Rebuttal

Below you can find the review text in black and our response in blue. Quotes from the manuscript are in parentheses.

Editor's remarks

I have now obtained two constructive and helpful reviews of your Stage 1 submission. Overall, the reviews are encouraging and the general consensus is that your proposal is a promising contender for eventual IPA. There are, however, a range of issues to address that cut across several of the Stage 1 criteria, including clarifications concerning the hypotheses (and contingencies between them), exclusion criteria and analysis plans, justification of including design features, and inclusion of positive controls. Another substantial concern raised is whether there is tension in your proposal between optimising it for unstated exploratory analyses vs. addressing the main confirmatory hypotheses. A Stage 1 RR should favour the latter as much as possible, but I am open to further discussion of this issue in your response in order to ideally reach a consensus.

On a purely technical note, please can you ensure that the main manuscript includes a direct URL to the Supplementary Information on the OSF so that it is easily accessible to the reviewers. When we invite to reviewers to evaluate RRs, we include only the link to the manuscript (rather than generic links to the parent OSF project), which avoids any risk of reviewers inadvertently assessing the wrong document or version. But because we only send the one direct URL to the manuscript, it is essential that the manuscript itself includes direct links to any such materials within the OSF project that are important for the review process.

We apologise for not providing an appropriate way of accessing the supplemental material and have now provided links throughout the manuscript to the supplemental information where relevant as well as below:

https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg

The password is as follows:

2B3DCkToT1Hb

On the basis of these reviews I'm happy to invite a Major Revision and response, which will be returned to both of the reviewers for a further look.

We are pleased with this outcome and are very grateful for yours and the reviewer's helpful comments which we have addressed below:

Reviewer 1

Morgan et al propose a registered report to assess the role of reward on sleep-dependent memory consolidation. Several studies have been conducted in this area with mixed results. They propose to conduct a large-scale online study to potentially reconcile these discrepant results, given several previous studies have had a relatively low N and therefore may have been underpowered. Overall, I am impressed with the level of care and detail that has gone into planning and presenting this Stage 1 RR and think the study could make a valuable contribution to the literature.

We are very pleased that the reviewer also believes that this stage 1 submission would make a valuable contribution to the literature. We found many of their comments extremely helpful and have provided responses below.

1A. The scientific validity of the research question(s).

The research question is scientifically valid. There is a lot of evidence in favour of sleep-dependent consolidation (i.e., an active consolidation process during sleep). Given the number of memory traces encoded during the day, it is likely that some process of selectivity is needed to consolidation specific traces. One possible driver of selective consolidation is reward at the point of encoding, and there is some understanding of the neurobiological basis for this relationship, involving dopamine and the VTA and hippocampus.

We thank the reviewer for appreciating the validity of our research question and for taking the time to understand the intricacies that are involved in examining the relationship between sleep-based consolidation and reward memory.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

The logic and rationale of the proposed hypotheses are appropriate. Given past research, both hypotheses are plausible. The authors present two clear hypotheses: (1) memory performance will be greater following a period of sleep than wake (i.e., a standard "sleep effect"), and (2) that the sleep effect will be greater for high than low reward stimuli. This latter hypothesis is the critical hypothesis related to reward and sleep-dependent consolidation. The former isn't presented as an control analysis to ensure the sleep manipulation has worked, but could be interpreted as such (see comments to criteria 1D).

1. Is a significant sleep effect in H1 critical to assessing H2? How will H2 be interpreted in the absence of an effect for H1?

The reviewer raises a valid point. In general, we would argue that a significant sleep effect in H1 is not necessary to assess H2, but in the absence of that effect our interpretations would indeed change. Take the example presented in the figure below which presents the most likely flavour of this hypothetical scenario (note that the figure depicts change scores for simplicity). In that figure it is clear that a significant effect of sleep at delayed testing is not present but a significant interaction between sleep and reward is. In this scenario, we can see that collapsed across rewards there is no difference in memory performance between the sleep and wake conditions. One might be inclined to conclude here that there is no evidence that sleep is a preferential state for consolidation, when compared to an equivalent period of wake, because there is not a main effect of sleep. However theoretical and empirical evidence would not support that conclusion. Rather, such an interaction in the absence of a significant main effect of sleep still indicates that sleep actively consolidates high reward and in addition increases forgetting of low reward items. Two theoretical frameworks in the sleep and memory literature would support such a conclusion.

The first, which we refer to in the manuscript on p. 3, lines 48 - 60 is the active systems consolidation hypothesis, where memory consolidation occurs preferentially during sleep as a result of spontaneous reactivation of memories. The second is the Synaptic Homeostasis Hypothesis, which proposes that during sleep synapses that were potentiated during wake learning are renormalized such that synapses, which are not potentiated enough are downscaled so much during sleep that the result is increased forgetting of that information across sleep (Tononi & Cirelli, 2014, Feld & Born, 2017). Ultimately, this would lead to less forgetting or increased consolidation of highly rewarded items,

greater forgetting of lowly rewarded items and a net equivalence of memories collapsed across reward between sleep and wake conditions. Overall this would be our interpretation of the data if it were the case that H1 was not significant but H3 (note we have added H2 and the old H2 has become H3) was. Fortunately, The structure of our experiment allows us to determine whether this could be the case empirically.

To account for this in our registered report we have now added a contingency in the interaction hypothesis for the case that H1 is not significant but H2 is in our design table on p. 10. We added the following:

“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.”

