Response to the Editor

We thank the editor and the reviewers for their invaluable feedback and apologize for the huge delay in our response due to an absence of work from the first author due to health problems. We have made changes to the manuscript based on the editor's and reviewer's comments. As there is no suitable pilot data that combines dream reports and emotional images available for hypotheses 1B and 2B, and 3, these hypotheses were dropped from the registered report. Furthermore, the manuscript has been slightly updated to include newer literature and some language edits. One co-author was removed as he could no longer commit time to the project due to the delays. We have also adjusted one inclusion criteria based on experience from other studies in the lab, showing that it was too strict (BDI cut-off from 15 to 20). We reply to the editor's and reviewer's comments point by point below. You can find our response in bold, sections from the manuscript in cursive, and changes sections in the manuscript highlighted in yellow.

1. For a RR, it is crucial to ensure an exact alignment between the power analysis and statistical methods that will test the hypotheses. Power analysis methods are available for LMEs (e.g. here) or should be reported using simulations. It was unclear to me what approach you had taken with the power analysis and which model(s) it applied to within the LMEs. The power analysis must be calibrated to be sufficient to meet PCI RR requirements for the most conservative hypothesis test among the set.

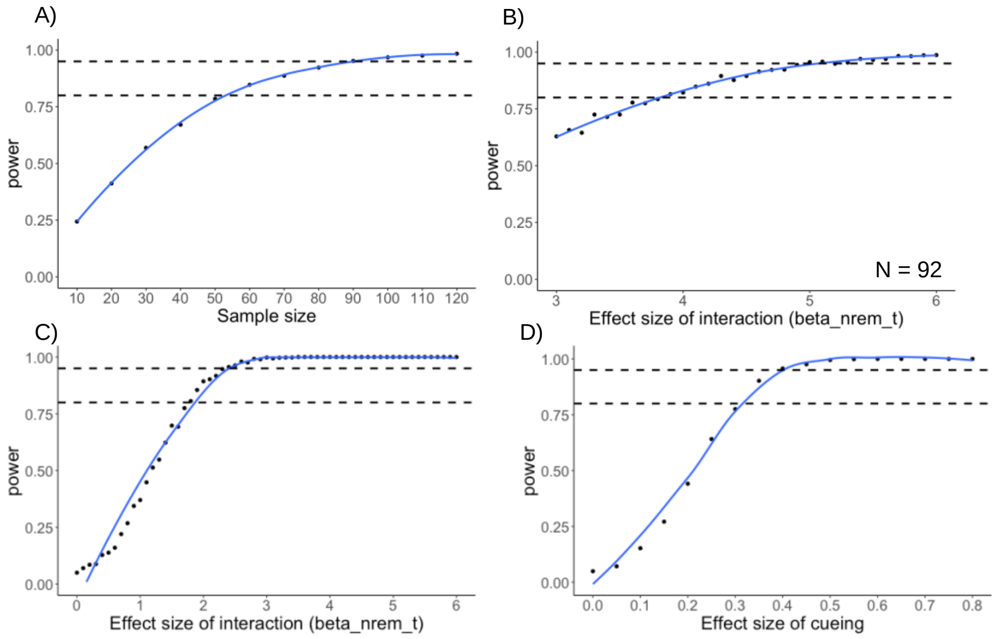
We have re-run our power analysis by using simulations for the multilevel models we will use in the registered report analyses and have now adjusted the text to reflect this new approach. Furthermore, we uploaded the R-Markdown for the simulations to our OSF.

For Hypothesis 1, we relied upon data from our previously published results to simulate 1000 new data sets (Schoch et al., 2019)**. We could show that with the betas we estimated from the previous data set, we would have 95% power to detect effects of the same size with 90 participants. To keep the sample size consistent with the ethics protocol, we opted to stay with 92 participants. We then ran a sensitivity analysis to show that with 92 participants, we have 95% power for b ≥ 5 (and 80% power for beta ≥ 3.9). The same analysis was done for the control model and showed sufficient power.**

For hypothesis 2, we did not have an exact corresponding data set available. However, we simulated data based on our previously published results on the incorporation of the task presented right before sleep into dreams versus the task shown 10 days before/later. Considering that TMR would give much higher specificity with cueing and that TMR shows effects in a similar range on memory performance as sleep does in general (hedges g for sleep = 0.28 for sleep (Newbury et al., 2021) and hedges g for TMR = 0.29 (Hu et al., 2020)), we think that

this is an adequately sensitive test. We show that we reach 95% power for n > 80, and our sensitivity showed that we have 95% power for b ≥ 0.4 (and 80% power for beta ≥ 0.3). These adjustments are now reflected within the manuscript text (pages 9 and 10), Table 1 (page 31, not included below due to size), and Supplementary Figure 2 (page 46).

