nature portfolio

Peer Review File

Sleep selectively and durably enhances memory for the sequence of real-world experiences

Corresponding Author: Dr Brian Levine

Version 0:

Decision Letter:

12th July 2024

Dear Dr Levine,

Thank you once again for your manuscript, entitled "Sleep selectively and durably enhances memory for the sequence of real-world experiences," and for your patience during the peer review process.

Your manuscript has now been evaluated by 3 reviewers, whose comments are included at the end of this letter. Although the reviewers find your work to be of interest, they also raise some important concerns. We remain very interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

As you will see, the reviewers ask for a combination of additional control analyses and clarifications about the methods used. We ask that you address these concerns in full in your revised manuscript.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within two months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

Giacomo Ariani, PhD

Senior Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Memory consolidation, sleep research, EEG

Reviewer #2: Memory consolidation, sleep research, EEG

Reviewer #3: Memory consolidation, sleep research, EEG

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
The authors investigated the differential consolidation of featural versus sequential aspects of an episodic memory formed during a real-world experience (an art tour). Most studies in humans have investigated episodic memory formation in the laboratory environment, often with the use of static stimuli such as word-picture associations. On the contrary, the experimental task used in this study allows participants to navigate in a real-world environment as the encoding unfolds, rendering the experience multimodal and extremely ecological compared to the standard laboratory procedures. Moreover, sleep's benefit on memory after a retention interval is usually observed in the form of decreased forgetting in the sleep group compared to the wake counterpart. In this work, instead, the authors show a selective sleep effect on the sequential aspects of an episodic memory in the form of memory enhancement. Finally, few studies so far have been investigating episodic memory across multiple delayed time points reaching up to 15 months as in the current work. This is a great research design to sutdy how real life episodic memory is formed and retained, and what role sleep has in the process.
The findings reported in this work are the result of multiple studies. The experiments have been cleverly designed and planned, and the main effects were replicated across the different studies, making the observed findings robust and overall convincing. This work contributes to narrowing the current gap between laboratory and naturalistic setting in the investigation of episodic memory formation, as well as the role sleep might play in selectively enhancing some specific features of episodic memory. I am very excited about the findings!

Major considerations:
Retrieval test modality:
1) Not being familiar with this type of memory task, I was wondering whether the mode of inquiry (true vs. false statements) could inadvertently favour the formation of false memories and create unwanted confounds in the data. Is there evidence for/against this in the data?

2) Did you look at the memory change over time and across retrieval types also for hits alone?

3) Did you correct for the unequal number of featural vs sequence item statements presented at the retrieval test, especially at T5?
Sleep group:
4) It wasn't immediately clear to me, while reading the provided material, to what extent the 10 additional participants of the sleep group in study 2B contributed to the results displayed. The text states that they underwent overnight PSG recordings in the sleep lab between T1 and T2. Were they only tested behaviorally at T1 and T2? If so, this should be stated more clearly, and it should be stated clearly where behavioral data of these participants is included in analyses.

5) Correlations between memory change and SWS hallmarks are reported for this sample and are both significantly positive for sequence and featural memory change. Did you also test the statistical difference in correlations' strength for coupled spindles count for sequence vs featural memory change? Behaviorally, what was the T1-T2 change for both retrieval types alone, as well as the difference in overnight retention between retrieval types for this specific sample (N = 49)? If I read tables S3-S5 correctly, the the behavioral finding does not fully hold in this subsample? It would be great if the authors briefly discussed/qualified this somewhere in the paper.

6) Likewise, in the Sleep group subsample with T5 (N = 33), it descripitvely seems like also featural memory increases between T1 and T2. Was this statistically tested?

7) Although the longterm benefit for sequence memory after 15 months is strongly visible and specific to the sleep group, the difference in slope direction at T2 for featural memory seems to vary depending on the sample taken into consideration for the behavioral analyses of the sleep group. A note of caution for interpretation/a note that additional research would help clarify these findings would be a valuable addition to the paper.