It should be noted that since this is not our pre-registered prediction, but is one that is plausible, we have provided a description of it in the supplemental information (in the file named “Morgan_PCiRRAlternativeHypothesis.pdf”) to support pursuing an assessment of H3 in the absence of a significant H1:

“Thus far we have proposed that sleep exerts its influence on memory as a function of reward, as shown in Figure 1 in the manuscript. This would yield a significant main effect of sleep (H1) and a significant interaction between sleep and reward (H3). However, one possibility is that there is not a main effect of sleep on memory, but an interaction between sleep and reward persists as shown in Figure 1 below. The absence of a main effect of sleep on memory does not preclude an impact of sleep on consolidation. Rather two sleep related processes, namely active systems consolidation and synaptic homeostasis may act together to yield a net null main effect of sleep. That is, reward modulates the benefit from reactivation during sleep and additional unspecific downscaling opposes this process and leads to generally lower performance. Ultimately this would lead to less forgetting or increased consolidation of more highly rewarded items, greater forgetting of more lowly rewarded items and a net equivalence of memories collapsed across reward between sleep and wake conditions. Overall this would be our interpretation of the data if it were the case that H1 was equivalent but H3 was significant and the structure of our experiment allows us to determine whether this could be the case empirically.”

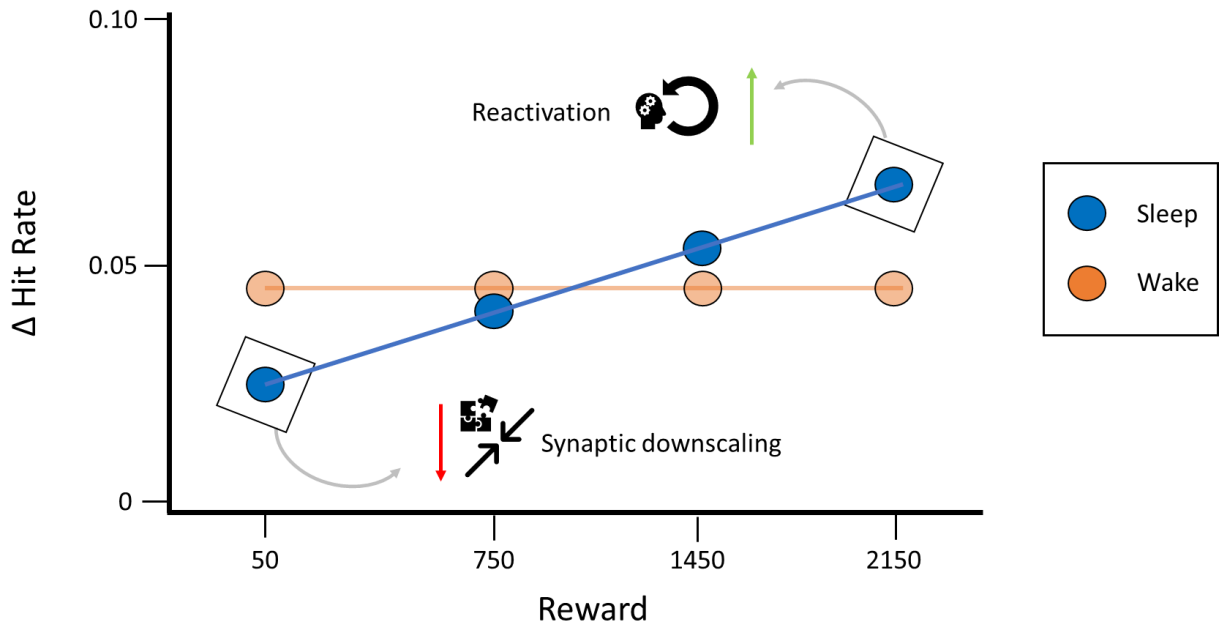

Figure 1. Hypothetical scenario depicting no main effect of retention (i.e. sleep vs. wake) at delayed testing, but an interaction between retention and reward.

2. In relation to H2, you appear to be predicting no sleep effect for low reward items (and an effect for high-reward). Given the ample evidence for sleep effects in studies that don't include a reward I wasn't sure about this. Is it not more likely that a sleep effect will be seen for both low and high reward, but that it is greater in the high reward condition?

Indeed, the reviewer is correct that there is much evidence demonstrating the benefits of sleep on memory in tasks, which do not explicitly tie memory performance to rewards. This may lead one to predict that the benefits of sleep would be observed across all reward levels, including low rewards. However, it is our interpretation of these experiments that every encoding procedure (be it explicit or implicit) engages motivational brain processes that may be driving consolidation. This is most prominently shown by virtually all rodent experiments in the sleep and memory field employing food deprivation or highly appetitive foods to motivate the animals during learning. In the few instances where this is not done, novelty or punishment is used as motivation. Humans are likewise motivated to learn, e.g., by extrinsic factors to receive money or by intrinsic factors motivating them to perform well as a participant. Therefore, it seems that those experiments actually do not reflect consolidation without motivation.

Moreover, although TMR is generally believed to enhance memory consolidation, research in rats has shown that it actually biases replay at the expense of the non-cued information (Bendor & Wilson, 2012. In humans, it was shown that reward biases replay towards the rewarded task, too (Sterpenich et al., 2021). Therefore, even if one assumes that memories can be learned and consolidated without motivation, it is our prediction that the competition for replay in our task will lead to lowly rewarded memories not benefitting from sleep. Luckily, our experiment will be able to speak to this question empirically. To do so we will additionally conduct a t-test comparing memory for low rewards at delayed testing between sleep and wake conditions. We will also conduct an equivalence test to determine whether they are statistically equivalent if the aforementioned t-test is not significant at $p < .020$. If there is a significant difference in memory for low rewards then we will conclude that sleep also benefits memory at low rewards.

We have added the following on p. 43, lines 861 – 865 to the manuscript in the data simulation section to justify our prediction that memory for low rewards will be the same:

"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]."

And we added the aforementioned analysis to test for this to our control analyses on p. 46, lines 936 – 944:

"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 motivated to learn. 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."

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

The methodology and analysis pipeline are appropriate, however I have several comments/questions in relation to this:

1. Overall the methodology and analysis pipeline are clear and explained in great detail. However, that detail came at the expense of clarity to me. I don't want to increase the length of the manuscript more than is necessary, or remove any of the detail (which is needed), however I wonder whether a summary/overview is needed at the beginning of the methods, or a reordering of the methods might help. For instance, I wanted more info about the actual experimental task (Figure 3 and 4 could have been presented earlier) and overall experimental design before I then tackled the detail.