*We conducted a power analysis using simulations[61] based on the results of our previous study[9]. Simulations were done in RStudio[62] and using the packages tidyverse[63], lme4[64], lmerTest[65], fitdistrplus[66], broom.mixed[67], faux[68]. For hypothesis 1, we simulated datasets containing 10 – 120 participants (across 1000 repetitions) based on estimates from the data of our previous study (n = 22). 95% power was reached with 90 participants (suppl Fig 2a). Using a sensitivity analysis with 92 participants and 1000 repetitions while varying the beta for the interaction of interest (NREM incorporation * time) from 3.0 to 6.0 (in 1.0 steps), we estimate that b ≥ 5 will be detected with 95% power and b ≥ 3.9 with 80% power (b = 5.14 estimated from the previous study, suppl Fig 2b). The same sensitivity analysis was done for the model controlling incorporation for chance level, estimating 95% power for b ≥ 2.4 and 80% power for b ≥ 1.8 (b = 7.12 estimated from the previous study, range tested 0 – 6.0, suppl Fig 2c). For hypothesis 2, we simulated datasets based on data from our previous study on incorporating the task into the dreams (comparison task from before sleep and the one 10 weeks before/after). We estimate that the effect size of TMR will be similar (based on similar effect sizes reported for TMR on memory performance compared to general sleep effects). For 92 participants (1000 repetitions), we showed that the sensitivity of our analyses was 95% for b ≥ 0.4 and 80% for b ≥ 0.3 (0.45 estimated from the previous study).*



Supplementary Figure 2. Effect size simulations for hypotheses 1 and 2. A) *For hypothesis 1, we used effect size estimates from our previous study to simulate 1000 datasets with 10 – 120 participants each. 95% power is reached with 90 participants.* ***B)*** *Sensitivity analysis with 92 participants and varying the effect size of the interaction (NREM incorporation*timepoint). With 92 participants, we reach 95% power with an effect size of b ≥ 5 and 80% power with effect size b ≥ 3.9.* ***C)*** *Sensitivity analysis with 92 participants and verifying effect size of the interaction (NREM incorporation*timepoint) for the model controlling incorporations for baseline. We reached 95% power with an effect size of b ≥ 2.4 and 80% power with effect size b ≥ 1.8.* ***D)*** *For hypothesis 2, we used effect sizes from data on task incorporation into dreams to estimate potential effect sizes for TMR. In the sensitivity analysis with 92 participants and varying the effect size of cueing from 0.0 – 0.8 (0.05 steps), we estimate 95% power with an effect size of b ≥ 0.4 and 80% power with effect size b ≥ 0.3.*

2. Rather than sampling 15% over the required sample size, commit to collecting the minimum required sample size to achieve the desired power, regardless of the exclusion rate. This will provide certainty concerning the sample size included in the analysis.

We agree and have adjusted the Registered Report to reflect that we will collect 92 eligible participants regardless of the dropout rate.

3. "Outliers will be inspected, but not removed unless there is a reason to believe they are due to measurement error. However, to ensure robustness of the results, models will be repeated with outliers (> 3 SD) removed." I found this a little vague. At what level of granularity (i.e. what cells in the design) will this rule of >3 SDs be calculated and applied? Will statistical outcomes with or without outliers determine the conclusions? How will the conclusions will adjusted if outlier removal alters support for the hypotheses? These contingencies need to defined precisely to eliminate risk of bias.

We agree that this was vague and have now added additional information. Firstly, outliers will be determined for each variable separately and removed cell-wise. Secondly, our main interpretation will be based on the models with the outliers included, however, the models without outliers will be used to determine if the effects are robust or dependent on a few participants with extreme values. We have adjusted the analysis section (pages 11 and 14) to reflect this.

First, we will examine outliers in each variable. Outliers will be inspected but not removed unless there is a reason to believe they are due to measurement error (e.g., the wrong task presented, audio not working, etc.).

…

To ensure the robustness of the results, models will be additionally analyzed with outliers (> 3 SD for each specific measure) removed at the cell level. While interpretations will be based on the models with outliers included, these additional analyses will be used to interpret if the effects are robust or dependent on a few participants with extreme values.

4. "If the data distribution of the residuals is non-normal, we will examine if a gamma distribution is a better fit. If problems persist, data will be transformed with a logtransform. Missing data will be estimated using full maximum likelihood." Define the precise conditions for concluding non-normality and the method used to determine it. Define precise contingencies under which different transformations will or will not be applied and corresponding downstream effects on analysis plans (if any).

   Now that we have done the simulations and were able to inspect residuals, we have removed this part of our analysis plan and plan to analyze our data without transformations.

5. “…due to bad sleep quality or problems with data recording.” Provide an objective set of criteria for determining that data is “bad quality”.

We now specify sleep quality in terms of sleep efficiency (< 70%), this is reflected in Supplementary Table 1 (pages 42 – 44) and Figure 2 (page 29, both not displayed in this document due to size).

6. A design table is included but please adapt this it to the PCI RR version, which has additional requirements (see here). I struggled to follow the structure and logic of the Control analysis section within the design table. Ensure that there is a one-to-one correspondence between statistical tests and interpretrations for all hypotheses (including the Control analysis section). If you predict certain relationships for the control analyses these should be include in the hypothesis column for that row. Perhaps insert a Variables column to the design table to keep the Analysis Plan column focused on the model(s) that will be tested. Alternatively, rather than adding a Variables column, move the description of variables out of the design table and into the Method section in the main text.

