sleep. *Psychology Learning & Teaching* ***20***, 40-57 (2021).

62. Berres, S., Erdfelder, E. The sleep benefit in episodic memory: An integrative review and a meta-analysis. *Psychological Bulletin*, **147**, 1309-1353 (2021).

63. Jacobi, F., Höfler, M., Siegert, J., Mack, S., Gerschler, A., Scholl, L., ... Wittchen, H. U. Twelve-month prevalence, comorbidity and correlates of mental disorders in Germany: the Mental Health Module of the German Health Interview and Examination Survey for Adults (DEGS1-MH). *International Journal of Methods in Psychiatric Research*, **23**, 304-319 (2014).

64. Lipinska, G., Stuart, B., Thomas, K. G., Baldwin, D. S., Bolinger, E. Preferential consolidation of emotional memory during sleep: a meta-analysis. *Frontiers in Psychology* **10**, 1014 (2019).

65. Schmid, D., Erlacher, D., Klostermann, A., Kredel, R., Hossner, E. J. Sleep-dependent motor memory consolidation in healthy adults: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, **118**, 270-281 (2020).

66. Ellis, B. W. *et al.* The St. Mary's Hospital Sleep Questionnaire: A Study of Reliability. *Sleep* **4**, 93–97 (1981).

67. Johns, M. W. A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep* **14**, 540–545 (1991).

68. Adan, A. & Almirall, H. Horne & Östberg morningness-eveningness questionnaire: A reduced scale. *Personality and Individual Differences* **12**, 241–253 (1991).

69. Horne, J. A. & Ostberg, O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiol* **4**, 97–110 (1976).

70. Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R. & Kupfer, D. J. The pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research* **28**, 193–213 (1989).

71. Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R. & Grant, M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* **88**, 791–804 (1993).

72. Irons, J. G. Bassett, D. T. Prendergast, C. O. Landrum, R. E. & Heinz, A. J. Development and initial validation of the caffeine consumption questionnaire-revised. *Journal of Caffeine Research*, **6**, 20-25 (2016).

73. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of Personality and Social Psychology*, **67**, 319-333 (1994).

74. Volk, R. J. Pace, T. M. & Parchman, M. L. Screening for depression in primary care patients: Dimensionality of the Short Form of the Beck Depression Inventory. *Psychological Assessment*, *5*, 173 -181 (1993).

75. Heydasch, T. Haubrich, J. & Renner, K.-H. Die Kurzform des Hagener Matrizen-Tests (HMT-S). Ein 6-Item Intelligenztest zum schlussfolgernden Denken. *Methoden, Daten, Analysen,* **7**, 183–208 (2013).

76. Hoddes, E., Zarcone, V., Smythe, H., Phillips, R. & Dement, W. C. Quantification of sleepiness: A new approach. *Psychophysiology* **10**, 431–436 (1973).

77. Dinges, D. I, & Powell, J. W. Microcomputer analysis of performance on a portable, simple visual RT task sustained operations. *Behavior Research Methods, Instrumentation, and Computers*, **17**, 652–655 (1985).

78. Lahl, O., Wispel, C., Willigens, B., Pietrowsky, R. An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, **17**, 3-10 (2008).

79. Aschenbrenner, S., Tucha, O. & Lange, K. W. *Regensburger wortflüssigkeits-test: RWT*. (Hogrefe, Verlag für Psychologie, 2000).

80. Bloch, K. E., Schoch, O. D., Zhang, J. & Russi, E. W. German version of the epworth sleepiness scale. *Respiration* **66**, 440–447 (1999).

81. Randler, C. German version of the reduced morningness–eveningness questionnaire (rMEQ). *Biological Rhythm Research* **44**, 730–736 (2013).

82. Backhaus J, Riemann D. Schlafstörungen bewältigen [Coping with sleep disorders]. Weinheim: Beltz Psychologie Verlags Union; 1996.

83. Roach, G. D., Dawson, D. & Lamond, N. Can a Shorter Psychomotor Vigilance Task Be Used as a Reasonable Substitute for the Ten-Minute Psychomotor Vigilance Task? *Chronobiology International* **23**, 1379–1387 (2006).

84. Strobel, A. Beauducel, A. Debener, S. & Brocke, B. (2001). Eine deutschsprachige version des BIS/BAS-Fragebogens von carver und white. *Zeitschrift für Differentielle und Diagnostische Psychologie*, **22**, 216-227 (2001).

85. Beck, A. T. Ward, C. H. Mendelson, M, Mock, J. & Erbaugh J. An inventory for measuring depression. *Archives of General Psychiatry*. **4**, 561–71 (1961).

86. Ujma, P. P. Sleep spindles and general cognitive ability–A meta-analysis. *Sleep Spindles & Cortical Up States*, **2**, 1-17 (2021).

87. Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods* **45**, 527-535 (2013).

88. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2019).

89. Payne, J. D., Tucker, M. A., Ellenbogen, J. M., Wamsley, E. J., Walker, M. P., Schacter, D. L., Stickgold, R. Memory for semantically related and unrelated declarative information: the benefit of sleep, the cost of wake. *PloS One* **7**, e33079 (2012).

90. Macmillan, N. A. & Creelman, C. D. *Detection theory: A user's guide, 2nd ed*. (Lawrence Erlbaum Associates Publishers, 2005).

91. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* **82**, 1–26 (2017).

92. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278 (2013).

93. Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious mixed models. (2015). *arXiv preprint arXiv:1506.04967* (2015).

94. Heisig, J. P. & Schaeffer, M. Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review* **35**, 258–279 (2019).

95. Bates, D. Mächler, M. Bolker, B. & Walker, S. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**, 1–48 (2015).

96. Rights, &. S., Jason D. Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological Methods* **24**, 309–338 (2019).

97. Brysbaert, M. & Stevens, M. Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition* **1**, 9 (2018).

98. Green, P. & MacLeod, C. J. SIMR: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* **7**, 493–498 (2016).

99. Schäfer, S. K. *et al.* Sleep's impact on emotional memory: A meta-analysis of whole-night, nap, and REM sleep effects. *bioRxiv* (2019).

100. Sterpenich, V., van Schie, M. K., Catsiyannis, M., Ramyead, A., Perrig, S., Yang, H. D., ... Schwartz, S. Reward biases spontaneous neural reactivation during sleep. *Nature Communications*, **12**, 1-11 (2021).

101. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* **1**, 259–269 (2018).

102. Wixted, J. T. & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, **18**, 10-65 (2017).

Data availability

The authors commit to sharing the raw data and materials on acceptance of our Stage 2 manuscript.

Code availability

The authors commit to sharing all code on acceptance of our Stage 2 manuscript.

Acknowledgements

This work was supported by a grant from the German Research Foundation to GBF (Deutsche Forschungsgemeinschaft GZ: FE 1617/2-1). The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

The authors declare no competing interests.