Minor comments:
1) How different were the items to be learned from each other in terms of features and art modality (painting, sculpture, miniatures etc…)? Did you control for subjective emotional reactions to the art as a confound?

2) From the methods section, it seems that each art piece item was accompanied by the audioguide providing semantic information about the item, prior to visual inspection by participants. However, in the audioguide example that I listened to (103_n, the following ones weren't working for me), visual examination for encoding the item's features immediately followed the title and was instead preceding further information about the art piece provided by the audioguide. This is a very minor detail, but I would consider inverting the order in the explanation of the tour to make sure the reader understands the tour's procedure as accurately as possible/clarifying why this may have been the case for some items but not others (I was not able to verify).

3) The reported N does not match for the same studies in different parts of the manuscript: Study 1 (N = 53 in the intro and supplementary methods but N = 57 in figure 2a note), Study 2a (NSleep = 39 and NWake = 38 in the general intro but then Nsleep = 44 and Nwake = 46 in the Supplementary Methods and Figure 2c and 2d). The N in figure 2 should probably report the final sample included in the analysis and not the overall sample before data exclusion? Please clarify.

4) It would facilitate getting a quick overview of the main findings if there were asterisks showing where the change whithin, or at least the difference between retrieval types, was significant. For example, in figure 2C Sleep at T4 or 2A at T3 and T4 between retrieval types.

5) Second page of the discussion, second to last line, Study 2a Wake and Study 2a Sleep seem to be inverted regarding the switch in direction across time points?

Reviewer #2:
Remarks to the Author:
The authors report results from human participants recalling a real-world art tour while varying the retention interval and the presence of sleep. Intriguingly, they found that memory for the sequential order of experiences in the tour improved over short delays, which they show relies on sleep. Memory for individual features - which were matched for overall difficulty - were maintained but did not show a raw improvement over sleep. These results show some of the first – and arguably the most convincing – evidence that sleep can outright enhance memory, rather than simply reducing forgetting. The authors also show that sequence memory remains better for those in the sleep group compared with the wake group at much longer intervals (up to 1 month initially and up to 15 months in a follow up sample), showing that differences in initial sleep vs. wake have consequences well after both groups have gained equivalent amount of sleep. Overall, I found the paper exceptionally well-written, and the data are mostly convincing. That said, I have a few comments and questions I hope the authors can clarify.