The design table was adapted to the PCI version (page 31 – 33) and restructured to make it more transparent, including clear hypotheses and interpretations for each subpoint. Some control analyses were removed as they would not directly affect interpretation. The variable descriptions have now been moved to the statistical analysis section in the main text as suggested by the editor.

7. Hypothesis 1 is multi-pronged (with multiple sub-hypotheses) and the testing of Hypothesis 3 is contingent on support for Hypothesis 1. This makes it especially crucial to define the precise conditions under which Hypothesis 1 will be considered to be confirmed or disconfirmed based on combinations of outcomes for each of its sub-hypotheses. This needs to be made crystal clear.

Hypothesis 3 has now been removed; therefore, this contingency does not exist anymore. However, we have clarified when hypothesis 1 is considered confirmed or not in Table 1 (pages 31 - 33).

8. The design table splits the hypotheses into sub-hypotheses (which is good) so please do the same in the introduction. I suggest listing them at the end of the Introduction in bullet point form to achieve maximum clarity.

Due to the dropping of the hypothesis relating to emotional memory, there are no more sub-hypothesis, however the hypotheses are now listed as bullet points at the end of the introduction, nonetheless (pages 5 – 6).

In this study, we will test the following hypotheses in a sample of 92 participants:

- *Hypothesis 1a) Incorporations of the picture categories of the memory task into NREM dreams, but not REM dreams, are associated with* *improved performance on the memory task the next morning and 4-days later.*

- *Hypothesis 2) TMR leads to the subsequent incorporation of the associated image categories into dreams during both NREM and REM sleep stages.*

9. Use of Bayes factors needs to be substantially elaborated, making clear the priors and any other relevant parameters, and the precise conditions under which Bayes factors will be reported. Bayes factors are mentioned in the study design table but, unless I missed it, are not mentioned anywhere else in the manuscript.

We now include more information on the use, parameters, and interpretation of the Bayes Factors in the Statistical Analysis section (pages 13 – 14).

If the initial NHST results in a p-value above our 0.5 alpha threshold for the specified fixed effects, we plan to explore further the extent to which our data provides evidence against/for our hypotheses by using Bayesian methods, specifically Bayes factors BF01 to quantify how much more likely the null hypothesis is relative to the alternative hypothesis. We will use the bmrs[84] and BayesFactor package[85] to implement the Bayesian analyses. We will use a balanced null comparison to test for the presence/absence of the fixed effect[86]. We will follow the guidelines proposed by [87] and consider the evidence to be: inconclusive/null if BF01 = 1; weak in favor of H0 if 1 < BF01 < 3; moderate in favor of H0 if 3 < BF01 < 10; strong in favor of H0 if 10 < BF01 < 30; weak in favor of H1 if 1/3 < BF01 < 1; moderate in favor of H1 if 1/10 < BF01 < 1/3; strong in favor of H1 if 1/30 < BF01 < 1/10.

10. The Statistical Analysis section needs to be greatly expanded to provide comprehrensive detail, rather than relying entirely on the design table (this is also noted by the reviewers).

We have expanded the statistical analysis section (pages 11 – 15) to include our analysis plan in detail, including the variable descriptions, which have been moved from Table 1 (pages 31 – 33), and further explained the steps in the design table (both not included due to size).

11. The exclusion criteria are quite complex so please add a CONSORT-style diagram to illustrate the rules under which data / participants will be excluded at different stages in the data acquisition and analysis.

We have now added a CONSORT-style diagram to illustrate the inclusion/exclusion criteria across the whole study (Figure 2, page 29).

Reviewer #1 (Remarks to the Author):

This manuscript presents the authors' proposal for conducting a study that specifically examines the extent to which task information is incorporated into dreams and how that relates to subsequent memory performance. This is an interesting question, and previous studies investigating this question have uncovered conflicting evidence. This is also a difficult question to address since the contents of dreams are difficult to probe.

In this design, the authors propose to have subjects perform a simple visual association task prior to sleep. Then during sleep, at four points during NREM and also four points during REM, the experimenters will wake the subject and ask them to report their dreams. This will address the question of whether task information is incorporated into dreams, and subsequently how much this correlates with memory performance. In a second version of this paradigm, the experimenters will introduce an audio recording of words presented in the task during sleep prior to waking(targeted memory reactivation) in order to assess whether this manipulation will increase the likelihood of incorporating task information in the dreams, and the subsequent effects on memory. An additional question to be pursued here is whether the incorporation of task elements during REM sleep change the emotional valence of those images.

This is a well designed study and an interesting question, and it would potentially provide a valuable result to the community of researchers studying the relation between dreams and memory. This is also, however, a challenging study largely because of the challenges inherent in probing the contents of dreams. This entirely relies upon the subjects ability to recall and report details of their dreams. The task design used here (training on reporting prior to the experiment and the experimental paradigm itself) partly addresses those limitations, and so there is little more that can be done about this. One strength of the design is the relatively large numbers of participants that are expected to be enrolled. This could help mitigate some of the concerns regarding the data that are being captured.