We have now moved Figures 3 and 4 to an earlier place in the manuscript and have added some additional information to Figure 3 so that the reader has more detail about the experiment earlier on. Note that Figure 3 is now Figure 4 since the order has changed in the manuscript.

In Figure 3's legend the Motivated Learning task is now explained in more detail p. 17, lines 286-292:

"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)."

and Figure 4's legend now explains the following p. 19, lines 2994-297:

"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. "

2. Data collection and demographically diverse sampling is ambitious (high N, 8 month data collection window, wide range of individuals). I wonder whether the authors have any backup plans in terms of ensuring they reach their target N (e.g., if they can't collection enough data in one demographic area, will they sacrifice this aspect of the study to ensure they reach the target N, or will they sacrifice N to ensure a representative sample)?

The reviewer raises an important point. Of note, this online experiment will allow us to sample from the entire German population of 83 million, which is much more than can be reached by recruiting in even a large city. Our stratification gives us access to at least 40 million individuals who are eligible to take part in this experiment. Nonetheless, we have lengthened our recruitment phase to 12 months.

We will also be engaging in a diverse advertisement campaign using and paying for Meta Advertisements, which will enable a faster paced recruitment process compared to only asking people to share the experiment online. Also, we will use our well-established contacts to national media to further advertise the experiment to the public. We have now made this clearer in the manuscript on p. 11 lines 208 – 210:

"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."

Additionally, we have now added a "refer a friend" option to participants, where they are guaranteed a small voucher if they recruit one of their friends to participate and that friend successfully completes the experiment with the intention of maximising our recruitment efforts. This has now been added to the manuscript on p. 11, lines 211 – 213:

"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."

Additionally, we have also added the following contingencies to ensure our data collection is successful:

"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 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."

Importantly, we do not expect to have to implement these, however it is a precautionary plan which will certainly ensure we complete data collection in the unlikely event that we encounter recruitment difficulties. We have now included the above contingency planning to the manuscript on p. 13, lines 246-254.

Finally, we have now removed the 60-69 age category from our sampling plan, which will make it easier to collect the whole sample from the beginning.

3. Participants will be told not to take a nap in the wake condition. Is this potentially problematic to individuals who do typically nap during the day (e.g., older populations)?

In response to the reviewers comments above and also a comment by reviewer 2, we have now removed the oldest age group that this is most likely to impact (i.e. 60 – 69 year olds) and therefore participants will still be asked to not nap during the day. This restriction is routinely applied in sleep research since even ultra-short daytime naps may consolidate memories (Lahl et al., 2008). Not restricting napping would therefore threaten the validity of our interpretation in absence of sleep effects. We have added this rational to the manuscript on p. 21, lines 351-354.

"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]."

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

There is appropriate methodological detail and the analysis pipeline is appropriately explained. Again, I have a few comments/questions though:

Methods:

1. The N reported in the abstract is different from that reported in the main text, so please correct this.
   We apologize for this oversight due to editing and have now corrected the N reported in the abstract to match the N reported in the main text.

2. The two sessions will be conducted a maximum of 4 weeks apart. What is the minimum gap between sessions?
   There will be a minimum gap of 1 one week between sessions and we added this in the manuscript on page 19, lines 303-304:

"Participants will complete the sessions separated by a minimum of 1 week and a maximum of 4 weeks"

3. On p. 18 you describe the total number of stimuli etc., but I couldn't see information on the number of encoding trials per item. I presume it is one?

Yes, that is correct, the number of encoding trials per item is one. To clarify this we have now written the following on p. 23, lines 401 - 403:

"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]."

4. On p. 28 you explain the 4 validation questions. It isn't clear to me whether inclusion only occurs if all 4 questions are answered correctly, or exclusion occurs if all 4 questions are answered incorrectly.

   We apologize for being unclear. Participants are given at most two opportunities in one session to correctly answer all four of the validation questions. If they answer them correctly on their first attempt then they can continue the experiment. If they answer at least one of the questions incorrectly they will be asked to answer the same four questions a second time. If they correctly answer all four questions on the second attempt then they can continue with the experiment. Finally, if they answer at least one of the questions incorrectly again on the second attempt then they are excluded from the experiment. This approach was used successfully in our previous experiments and led to only a small amount of exclusions (16% of 550 participants failed this validation procedure).

   To make this clearer we have amended the "validation questions" section of the manuscript, which is now on p. 32 - 33 , lines 631 – 653 and have added an additional table (Table 2) in the manuscript, where we list all of our inclusion and exclusion criteria.

   "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."

5. On p. 29 you explain the "seriousness check". I presume during this you still make clear that their answers to these questions will not affect their inclusion in the prize draw, otherwise they have a stake in saying they were being serious.

   We agree that participants should not be under the impression that they could be disadvantaged if they provide an honest answer. Therefore, we have clarified this on p. 34, lines 662 – 673:

   "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"

Analysis:

1. On p. 29 you explain exclusion criteria in relation to d'. I think this is d' collapsed across all conditions (which you do mention later), but I think this should be made explicit here.

   We agree with the reviewers comment and have made this explicit in Table 2 detailing inclusion and exclusion criteria, which has replaced the paragraph on exclusion criteria, p 13:

   "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"

2. In relation to both H1 and H2 the hypothesis is directional, but the interaction analysis isn't. You state you will follow up significant interactions with posthoc tests to appropriately characterise the interactions, however there is a slight gap in relation to the interpretations. For example, in Table 1 you state "If there is no difference in memory performance between the sleep and wake groups…". What if a difference is seen but it is the opposite to that predicted for H1? If this occurred, I don't think you have made clear what the conclusion would be. This probably just requires a slight tightening of the wording in relation to the last two columns of the table to ensure you have covered every statistical eventuality.

We apologise for our lack of clarity in the manuscript on Table 1. As the reviewer has proposed we have tightened the wording in Table 1 under the section "Theory that could be shown wrong by the outcomes" presenting an explanation of how the data would be interpreted if the opposite of our predictions where true. Specifically, in response to the reviewers comment on H1 we made the following addition to Table 1:

"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."