Larger issues:
P. 12: "Given our prior results concerning the behavioural dissociation of sequence and featural memory change overnight, these measures are visualized separately in relation to coupled and uncoupled spindles (Figure 4). Overnight sequence memory enhancement was uniquely associated with coupled spindles, and significantly more so than uncoupled spindles. Featural memory showed a similar pattern, although the correlations between coupled versus uncoupled spindles were not significantly different (see Figure 4)." The issue generally with single-night sleep physiology correlations is that there is no baseline comparison, so one cannot say how memory specifically correlated with physiology as a results of learning (and in support of post-sleep memory). Therefore, correlations between memory and spindles / higher SO-spindle coupling, etc. could simply result from an underlying factor. For example, let's say individuals who have better memory, or higher IQ, simply have more spindles or higher SO-spindle coupling, which is something that has been found (though debated) (Ujma, 2021, Sleep Spindles & Cortical Up States). This would mean they have better memory and also more spindles, but the spindles may not be contributing at all to the memory performance. Overall, the physiology results nicely complement their behavioral findings, but I believe the authors should point out this important caveat in their discussion.
P. 10: "While the Study 2a long-term (T5) follow-up participants were composed of a subset of the full sample, they were demographically similar to the main sample (see Supplementary Materials) and there was no a priori reason to expect that these individuals would show a disproportionate advantage for sequence over featural memory." Did this subset differ from the original groups in memory performance at any of the intervals? This might suggest they are a biased sample. It does appear visually that featural memory in the Study 2a Sleep group increases across the 12 hours (Fig 3), which it does not appear to do as strongly in the main group (Fig 2), but it isn't clear if this is significant.
P. 18: In the T5 test, did the sequence and/or feature questions have the EXACT same question (comparing the same elements) or were they mixed and recombined to ask about different parts of the tour (using the same parts from before)? That is, was it the case that, e.g., a sculpture called "Sails" was compared in time to the shuffleboard game, but that at the T5 test, the sculpture called "Sails" was compared to the sculpture called "One Nine North" (a different part of the tour), or was the EXACT same comparison made (e.g., "Sails" vs. shuffleboard again)? If the exact, there could be testing effects (e.g., Rowland, 2014) that should be mentioned as a caveat here, but if not, that is less likely.
P. 21: "This procedure involved building a time window of 4 seconds around the negative peaks of the detected slow waves. The spindles were identified as coupled slow wave-spindle (SW-SP) complexes if the spindle peak latency fell within a 4-second time window of the slow wave peak detected in the corresponding slow wave channel. Alternatively, spindles were marked as uncoupled if the spindle peak latency occurred outside the 4-second time window." In my opinion, a 4-second window is too wide, as most slow waves in SWS (and outside of deep anesthesia) occur at a higher frequency that would be more amenable to a shorter window. Even a 3-second window (~ a 0.66 Hz slow wave cutoff) would capture some spindles that start in the next slow wave. The distribution even appears in their Figure S6 – almost all coupled spindles are within the first +/- 1 second. The danger here is that one could be frequently capturing spindles that occur after the slow wave, which may or may not meet the definition of a slow wave. Personally, I think the better choice would be to use a method where they find the time of troughs between verified slow oscillations and the timing of the spindle start and only consider it to be coupled if it occurs between these troughs. This avoids the issue of needing to place any frequency cutoff to begin with, and it also avoids the issue of spindles occurring within the next slow wave. I see that they use an established method (from the Baena Perez studies, who is a co-author), so I am curious if there is anything from those studies that would allay this concern. Please either explain why this concern is misplaced or consider

changing the method for coupling detection.

Moderate issues:
Fig 2 caption and in manuscript: The authors report what qualitatively looks to be a negatively accelerating forgetting function, but I think they could actually fit it. They mention a linear decline against log spacing, so they could test the fit for y = log(x) +b, where y is memory and x is retention interval, and report the R-squared value. I imagine this will break for the sequential memory case, but I believe that is fine, as it is clear sequential memory differs from the typical forgetting curve, and they aren't making an argument that sequential memory fits this pattern anyway.
P. 14: “Specifically, spindles coupled to slow waves uniquely predicted overnight memory enhancement, aligning with contemporary models of hippocampal-neocortical communication in sleep-related mnemonic processing (17, 24) (for review, see (19)). Contrary to our hypothesis and behavioural evidence for sleep’s selective effect on sequence memory, SWS duration and neurophysiology predicted both sequence and featural memory change – although there was evidence for greater specificity to spindle-slow oscillation coupling for sequence relative to featural memory.” Despite my largest comment on needing to weaken the physiology interpretations, the authors’ own work may be useful in speculating about something else here. SO-spindle coupling changes in aging (Helfrich et al., 2018, Neuron), and given their prior work on aging in the current paradigm (Diamond & Levine, 2020, Psych Sci), there could be room for meaningful speculation about performance on this task in aging and how it relates to changes in sleep physiology.
P.21: “Consideration of fast and slow spindles separately did not add new information to hypothesis-driven analyses.” Is this a statistical model comparison or something else? Please clarify.