We appreciate that the reviewer agrees that we optimized our study design to address the limitations of dream reports. We also want to highlight that we use participants with high dream recall frequency. In our previous study with the same criteria, we got dream reports from 87% of awakenings. Furthermore, while dream reports have limitations, all evidence so far points to them generally being reliable (Windt, 2013, 2015)**: they align with neural data** (Siclari et al., 2017)**, with movements in REM sleep behavior disorder** (Valli et al., 2012) **and with lucid dreaming reports** (Konkoly et al., 2021)**. We have also ensured that the methods we use are optimally tailored to the specific demands of the study and represent the gold standard in current dream research (i.e., dream reports gathered**

immediately following the awakening from the targeted sleep stage, combined with specific and carefully worded questions and training of participants). Moreover, given that we are interested in incorporating task-related information in dreams, there is no alternative to using retrospective dream reports.

Another issue is that the images and words used in the memory task are drawn from a large dataset and are likely very common, which makes it unclear whether the contents of a dream are explicitly related to the task, or just coincidentally overlap with some elements of the task (for example, if the task contains a picture of a dog, and you dream of a dog, is it because of the task, or is it because you happen to be dreaming about a dog). This could be partly addressed by using only very unique and rare images or items for the memory task.

The issue that potentially dreams unrelated to the task could be misinterpreted as incorporations is an important one. The approach we adopted ensures that we can detect deviations from the baseline regarding the presence of the studied categories. Even if some of the categories are commonly present within dreams, we are testing checking if their occurrence is changed due to our manipulation by using two approaches: 1) For the spontaneous incorporation, we make use of the two measurement nights within the same subject, which gives us a baseline of incorporation. The control model (Multilevel model correcting for baseline incorporation) reflects this. All categories are counterbalanced across participants to address potential differences in the natural occurrence of these categories. 2) For the TMR nights, we use three categories which enables us to have one category cued in NREM, one in REM, and one uncued category (counterbalanced across subjects).
Furthermore, we took specific caution to choose images that relate to only one of the categories and do not contain any other categories.
While having completely distinct images could be an interesting approach, it would be extremely difficult to find 200 images showing uncommon images without overlapping. Furthermore, using familiar images ensures that they are equal for all participants, while generally unfamiliar images might still be familiar to some but not others.

Reviewer #2 (Remarks to the Author):
The authors propose a study investigating the relationship between dream content, memory retention, and emotional processing. They do a nice job of covering the literature (though see a few notes below), the planned data collection seems sufficient to address their questions, and the planned analysis are thorough and clear. The proposed study is interesting and does not have any obvious flaws (provided some of the answers to the below concerns are reasonable), though I do question whether the study’s importance, even if all hypothesized results are obtained, warrants publication in Nature Communications. That said, I will leave that up to the editor to decide and I offer my comments/concerns below.

We thank the reviewer for their positive assessment of our study design. We want to highlight that we think the study will bring considerable conceptual advances not just for dreaming research but also for sleep research, memory research, and psychology in general. However, implications may also reach well beyond these fields. First, the function of dreams has long been a topic of interest and continues to be controversial (Scarpelli et al., 2019; Schredl, 2018), **and this study promises to advance this central issue of dreaming research. Second, if dreams are relevant to memory consolidation, theoretical models of sleep-dependent memory consolidation need to account for the role of dreaming, i.e., link what happens on the biological level with the conscious experience. Third, if dreams are essential for emotional processing, this impacts psychology and psychotherapy significantly, which might rely on a scientifically-grounded analysis of dreams and dream content. Fourth, identifying the functions of dreams will advance our understanding of the processes that underlie dream generation and shaping, with critical translational implications. Indeed, alterations in dream content and frequency are commonly reported not only in sleep disorders but also in neurologic (e.g., epilepsy, stroke, dementia) and psychiatric conditions (e.g., schizophrenia, post-traumatic stress disorder). Such alterations and their significance are still poorly understood but may represent a sign of compromised or dysfunctional sleep-dependent processes that might, in turn, contribute to the clinical symptoms observed during wakefulness. Finally, it should not be forgotten that dreams have always represented an important component in human history and societies, attracting the interest of many individuals and sometimes influencing their behavior and beliefs. In this light, we are convinced that our present study has the potential to bring forward important advances affecting multiple fields of research and may thus be of interest to a broad and diverse audience.**

We have now highlighted this also in the manuscript on page 5:

Considering that the function of dreams has long been a topic of interest and continues to be debated [55,56], this study will provide a large empirical dataset as a basis to understand two potential functions of dreaming: memory and emotional processing.

Hypotheses:
Do the authors have a hypothesis regarding whether TMR during REM sleep (or TMR in general, since it will be applied during both SWS and REM on the same night) will further lower emotional and arousal ratings beyond ratings given on the spontaneous night? It seems reasonable to at least consider this for a planned post-hoc test, if there is no hypothesis.