3. On p. 31 you describe the GLMMs, and that they will include "all interactions, main effects and random slopes for each participant for all parameters". What about intercepts in this model?

We will be including random intercepts and this is implicitly written in the model below, shown on p. 36, lines 707 - 708:

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

This was missing in our description of the model so we have adjusted the sentence p. 36, lines 709-712-7

"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."

4. On p. 31 you say “reward will be a mean-centred continuous predictor”. I presume this is a linear predictor, based on the (mean-centred) raw reward values, but it should perhaps be made clear.

Thank you for spotting this. This parameter will be scaled in the data analysis such that each step change in reward reflects an increase of 1000 gems. We have now also amended the statement which the reviewer has quoted to the following p. 36, lines 713-715:

“Reward will be scaled such that a change in reward values reflects an increase of 1000 gems collapsed across duration categories.”

Likewise the reward parameter was scaled in the data simulations the same way and so we have also amended this on p. 41, lines 812 - 813:

“The reward parameter was scaled such that a change in reward values reflects an increase of 1000 gems collapsed across duration categories.”

We have also provided further clarification around what this means in terms of the interpretation of the data simulation on p. 41, lines 816 - 819:

“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.”

5. On p. 32 & 38 you discuss adding covariates to the analysis (e.g., memory performance at immediate test). It isn’t clear whether these will be added to the primary analysis, or whether the primary analysis will be conducted without these covariates and you will then run a second analysis including the covariates.

We will first run the primary analyses as they are shown in the analysis strategy without the covariates, as to give the reader the chance to appreciate the results without covariates. Then, in the case that control variables are not equivalent between the conditions as determined by equivalence tests, we will run additional analyses including those control variables as covariates in the primary analyses and interpret the results structure appropriately. We have added this information to that paragraph on p. 45 lines 911 - 919:

“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.”

6. Why is the p-value set at p<.02?

The p-value is set to .020 as this is the minimum requirement for the journal Cortex for registered reports and we consider Cortex a potential overlay journal for our PCI RR.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

I do not think the authors have presented a positive control to ensure the data are of sufficient quality to assess their main hypothesis. There are multiple checks of data quality, which will help in relation to this. One potential possibility is to only test H2 conditional on H1 being true (i.e., a significant sleep effect, regardless of reward, is present). However, this might be too limiting. Another possibility is to assess a reward effect at immediate test. There might be other possibilities that I have missed.

We agree that it is important for us to include positive controls in our analysis. We also agree that it would be too limiting to test H3 conditional on H1 being correct since it is plausible that an interaction that points to sleep preferentially consolidating high rewards over low rewards may still be present as was previously discussed above on p. 2 of this rebuttal. However, we have added positive controls to the manuscript on p. 44 - 45, lines 891 – 905. Note, that we will assess the reward effect at delayed recognition (instead of immediate recognition), as it is possible that the reward effect in this task only develops over the retention interval, as has been previously discussed (Adcock et al. 2006; Feld et al., 2014).

"**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."

Further comments:

1. The first sentence of the abstract is overly complex and could be simplified for clarity.

We have now rewritten the intro of the abstract to make it more accessible to the reader on p. 2, lines 23 - 25 to the following:

"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 sample sizes (among other factors) likely drive these inconsistencies."

2. The mention of psychiatric disorders in the abstract does not seem appropriate to me.

We have decided to keep the point about psychiatric disorders in the abstract since this is a key practical implication of our research and reviewer 2 has even asked us to expand on this topic in the manuscript. In response to reviewer 2's comment no. 20 on we have added some details into the introduction that help justify that sentence in the abstract on p. 6 – 7, lines 139 - 150:

"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."

3. In the introduction there is quite a bit of discussion about dopamine and the VTA. Although interesting and relevant I wonder whether this could be reduce somewhat. There could also be better signposting on which studies in the introduction are human vs non-human research.

Throughout the introduction we have now pointed out, which study has been conducted with humans and which study has been conducted with non-human subjects. We have also shortened the paragraph about the neurophysiological underpinnings and added a rational, why we think it is relevant to the introduction of our manuscript on p. 4, lines 83 – 89:

"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."

Reviewer 2

Summary

Memory consolidation is supported by sleep and recent theoretical approaches have suggested that sleep might work in a selective manner to strengthen information that is most relevant or salient to the individual. However, empirical evidence related to this selective memory benefit of sleep is fairly mixed, potentially due to a diversity of experimental methods and low statistical power. In this Registered Report, Morgan and colleagues plan to carry out a timely and well powered investigation of the effects of sleep on memories associated with high and low rewards. They will use an online experiment with a gamified task to determine whether a benefit of sleep for recognition memory is amplified for information associated with high relative to low reward. My overall assessment of the manuscript is positive, but I have a number of comments and queries, which are separated according to the subheadings of the reviewer guidance notes. I have also included some more general comments at the end of my review.

We are pleased that the reviewer is positive about our stage 1 submission and we thank them for their helpful and detailed feedback. We have provided our responses below in blue.

Please note that I was unable to retrieve the authors supplementary materials, which were referenced several times in the main manuscript. I will need to see this in the resubmission before providing another opinion on the suitability of the manuscript for stage one acceptance.

We apologise for this. We have now provided a direct link to the supplemental in our manuscript so that the reviewer can determine the suitability of this manuscript. The supplemental can also be accessed from here using the following link:

https://cloud.zi-mannheim.de/index.php/s/jDnY35CM4WMdQCg

The password is as follows:

2B3DCkToT1Hb

The scientific validity of the research question(s)

1) From the overview of the literature in the Introduction, it appears that a potential source of inconsistency in previous findings relates to the memory system in question (e.g. declarative vs procedural memory). Is it the case that studies showing a positive effect of sleep on high vs low reward memories are in the procedural domain? Can the authors be sure that a declarative memory protocol is the optimal approach for studying effects of sleep and reward on memory consolidation?