Small issues:
P. 4: “Sleep may in some cases improve non-episodic memory tasks like motor sequence learning or statistical learning (28, 29) (but see (30))…” I recognize the authors may be trying to avoid too strong a claim (in attempts to avoid offending anyone), but I would state the lack of raw procedural learning improvements across sleep more strongly. The only studies to find improvements (e.g., Walker et al., 2002) have clear fatigue confounds (from the context of today, after the methods measuring fatigue effects have become clearer). Not that they would need this to be the case, but I think this ultimately makes their findings of raw improvement more novel.
P. 4: “One idea sleep serves as “memory triage” system, tagging only the most relevant pieces of information to persist over time” Please fix this sentence.
Paragraph spanning p. 4-5: It isn’t ‘once-presented’, but the sequential memory paradigm in Petzka et al. (2022, Current Biology) seems relevant here.
P. 5: The authors may find some of the following studies helpful from the targeted memory reactivation literature, which examine memory over longer periods of time: Knötzele et al. (2023, Sci Rep); Neumann et al. (2020, Sci Rep).
P. 8: “The preferential effect of sleep on sequence versus featural memory could not be…” Up to this point, they’ve only shown a 24-hr increase, which may or may not be attributable to sleep. Please fix to not (yet) focus on sleep.
P. 9: “Here, we first replicated the findings observed in Study 1 in the Sleep condition, where...” -> “whereas”
P. 11: “(N = 49 after exclusions, see Supplementary Methods” Please add ‘)’.
P. 13: “Sleep’s influence on the fate of our memories has been studied for at least a century.” The authors should cite a study supporting this claim.

From the Supplement:
Table S3 – Were the stage correlations corrected for multiple comparisons, or was there an a priori hypothesis predicting SWS duration would lead to behavioral improvements? These are not the featured sleep physiology analyses, so I do not believe these carry that aspect of the paper, but the authors should state either way.

Reviewer #3:
Remarks to the Author:
In this study Diamond and colleagues investigate to what extent sleep transforms multidimensional real-world experiences (i.e., art tour related memories). They show that sleep selectively enhances sequential aspects of memories, as compared to perceptual details. Importantly, this effect grew stronger over up to a year post-encoding and was not observed after wakefulness. In addition, memory enhancement was associated with slow-wave-sleep duration and SO-spindle coupling. These findings are highly intriguing and shed light on several important features of memory consolidation (which memory aspects preferentially benefit from sleep? How durable are those effects?). While I am generally very impressed with the high quality of the work, I do have some questions ± the EEG analysis.

Comment 1: Why is there a significant discrepancy in the number of detected events? The number of SOs detected appears excessively high. As shown in Fig. S4, the mean of detected SOs is approximately 7000. Assuming participants spent around 5 hours in NREM sleep, this would imply that SOs occurred roughly every 2.5 seconds, which seems improbable, especially during sleep stage N2. Conversely, the number of spindles detected seems quite low (mean around 450) compared to what has been reported in the literature. Could you please provide more details on the spindle detection methodology?

Comment 2: Related to the above, Fig. 4 suggests that approximately half of the spindles were coupled with SOs. This ratio appears quite high compared to previous findings (e.g., Mühlroth et al., 2019; Schreiner et al., 2021). This may be due to the authors' use of a very liberal criterion for identifying coupled events, where spindle peaks within a 4-second window around an SO downstate are considered coupled. I find it unlikely that a spindle occurring 3 seconds after an SO could be considered coupled. I recommend applying a more conservative criterion, such as ±1.5 seconds, to identify coupled events more accurately.

Comment 3: The authors used different electrodes to detect SOs (frontal) and spindles (central), which makes sense when analyzing these oscillations independently. However, for coupled events, I recommend that the authors focus on a single site (see, e.g., Helfrich et al., 2018).

Comment 4: In the methods section the authors state that the lag variance between SOs and spindles was calculated for each individual as a measure of coupling strength. Could the authors report the results of this analysis? Same goes for uniformity tests (Rayleigh and V-test).

Comment 5: was the correlation between SWS duration and memory change corrected for multiple comparisons?

Version 1:

Decision Letter:

Our ref: NATHUMBEHAV-24062188A

7th October 2024

Dear Dr Levine,

Thank you for submitting your revised manuscript "Sleep selectively and durably enhances memory for the sequence of real-world experiences" (NATHUMBEHAV-24062188A). It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Giacomo Ariani, PhD
Senior Editor
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I thank the authors for the careful responses to my comments and further explanations regarding the design. I appreciate that parts of the discussion have been toned down, and further points have been added, particularly regarding the progression of featural memory over time for the separate study samples. I think the behavioral effects on sequence preservation in real-world memories are highly exciting, and I am looking forward to seeing this work published. I have no further major comments.