We thank the reviewer for bringing up this critical point; while we would, in general, agree, we had to remove hypotheses regarding emotional processing from the registered report, as we do not have adequate pilot data to ensure we have enough sensitivity to detect the effects we are interested in.

Are there predictions regarding the follow-up memory recall test?

The follow-up memory recall is incorporated into the models, which take all time points (before sleep, morning, 4-days later) into account. This allows us to model the change from evening to morning and from evening to 4-days later. Our predictions for the follow-up test are in the same direction as for the morning recall. We have now specified this more clearly in our hypotheses on pages 5 - 6 and Table 1 (pages 31 – 33).

- Hypothesis 1a) Incorporations of the picture categories of the memory task into NREM dreams, but not REM dreams, are associated with improved performance on the memory task the next morning and 4-days later.

Are there any physiological predictions, such as correlations between stages or hallmarks of sleep like slow oscillations and spindles?

This manuscript will not include physiological analyses as the current registered report focuses on behavioral results rather than neurophysiology. However, the present work will pave the way to further investigations of physiological events if we can confirm our hypothesis regarding memory consolidation.

Introduction:
2nd paragraph: Oudiette et al. (2011) is a relevant study here.

We agree and have incorporated this study on page 3.

3rd paragraph: “less” -> “fewer”

We corrected this.

Final paragraph: Konkoly et al. (2021) is highly relevant here, as is Horowitz et al. (2020).

We included the two studies in our manuscript. See page 5:

During the sleep onset period, dream content has been successfully biased by using auditory stimulation[55] and during lucid dreams, participants were able to reply to questions presented aurally (among others)[56].

Methods:
Please define EGG for the readers in the main text.

We now define EEG on page 6.

For the adaptation night, participants will be invited to the Donders electroencephalography (EEG) laboratory at 21:30.

"In experimental session A, participants will be woken up a maximum of four times from NREM and four times from REM sleep, 15 minutes into each sleep stage. A free dream report for the last minute of sleep will be elicited during each awakening, followed by ratings on several scales." Why 15 minutes? What happens if this stage is broken up by another sleep stage (e.g., 8 minutes of REM, 2 minutes of stage-2, and then more REM)? I would advise the authors to lower this number unless they have a strong justification for it, considering the frustrations that will likely ensue if they enforce it strictly.

We appreciate the input from the reviewer and agree that we need to loosen the restrictions to make the study more feasible. By using 15 minutes, we wanted to have a compromise between allowing sleep-stage dependent memory consolidation to happen (which relies on undisturbed sleep) while simultaneously increasing the chances of being able to do as many awakenings as possible. We now change this definition to that the awakening should happen at least 15 minutes into the sleep stage, but that the last 1 minute before the awakening should not include any wake or other sleep stages, i.e., REM in NREM and vice versa. We believe the last part is essential to provide evidence of the difference between NREM and REM dreams. We thank the reviewer for raising this point and improving our study design. This is now reflected in the manuscript on page 7:

In experimental session A, participants will be woken up a maximum of four times from NREM and four times from REM sleep, at least 15 minutes after the first start of the respective NREM/REM sleep stage.

And with more detail in the supplemental methods on page 37:

The participants will be woken up to 8 times during the night following an awakening protocol (on project OSF) – four times from NREM and four times from REM sleep (at least 15 minutes into each sleep stage). For NREM sleep, N2 will be used as the start of the sleep stage, however, the awakening can be done in any NREM (N1, N2, or N3) sleep stage. The preceding 1 minute of each awakening should not contain any wake or the opposite sleep stage (i.e., REM for a NREM awakening and NREM for a REM awakening).

"The words will be presented for approximately 10 minutes before each awakening." How long will the authors wait between the final word presentation words and awakening? This is a critical detail and could very well influence their results.

Additionally, do the authors have a hypothesis regarding whether dreams will be more likely for more recent (vs. less recent) associated memories?

We agree that this is important to specify. The awakenings will happen more or less immediately after the final word presentation, we have now specified this as a time frame of 10 - 30 seconds. As all the cues will reflect one specific category, we will analyze incorporations of the category rather than specific images, therefore, we do not plan to analyze time effects within the TMR.

This is now specified in the manuscript on page 38:

The participants will be awoken between 10 – 30s after the last TMR at least 15 minutes into each sleep stage.

“maixmally” -> “maximally”
“108 healthy male” -> “One hundred and eight...” Please do not start a sentence with a number.

Thanks, we fixed these mistakes.

“When the participant is lying in bed, we will do a resting-state EEG measurement (1.5 min eyes open, 1.5 min eyes closed, 1.5 min eyes open, 1.5 min eyes closed).” Can the authors please explain the rationale for this? Is there a hypothesis linked to this measurement?

There is no hypothesis linked to resting-state measurement, however, we think this is an interesting additional data point for potential future research questions for which this dataset can be used (the dataset will be made freely available, therefore we also include measures that we think can add value later on).