We agree that the distinction between the impact of sleep and reward on procedural vs. declarative memory is important to consider when deciding on the optimal approach for our research questions and especially when discussing the results of our experiment. Generally speaking there is strong evidence that indicates at least a small to medium effect of sleep on declarative and procedural memory (for meta-analyses see Berres & Erdfelder, 2021; Lipsinka et al., 2019; Schmidt et al., 2020). To our knowledge there is also only one study that has examined the impact of sleep on rewarded memories using a procedural memory procedure (Fischer & Born, 2009). The majority of research has been performed using declarative (or hippocampus dependent) tasks with some tasks finding the effect and others not (Baran et al., 2013; Bennion et al., 2016; Fischer & Born, 2009; Igloi et al., 2015; Oudiette et al., 2013). So it cannot be concluded that the tasks finding the modulating effect of rewards on sleep-dependent memory consolidation are from the procedural domain and since there are more studies in the declarative domain and the hippocampus has been strongly implicated for the sleep effect on memory we focus on this for now. We now give the rational for this decision in the design section in the manuscript on p. 16, lines 269 – 273:

"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."

We also added a note to the design table that our interpretations are limited to memory from the declarative domain.

"Table 1. Design table. Of note, 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."

2) A more general point relates to the issue of whether reward actually has any impact of consolidation, irrespective of whether it occurs over sleep or wakefulness. In their Introduction, the authors state: "Either sleep selectively consolidates information associated with high rewards or reward related processes during encoding together with sleep-independent consolidation processes

initiated shortly after learning are sufficient to enhance reward memory". A third possibility is that the benefits of reward occur only at encoding and are not enhanced at all by consolidation. Can the authors rule this out?

This is an important point! Our design is specifically designed to be able to rule this out, since we collect data about the effect of reward immediately after encoding (immediate recognition task) and after a retention interval (delayed recognition task). Our model assumes a time point x retention x reward interaction, as is indicated by our hypothesis (i.e., that rewards specifically affect consolidation during sleep retention and not during wake retention). The model therefore assumes that the slope of the reward effect remains the same for immediate and delayed recognition in the wake retention condition (as can be seen in Figure 1). If the slope of the reward effect remains the same between the two assessments of memory performance also in the sleep condition, this would lead to a non-significant time point x retention x reward interaction and therefore indicate that rewards do not affect sleep-dependent memory consolidation. Then, if the time point x reward interaction is also not significant, this would mean that rewards also do not affect memory consolidation in a time-dependent fashion. We have made this clearer and have added the following to our introduction on p. 5, lines 111 - 116:

"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."

And analysis strategy on p. 11lines 208 -213:

"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 retention 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 × retention × reward interaction is significant, we will follow 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)."

The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

3) The use of a recognition paradigm raises questions about the plausibility of the authors proposed hypotheses. A number of previous studies have shown that recognition paradigms are fairly insensitive to sleep-memory effects, and this lack of sensitivity could pose problems for the authors predictions. Relatedly, H1 focuses on performance in the delayed test, whereas sleep-memory effects are typically shown via the change in performance between immediate and delayed tests.

We thank the reviewer for raising both of these important points and address them in turn.

A recent meta-analysis of the impact of sleep on memory consolidation identified that the impact of sleep on recognition memory, while smaller than for free and cued recall, is still a small to medium

effect (Berres & Erdfelder, 2021). Our experiment is sufficiently powered to detect effect sizes much smaller than those identified in the meta-analysis. Therefore, we are confident that we have enough sensitivity to detect the sleep effect even using recognition memory. We have added this information to the manuscript on p. 8 lines 166 - 169:

"Reward memory will be measured using a paradigm adapted from earlier studies[30,36,37] and recently validated in our laboratory to yield positive effects of reward on memory performance (see supplementary material), 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."

It is true that many experiments calculate a change in performance between immediate and delayed testing. In a 2x2 ANOVA, evaluating the interaction is identical, statistically speaking, to a t-test on the change score. Our approach of assessing an interaction of timepoint X retention X reward in a linear mixed model is likewise statistically very similar to using a change score, with the added benefit of allowing us to show in the same analysis whether immediate and delayed recognition are different (i.e., if memory declined over the retention interval). Nevertheless, to determine the impact of using a change score we performed our data simulation again using the same parameters as last time (shown in Table 3) except that the evaluation of that simulation was performed using the change in hit rate between immediate and delayed testing:

```
Δ hit rate ~ retention:reward + (retention:reward | subject)
```

The power analysis indicated that a sample size of N = 1600 can detect at least an unstandardized effect size of .015 with 95% power at an alpha level of .020 (same input parameters as our original simulation). In the absence of any retention X reward interaction our data simulation yielded false positives on 2.1% of occasions. The change score would mean a reduction of 150 participants (or 8 %) compared to our current plan to achieve the same power, see p. 44, lines 880 – 890. Due to the added benefit of being able to evaluate the difference between immediate and delayed testing in the main analysis, we have decided to keep our analysis plan.

The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

4) One of my foremost concerns relates to the authors plan to collect a large amount of data for several exploratory analyses, seemingly at the cost of the main research question. In particular, the authors will not exclude participants on the basis of an existing mental health condition, sleep disorder or use of medication (among various other factors that are known to affect sleep) so that they can carry out exploratory analyses of how their findings are influenced by these factors. In my opinion, they should design their experiment in such a way that it is optimised for their main research goal (to test the selective effect of sleep (vs wake) on high (vs low) reward memories) and then carry out exploratory analyses where possible. Accordingly, they should extend their exclusion criteria to other factors that are known to affect sleep, so that they can maximise experimental control. Along the same lines, given the known effect of age on sleep architecture (such as slow-wave sleep, which has been linked to the preferential consolidation of salient information), I recommend that the authors restrict their sample to a narrower age range (e.g. young adults). This, again, will optimise their ability to test the main research question and provide a strong platform for future research in which the

effects of age, mental health condition etc, can be studied in the context of the selective memory function of sleep.