“In both experimental nights, participants will be instructed to signal if they have a period of lucidity. Dream reports with lucidity will be removed from analyses (score >= 4 on the lucidity scale).” Can the authors also explain the rationale for this exclusion practice?

The idea behind the exclusion is that potentially the mechanisms of memory consolidation during lucid dreaming are potentially different than during non-lucid REM dreaming. While there is still a discussion on whether lucid dreams happen during a mixture state between non-lucid REM sleep and wake (Voss et al., 2009) **or activated REM** (Baird et al., 2022)**, to ensure that dream reports are homogenous within each sleep stage, we will exclude dreams with high lucidity scores (scale 1 - 5). We have, however, now removed the instruction for participants to signal lucidity as this might be difficult without practice (see point raised by reviewer 3) and will only remove clearly lucid dreams (score = 5).**

Statistics

What will authors do if control analyses do not go as planned e.g., H1a model is supported but, dream length is significant?

We have realized that calling these analyses control analyses can be confusing. Firstly, we removed the model for word length because there was no clear effect on interpretation. Secondly, we now have a primary (without baseline adjustment) and a secondary model for H1 (with baseline adjustment). We have now specified in Table 1 how we would interpret differing results between the primary and secondary model, specifically if either model is significant this adds support to hypothesis 1, however, if the primary model is not significant, but the secondary model is, this suggests that baseline-adjustments are needed to see an effect. We now clarify this in the analysis section on page 13 as well as Table 1 (pages 31 – 33):

If the interaction NREM_Dream_Incorporations:Timepoint is significant in either model, we will interpret this as evidence for H1 that NREM dream incorporations are significantly associated with memory performance after sleep. If the interaction REM_Dream_Incorporations:Timepoint is significant in either model, we will interpret this as evidence against H1 that REM dream incorporations are not significantly associated with memory performance after sleep. If the interaction is only significant in the secondary but not primary model this means that baseline adjustment for dream incorporations is necessary to detect association with memory performance.

Why these interactions? “NREM_inc_cor:Timepoint +REM_inc_cor:Timepoint” Aren’t the dream incorporations (and all related sleep variables) going to be the same, as attributed to both evening and morning? Please clarify if I’ve simply misunderstood, as it will perhaps be unclear to other readers too.

The interaction specifies that our main hypothesis does not reflect performance in the morning/evening but rather the change in performance relative to baseline (i.e., morning compared to evening, follow-up compared to evening). Specifying this as an interaction allows us to take into account that participants start with different baseline levels.

We have now added an explanation to the analysis section (page 13) :

NREM_Dream_Incorporations:Timepoint (interaction) Interaction effect to quantify changes between baseline (evening) and morning/follow-up dependent on incorporations into NREM dreams.

REM_Dream_Incorporations:Timepoint (interaction) Interaction effect to quantify changes between baseline (evening) and morning/follow-up dependent on incorporations into REM dreams.

What is different in the two main H1b, “Check if decrease across time is significantly dependent on valence” sections?

As mentioned previously, H1b is no longer part of this registered report.

Reviewer #3 (Remarks to the Author):

The authors propose a study to assess the role of dreaming for memory consolidation. This issue is of great significance for the cognitive neuroscience community. Despite numerous studies on the interplay of sleep and memory, this important research question could still not be satisfactory answered.

Overall, the proposed study design is suitable to investigate if memory consolidation actually benefits from dreaming. The introduction provides a comprehensive overview of the relevant literature and the proposed hypotheses are plausible.
Furthermore, the proposed methodology and analysis pipelines are sound and feasible. The authors provide a careful statistical power analysis base on conservative assumptions, finally leading to a relatively large number of 108 participants required for the study.

The methods are described very clearly. Besides my comments below, which need to be addresses prior to the study, the authors provide all necessary detail to prevent undisclosed flexibility in, and to enable exact replication of the proposed experimental procedures and analysis pipelines.

We thank the reviewer for their positive assessment.

Comments regarding the proposed statistical analysis:
Compared to the degree of detail of the study design description, the analysis paragraph is rather short and contains not enough detail.

For instance, the statement “If the data distribution of the residuals is non-normal, we will examine if a gamma distribution is a better fit. If problems persist, data will be transformed with a logtransform” needs more elaboration. Why is the gamma distribution the second best choice? Which parameters are expected to be normal, gamma, or lognormal distributed? How will the authors test for several distributions?

We agree that the statistical analysis section was not detailed enough, as also mentioned by the editor. This has now been vastly expanded. Furthermore, as we

have now run simulations and examined residuals, we are confident that we can use a normal data distribution and have therefore removed this section from our analysis plan.

Also the statement “Missing data will be estimated using full maximum likelihood.” needs further explanation. Why does missing data need to be estimated at all, instead of just being skipped? How exactly will missing data be estimated in an unbiased way and how is it assured that the estimated data does not affect the overall results? In order to prevent undisclosed flexibility in, and to enable exact replication of the analysis pipeline, this issue needs to be addressed.

We agree with the reviewer and have removed this statement. Missing data will now be excluded, however, participants will be included as long as one timepoint is available.

Comments regarding the (supplemental) methods section, the following issues should be addressed prior to the proposed study:

The authors state that “The adaptation night is scheduled as closely as possible to the first experimental night (the night before the first experimental night, maximally seven nights before)”.
However, I strongly recommend the adaptation night to be always immediately before the first experimental night.

We agree with the reviewer that this is the ideal study design. However, to make the study feasible, we also want to include the possibility of having a larger delay between the adaptation and experimental night, as, e.g., participants or experimenters could fall ill and require rescheduling or the laboratory being occupied. As this is a complex study with multiple nights per participant and a high target sample size, some flexibility is necessary, so data collection remains feasible. Furthermore, most studies in the sleep and memory field do not have experimental nights immediately following adaptation nights, especially considering that larger delays between the experimental nights are necessary (two weeks in our study).

Regarding this procedure: “three trials are conducted where words are presented at increasing sound levels (from 20 dB in 5 dB steps) until the participant shows an arousal.”
- It should be “20 dB SPL”
- How is the speech transduced? Open field or via in-ear head phones? If open field, then what is the distance between the loudspeaker and the participant’s ear? This is an important detail since the sound pressure level (SPL) decreases quadratically with the distance.

We agree with the reviewer that this was not well specified in the previous version. Additionally, we made the following changes: Firstly, we adjusted the

start of the sound levels due to unreliability at the very low levels (20 - 25 dB SPL) to 30 dB SPL. We also specify the maximum upper level of 65 dB SP. Furthermore, we specify that sounds are presented on the loudspeaker and that the loudspeaker is always at the same distance from the head (230 cm). dB SPL Levels were measured at the position of the participants head. Lastly, we changed the criteria from arousal to K-complex in NREM sleep, as we want to minimize sleep disturbance. Lastly, as we realized in further pilots that audio thresholds vary across the different nights, and within one night, we will now start each TMR at 30dB SPL and increase until a K-complex/arousal is elicited (maximum 65dB SPL). This is now changed within the manuscript on page 7:

Words are presented on two speakers 100 cm from the head on each side.

And page 8:

Words are presented starting from 30dB SPL via two loudspeakers situated 230 cm from the head of the subject. Sound levels will be increased until a K-complex (NREM), or arousal (REM) is elicited in each sleep stage and then kept at that sound level (NREM) or one below (REM) or to the maximum of 65dB SPL.

It is planned that "Afterward, they fill out a questionnaire about their sleep … and a question about spontaneous, non-experimenter awakenings)."
Will this be cross-validated using the EEG data? Which purpose does this question serve? Whether participants can recall non-experimental awakenings, or whether they had any at all? Please clarify.

This questionnaire will not be analyzed for this registered report, we only clarified this to make transparent that we adjusted the established S-F/A questionnaire, which normally includes the question about any awakening.

"After 3 minutes of stable NREM and REM sleep"
Which NREM sleep stage exactly? N1, N2, N3 ? I guess N3 would be the most obvious choice, however earlier in the report, the authors mentioned N2. Please clarifiy.

We now specify that TMR starts at N2 or N3 (but not N1) sleep. We allow for N2 and N3 to increase feasibility. As N3 decreases across the night, if we limited TMR to N3 this would strongly reduce the chances of doing TMR across four NREM cycles. Awakenings can be done in any NREM sleep stage (N1, N2, N3) as is now also specified within the manuscript see page 8:

Words will be presented for maximally 15 minutes before each awakening after 3 minutes of stable sleep (NREM2/NREM3 or REM) has been reached. Words are presented starting from 40dB SPL via two loudspeakers situated 230 cm from the head of the *subject. Sound level will be increased until a K-complex (NREM) or arousal (REM) is elicited.*

As well as in the supplementary methods page 38:

After at least 3 minutes of stable NREM (N2 or N3) and REM sleep, experimenters will play audio cues for 5 to 15 minutes using two loudspeakers placed at 230 cm from the participants' heads (position kept consistent across participants). Words associated with one specific image category will be used for cueing in each sleep stage (randomly chosen for each participant). Words from the category will be presented randomly every 8,000 to 8,200 ms.

“experimenters will play audio cues for a maximum of 12 minutes at the detected audio threshold using two loudspeakers placed next to the participant’s head.”
What will be the distance between the loudspeaker and the ear? Since the sound pressure level (SPL) decreases quadratically with distance, it is very important that this distance is either always kept constant or recorded for each participant on the night in which the threshold values are determined. This is the only way to ensure that on the second night the distance and thus the sound pressure level is comparable and actually corresponds to the estimated threshold value.

We now specify the setup in more detail on page 38:

Cueing will start at 30dB SPL and increase in 5 dB steps until the participant shows a K-complex (NREM) or arousal (REM). Audio will then be played at the level (NREM) or one step below the level (REM) for the remainder of the sleep cycle. Audio levels will be determined for each cycle as thresholds vary throughout the night. Audio cues will be stopped if participants show a sign of arousal or change into a different sleep stage. The participants will be awoken between 10 – 30s after the last TMR at least 15 minutes into each sleep stage. The protocol for the awakenings is identical to session A.