The reviewer rightfully points out a series of important issues that give the impression that we are valuing exploratory analyses over confirmatory analyses. Below we address their concern point by point:

Our decision to include participants with existing mental health condition, sleep disorder or use of medication was not primarily made with the intention to conduct exploratory analyses. Rather we are collecting a representative sample that will allow us to generalize to the population that would be the target of interventions. In this experiment we are certainly trying to find a balance between experimental control and recruiting a representative sample. One of the major problems with sleep and memory research is that it is typically conducted in young highly educated and healthy student populations. One of our primary goals is to overcome some of this issue especially since we have the broader intention to use the findings of this experiment to derive a precise effect size estimate that can be brought into clinical settings. Of note, 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 (Jacobi et al., 2014). In addition, populations in clinical settings are much more diverse than is currently represented in the sleep and memory literature. To explain and justify this further, we added the following to our methods section in the manuscript on p. 18, lines 227 – 241:

"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."

In addition, we have extensive experience performing online experiments having conducted a number of online, web-based experiments over the past 6 years including a large-scale registered report (Morgan et al., 2019; N = 4,000). In that time, we have learnt that one major limitations to collecting data online and employing numerous exclusion criteria is that we substantially reduce our pool of participants. For example, in our registered report we spent over 2 years collecting data in an online experiment and one of the major contributors to the length of that project was exclusion criteria. During that time, we excluded over 10,000 participants in our pre-screening questionnaire and despite that we did not identify that there was a significant difference in memory between sleep and wake conditions. So, in that one example, it is unclear whether or not implementing those exclusion criteria was worth it – we were unable to find a benefit of sleep on memory despite having a "homogenous sample". Additionally, there is no standardisation for which disorders researchers should exclude with many labs implementing different criteria but still being able to detect a sleep effect regardless.

Nonetheless, we understand the points that the reviewer has made on this matter and therefore we have implemented the following changes to find a balance between generalisability and sensitivity:

- The age range for our stratification will be limited to 20-59 years of age and so Figure 2 has been adjusted accordingly as has our stratification presented in our supplemental material
- Participants who nap or consume alcohol in between the retention interval in session 1 and 2 will be excluded from the data analysis, this has been added to a new table which clearly labels our inclusion and exclusion criteria on Table 2, p. 13.

Notably, it is the case that our sample size in the young adults (20-39 years old) remains much larger than in any other study on the effect of rewards on sleep-dependent memory consolidation (n = 760). Therefore, we can conduct an analysis limited to these participants and exclude mental disorders post-hoc to probe whether the effect is evident in a "homogenous" sample like previous research, which we have added to the control analyses on p. 46, lines 945 - 949.

"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 possibly only evident for young healthy adults."

5) Along the same lines as the above comment, given that encoding strength is known to influence the effects of sleep on memory consolidation, keeping encoding strength consistent at learning would presumably be of benefit to the main research question.

We agree that this is important. However, it is not feasible to keep encoding strength consistent in the present paradigm, as it does not allow, e.g., criterion learning as is often used for paired associates. However, we are using different durations for our stimuli, which will allow us to explore this question. We have added this information to the manuscript (quote, see our answer to concern 6 below).

6) Related to the above point, I question how the authors can separate the effects of reward from encoding strength at learning? Presumably high reward items will be encoded more strongly at the learning phase than low reward items, which would influence their consolidation in sleep. How do the authors intend to control for this?

We agree that this is an important consideration. For this reason, we have included duration of stimulus presentation at encoding as an important exploratory condition. In the literature, it has been argued sleep benefits low memory strength items the most (Drosopoulos et al., 2007; Payne et al., 2012). In contrast, the reward effect is suggested to be largest for high reward items (which would be the ones with the highest memory strength according to the reviewer). We have added this reasoning to the manuscript on p. 35, lines 687 - 692:

"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."

7) At several points in the manuscript the authors say that data will be collected for exploratory analysis, but they do not provide much insight on the nature of the questions they intend to address. For example, it would be interesting to know the authors' motivations for collecting data on participants' experiences during the day, and the question(s) they intend to address in this exploratory analysis.

Our study is a considerable effort both to us and the participants involved, to balance the cost-benefit-ratio, we have added some measures for exploratory analyses. We are more than happy to provide the reviewers insights into the exploratory analyses that can be performed on the data we collect. Throughout the manuscript we now provide an example of how the data that is collected for exploratory analyses may be used throughout the Materials section. Of note, these exploratory analyses will mostly not be included in the stage 2 report, but be left for dedicated publications as this would go beyond the scope of a registered report.

Caffeine Consumption Questionnaire:

"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" p. 28, lines 520 – 522

Epworth Sleepiness Scale (ESS):

"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." p. 28, lines 531 – 533:

Reduced Morningness Eveneingness Questionare (MEQr):

"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." p. 29, lines 548 – 551

St Mary's Hospital Sleep (SMHS) Questionnaire:

"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." p. 29, lines 559 - 561

The Pittsburgh Sleep Quality Index (PSQI).

"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." p. 30, lines 568 - 570

Becks Depression Inventory – Short Form (BDI - SF).

"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." p. 32, lines 613 - 615

8) The plan to use a within subjects is statistically optimal, but I am quite concerned about potential attrition rates, given the large number of questionnaires to complete at the recruitment stage, the lack of guaranteed financial compensation, the long delay between completing the sleep and wake conditions, and the delay between the initial recruitment screening and the first main experimental session (minimum of 24 h). The proposed use of a stratified sample could compound this even further,

as the authors will need to "turn away" a lot of participants. Can the authors provide assurance that this plan is feasible?

Attrition rates are indeed a concern that we must take seriously in order to assure that recruitment in this experiment is successful. We believe the following will mitigate high attrition rates:

- We have implemented a refer a friend option, which guarantees that participants will receive a 5-euro voucher, if they and their friend both complete the full experimental procedure
- Based on experiments, which are half of the length of the experiment at hand, attrition rates were as follows: of 265 participants that joined the experiment only 63 failed to return for both parts of the experiment, an attrition rate of 23%, Therefore, we believe it will be unlikely that attrition rates will differ substantially from this experiment.
- Our gamified Motivated Learning Task has received an overwhelming number of positive comments with many participants reporting that they wanted to complete the experiment not for payment, but because they enjoyed the aesthetics of the task and found it very engaging.
- Rather than compound attrition rates or limit participation, our stratification actually opens up the experiment to many more individuals than typical sleep and memory experiments, since we are not restricting the sample to highly educated healthy young adults.