“In both experimental nights, participants will be instructed to signal if they have a period of lucidity. Dream reports with lucidity will be removed from analyses (score >= 4 on the lucidity scale).”
- How will participants signal periods of lucidity? I guess via voluntary eye movements according to a pre-defined code. However, how can it be assured that participants are actually able to do so during lucid periods. Usually, this requires training even for frequent lucid dreamers. Please clarify! - Please cite literature on lucidity scale.

We agree that without practice, dreamers might not be able to signal lucidity. Therefore we now changed to exclusively use the lucidity rating as an exclusion criterion, and only when a score of 5 is reported (reported after awakening). The lucidity scale is the same as used within (Stumbrys et al., 2013)**. We added this citation now also to the manuscript.**

“The inclusion criteria to participate in the study are … high English language proficiency”.
- How will the proficiency be assessed?
- Wouldn’t it be better to restrict the inclusion criteria to English native speakers? In neurolinguistic studies, it is of great importance whether participants are L1 or L2 learners of a given language. To some extent this also applies to memory consolidation studies.

We agree that including only native English speakers would be preferable. However, this would affect the possibility of reaching our target sample size. Indeed, the study will be performed in the Netherlands, in an international University town, and only recruiting native English speakers (or even native Dutch speakers) in this context would be a too restrictive selection criterion. Therefore, we will use high English language proficiency, as measured by the Boston Naming Test (as specified in Supplementary Table 1). While the language level will potentially influence the performance of the memory task, our main effects of interest (interaction effects) focus on relative changes in performance (from evening to morning/follow up) and takes individual differences into account.

References:

Baird, B., Tononi, G., & LaBerge, S. (2022). Lucid Dreaming Occurs in Activated REM Sleep, Not a Mixture of Sleep and Wakefulness. *Sleep*, zsab294. https://doi.org/10.1093/sleep/zsab294

Hu, X., Cheng, L. Y., Chiu, M. H., & Paller, K. A. (2020). Promoting memory consolidation during sleep: A meta-analysis of targeted memory reactivation. *Psychological Bulletin*, *146*(3), 218.

Konkoly, K., Appel, K., Chabani, E., Mironov, A. Y., Mangiaruga, A., Gott, J., Mallett, R., Caughran, B., Witkowski, S., & Whitmore, N. (2021). *Real-Time Dialogue between Experimenters and Dreamers During rem Sleep*.

Newbury, C. R., Crowley, R., Rastle, K., & Tamminen, J. (2021). Sleep deprivation and memory: Meta-analytic reviews of studies on sleep deprivation before and after learning. *Psychological Bulletin*, *147*(11), 1215. https://doi.org/10.1037/bul0000348

Scarpelli, S., Bartolacci, C., D'Atri, A., Gorgoni, M., & De Gennaro, L. (2019). The Functional Role of Dreaming in Emotional Processes. *Frontiers in Psychology*, *0*. https://doi.org/10.3389/fpsyg.2019.00459

Schoch, S. F., Cordi, M. J., Schredl, M., & Rasch, B. (2019). The effect of dream report collection and dream incorporation on memory consolidation during sleep. *Journal of Sleep Research*, *28*(1), e12754. https://doi.org/10.1111/jsr.12754

Schredl, M. (2018). Functions of Dreaming. In M. Schredl (Ed.), *Researching Dreams: The Fundamentals* (pp. 175–181). Springer International Publishing. https://doi.org/10.1007/978-3-319-95453-0_9

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., & Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, *20*(6), 872–878. https://doi.org/10.1038/nn.4545

Stumbrys, T., Erlacher, D., & Schredl, M. (2013). Testing the involvement of the prefrontal cortex in lucid dreaming: A tDCS study. *Consciousness and Cognition*, *22*(4), 1214–1222. https://doi.org/10.1016/j.concog.2013.08.005

Valli, K., Frauscher, B., Gschliesser, V., Wolf, E., Falkenstetter, T., Schönwald, S. V., Ehrmann, L., Zangerl, A., Marti, I., Boesch, S. M., Revonsuo, A., Poewe, W., & Högl, B. (2012). Can observers link dream content to behaviours in rapid eye movement sleep behaviour disorder? A cross-sectional experimental pilot study. *Journal of Sleep Research*, *21*(1), 21–29. https://doi.org/10.1111/j.1365-2869.2011.00938.x

Voss, U., Holzmann, R., Tuin, I., & Hobson, A. J. (2009). Lucid Dreaming: A State of Consciousness with Features of Both Waking and Non-Lucid Dreaming. *Sleep*, *32*(9), 1191–1200. https://doi.org/10.1093/sleep/32.9.1191

Windt, J. M. (2013). Reporting dream experience: Why (not) to be skeptical about dream reports. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00708

Windt, J. M. (2015). Dreams and Dreaming. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/dreams-dreaming/