9) There is quite a bit of flexibility in the time windows that participants can complete the morning and evening sessions, which might be a further source of noise. For example, if in the PM (sleep) condition a participant completes the first test at 7pm, goes to bed at 11pm, wakes up at 8am and is tested again at 11am, there is 7 hours of wakefulness between the immediate and delayed tests. Now, if the same person in the AM (wake) condition does the first test at 11am and the second test at 7pm, there is nearly the same amount of wakefulness as there is in the PM (sleep) condition. Can the authors provide a tighter control of the timings to mitigate this issue?

We agree with the reviewer that there is too much flexibility in the time windows that participants can take part in. Therefore, we have now amended the manuscript on p. 19, lines 309 – 315 with the smaller windows of opportunity for participation:

"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."

And we have also adjusted the times on Figure 4 on p. 18 which depicts the procedure of the experiment.

10) In subsidiary analyses, the authors plan to examine the sleep-memory effects at different levels of reward, with the prediction that an effect of sleep will be observed for the high rewards but not the

low rewards. I recommend the use of Bayesian approaches to quantify evidence for the null for the low reward levels.

We agree that we should be using an analysis to determine whether or not memory at lower reward levels between sleep and wake conditions at delayed testing is equivalent. Our preference is to stay within the Null-Hypothesis-Significance-Testing (NHS) framework to keep the analysis plan consistent, since this is our approach already elsewhere in the manuscript. Therefore, we will perform an equivalence test (Lakens et al., 2018) for memory of low reward items between sleep and wake conditions. On p. 46, lines 936 – 944 we have added the following:

"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 motivated to learn. 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."

Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

11) Will participants be excluded if they indicate that they have undertaken a prohibited action (e.g. napped in the wake group, consumed alcohol in the preceding 24 h)? Please make this clear.

Yes, these participants will be excluded. We have now provided a list of inclusion and exclusion criteria in Table 2, p. 13., which includes napping and alcohol consumption.

12) What will happen if the participants are continuously too slow on the flanker task? Will they be excluded?

We have now clarified that participants who are too slow on the flankers task will be excluded from the data analysis. However, in our pilot data we did not exclude participants due to this. Specifically we state:

"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."

And reiterate this in Table 2 on p 13.

13) Are participants informed about the different levels of reward in the instructions or must they learn this adaptively? I think the former would be a better control.

Participants are informed about the different levels of rewards in the instructions and learn the reward contingencies associated with them as described on page 24, lines 434 – 455. We now clarify this a little earlier in the manuscript on p. 21 lines 348 – 349:

"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. ”

and have provided the table below in the manuscript to make this clearer to the reader one p. 33:

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.

14) How long will encoding and retrieval take, given the large number of trials? This feeds into the attrition issue raised above.

In our pilot experiment, which validated our motivated learning task (see supplemental material), the learning phase took an average of 18.50 minutes to complete and the test phase took and average of 27.60 minutes to complete (note that in the present study the test is only on half of the items for immediate and delayed recognition testing and therefore shorter). Our pilot experiments indicate that this number of trials is ideal to balance power (trial number influences precision and thereby power) and participant motivation. In addition, the pirates cover story has contributed to relatively small attrition in participants that had already started the experiment. In that experiment attrition rates were as follows:

A total of 263 participants completed the study phase of that experiment and a grand total of 200 went on to complete the test phase of the experiment. This means that 63 participants only completed the study phase leaving us with an attrition rate of 24%. In light of this we are not concerned about dropout rates since most participants complete the experiment at a rate that is comparably lower than other online research conducted online (see Moodie, 2018).

Nevertheless, for transparency we have now included the approximate length of the learning phase in the manuscript on p. 20 – 21, lines 347-349:

“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).”

And in the approximate length of the retrieval phase on p. 21 line 363:

“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].”

15) Why are participants asked to retrieve the reward if performance is based on the old/new decision? What is the purpose of the confidence scale? Why is the reward amount not with the boxes in the test phase?

Participants are asked to retrieve the reward to determine whether or not they have source memory of the reward that was shown to them during the study phase. This procedure corresponds to our pilot experiment and allows us to collect additional interesting source memory information. Confidence is routinely assessed when collecting recognition memory performance and some papers have used confidence to show stronger reward effects (note our analysis on this included under Control Analyses on p. 46 – 47, lines 950 - 963). We do not show the reward during retrieval as this might induce response bias in the participants. To clarify we have added the following to the methods section on p 24 lines 434 - 457:

"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."

16) How will people document their wake experience every hour? Are they expected to do this independently and then type it in at the tests?

Thank you for pointing this out to us, we apologise for not making this clear. We have now clarified how participants will document their wake activity in the manuscript of p. 34, lines - 658 - 660:

"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."

17) The authors state: "To prevent p-hacking, p-values will only be calculated once a model with good convergence is identified" – how do they define "good convergence"?

We define good convergence as the absence of non-convergence or poor-convergence. Using lmer in RStudio, the model can fail to converge and does not produce an output (non-convergence), or can produce a singular fit (poor-convergence), which generates a warning message produced by the lmer package. A singular fit indicates that the random effects structure is too complicated for the observed data to support it, in other words the model is overfitted to the data. To make this clearer and to explain what we mean by "good convergence" we added the following to p. 40, lines 796 -800:

"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."

Other Comments

Abstract

18) It is not clear from the abstract whether the authors expect there to be a main effect of reward (i.e. better memory for high vs low rewards in both the sleep and wake conditions) in addition to the interaction (with the magnitude of this reward benefit being stronger in the sleep condition than the wake condition). I encourage the authors to make this clear, so that the reader can fully understand the expected direction of the results.

Thank you for identifying this omission, we can confirm that we do predict there to be a main effect of reward, a main effect of retention and an interaction between reward and retention at delayed testing. To make this clear we have added the following to the abstract on p. 2, lines 32 – 35:

"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."

Note that the wording of these predictions has also been adjusted at the end of the introduction on p. 8, lines 170 - 175:

"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."

Note that the wording in the design table on p. 10 differs slightly as it explicitly refers to the hit rate instead of "greater retention" as per the recommendations of PCIRR.

In addition, the explicit prediction that high reward items will be better retained compared to low reward items is now added to the design table on p. 10 and is now referred to throughout the manuscript as H2 (whereas the former H2 has become H3).

Introduction

19) Some of the statements in the Introduction are quite strong and could be toned down slightly to reflect a more balanced view of the literature/current opinion. For example, "During sleep, memory traces that were encoded throughout prior wakefulness are replayed repeatedly and thereby strengthened" and "Importantly, sleep specific brain activity and especially the activity of hallmark oscillations (slow oscillations, hippocampal ripples and sleep spindles) that coordinate this replay drive

greater memory performance in those tasks" refers to only one (albeit influential) theoretical perspective.

We thank the reviewer for pointing this out. We have adjusted the statements that the reviewer refers to in addition to softening the overall argument made in that paragraph and also refer the readers to alternative explanations of the sleep effect on memory, see p. 3, lines 45 - 63:

"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 neocortex 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."

20) The link between the basic scientific questions (consolidation of reward related memories) and societal issues (e.g. additive behaviours) is interesting, but the authors only briefly touch on this. I think it would be helpful if they could explain a little more clearly about why the impact of sleep on rewarded information can guide maladaptive behaviour such as unhealthy eating, smoking etc – what is the proposed mechanistic link here? Relatedly, how would targeted interventions make use of the findings of this study?

We think that this is a great suggestion from the reviewer and we are more than happy to provide a more detailed explanation of the mechanistic link between the basic scientific questions, the societal issues and how targeted interventions would make use of the findings of this research on p. 7, lines 140 - 150.

"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."

21) The authors intend to use an online paradigm to study their research questions. The use of online paradigms to study sleep-memory effects has grown in recent years, with a number of studies showing a positive effect of sleep outside of the lab. I think the authors could mention this in their introduction, to highlight that the online method is an appropriate tool to measure the memory effects of sleep.

This is another great suggestion from the reviewer. We have added in a sentence in the introduction to highlight this important feature of web-based sleep and memory experiments on p. 7, line 157 - 159:

"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."

22) Given that a lot of readers will not know what the AM:PM design is, I encourage the authors to elaborate on this point in the Introduction

We have now provided a description of the AM:PM design and its suitability to test the sleep effect on memory on p. 7, lines 154 - 157:

"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)."

23) What are the authors referring to by "retrieval function" – page 7 line 142?

We have changed the wording to "general retrieval performance" on p. 8, line 176, as this better conceptually reflects the control variable we meant here.

"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)."

24) The authors state: "Regarding sleep, there is no consensus whether sleep enhances rewarded memories through additional dopaminergic neuromodulation during reactivation or rather dopamine sets a tag during learning that leads to enhanced reactivation without additional dopaminergic neuromodulation" – perhaps this should be removed or at least rephrased as it could give the impression that this is a question that the authors intend to address in their study.

We agree with the reviewers that what we have written gives the impression that we will address this in our experiment. Therefore, we added the following to p. 4, lines 86 – 89 to the following:

"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 neurophysiological mechanism be characterized."

Methods

25) It looks like the light and dark green colours in Figure 2 represent participant sex, but this is not clear.

This is correct and we have now added a figure legend explaining the colours to Figures 2a and 2b on p. 14 and 15 respectively.

26) The caption of figure 4 describes a 3-point Likert scale whereas the main text describes a 4-point scale.

We have now corrected this on p. 17, line 288.

"For each landscape image, participants decide whether an image is old (i.e., the image was shown during learning) or new (i.e., an image not shown during learning) and rate their confidence in their decision using a 4-point Likert scale (guess, somewhat sure, sure, very sure)."

27) The authors state: 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." - It should be clearer from the outset that the authors expect an overall decline in performance between the immediate and delayed tests, but that the magnitude of this decline will be smallest in the high reward/sleep condition.

We have now made this clear early on in the manuscript on p. 8:

"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 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."

References

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron*, *50*(3), 507-517.

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature neuroscience*, *15*(10), 1439-1444.

Bennion, K. A., Payne, J. D., & Kensinger, E. A. (2016). The impact of napping on memory for future-relevant stimuli: Prioritization among multiple salience cues. *Behavioral neuroscience*, *130*(3), 281.

Berres, S., & Erdfelder, E. (2021). The sleep benefit in episodic memory: An integrative review and a meta-analysis. *Psychological Bulletin*, *147*(12), 1309.

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General*, *136*(2), 169 - 183.

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

Fischer, S., & Born, J. (2009). Anticipated reward enhances offline learning during sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1586 - 1593.

Jacobi, F., Höfler, M., Siegert, J., Mack, S., Gerschler, A., Scholl, L., ... & Wittchen, H. U. (2014). 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*(3), 304-319.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, *17*(1), 3-10.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269.

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

Prolific Team (2022). How do I set up a longitudinal / multi-part study? . In Prolific Help Centre. https://researcher-help.prolific.co/hc/en-gb/articles/360009222733-How-do-I-set-up-a-longitudinal-multi-part-study-#heading-4

Morgan, D. P., Tamminen, J., Seale-Carlisle, T. M., & Mickes, L. (2019). The impact of sleep on eyewitness identifications. *Royal Society open science*, *6*(12), 170501.

Payne, J. D., Tucker, M. A., Ellenbogen, J. M., Wamsley, E. J., Walker, M. P., Schacter, D. L., & Stickgold, R. (2012). Memory for semantically related and unrelated declarative information: the benefit of sleep, the cost of wake. *PloS one*, *7*(3), e33079.

Tononi, G., & Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, *81*(1), 12-34.

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

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

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

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

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

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

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