Reviewer #2 (Remarks to the Author):

First, I want to commend the authors for a very nice revision. I appreciate their thoughtful comments regarding many points, including the new physiology analyses. I think the paper should be published, but I wanted to follow up on a few points from their response letter.

#1
“Any effect owing to the reuse of items would have been minimized due to the amount of time (i.e., more than one year) that passed between T4 and T5. This conjecture is supported by the fact that the forgetting curves (Figure 3) followed the expected trajectory considering the data observed for T1-T4. Most importantly, the critical conclusions from the T1-T5 analysis concerned a comparison of sleep vs. wake groups, who were exposed to the same test items. Therefore, any testing effect – if present – was controlled.”

Here I disagree. It is not at all clear to me that the effect of reuse would be so minimal as to not be a potential factor at T5.

In one respect, it seems the concern is being minimized due to the time scale. Such timescales resemble those in some large spacing effect studies (that involve retrieval tests) such as Cepeda et al. (2009) (Exp 2). That study is not particularly similar to this one in many ways, and Cepeda and colleagues provide feedback at the Session 2 test, but I mention it to show the profound effect testing on something after a delay of days to a few months can have on later memory of similar timescales (~ a year).

The authors presumably did not include feedback for their T1-4 tests (because there would have been no reason for them to do so -- but please state whether this is incorrect). Given that successful retrieval produces strengthening (e.g., Rowland, 2014), this means that any differences in performance due to sleep versus wake -- which indeed they observe and is a highlight of this paper -- would lead to a different number of items getting strengthened. For instance, about 85.5% of sequence items were remembered in the 12-hr sleep condition (T2), whereas about 78% were remembered in the 12-hr wake condition. This means that, of the items that were on that T2 test that were included in the later T5 test, an average of 7% MORE items would be strengthened in the sleep condition the wake condition, which could have an impact on a later test. The same principle applies for any tests with differential

performance (e.g., T3 & T4) (and would apply in the other direction if it were better after wake).

Also, although I appreciate their logic, I disagree that simply following the expected downward trajectory following from T1-4 means there's no cause for concern -- there could easily be a confounding effect of taking a prior test alongside forgetting (over the course of another year) that would lead to the observed results.

Therefore, it still seems to me that – because successful retrieval can have a profound influence on memory, whereas unsuccessful retrieval without feedback seems to NOT have this effect (e.g., Kornell et al., 2011) – the T5 results could be influenced by differential performance across sleep versus wake conditions on T1-4. If the authors still disagree or if I have misunderstood something, of course they should state this in the response. Otherwise, it seems worth highlighting as a limitation for this very long interval data point.

#2
Regarding the Ebbinghaus point, I first wanted to apologize that I misstated the function. I believe the log fit function would be y=a-b*log(x). When I estimate their memory results from the figure and plug them into code to perform curve fitting, I see an R-squared of 0.91 (or adjusted R-squared of 0.86, if the authors prefer that value). Here is the sample MATLAB code I used (requires the Curve Fitting Toolbox):

```
diam_ri=[1;24;168;672]; %define RI and mem values
diam_s1=[83;80.5;74.5;70.5];
ft=fittype('a-b*log(x)'); %define function fit -- change for different functions
options=fitoptions(ft);options.StartPoint=[1 1];options.Lower=[0 0]; % initializations
[fit_mod,gof]=fit(diam_ri,diam_s1,ft,options); %do fit
R_sq=gof.rsquare %print rsquare to terminal
Adj_R_sq=gof.adjrsquare %print adj rsquare to terminal
```

Regardless of the exact memory values (clearly my estimations were not exact), I think this fit should be much stronger than 0.197 – even a linear fit run by substituting the following line above …

```
ft = fittype('a-b*x');
```

…produced an R-squared of 0.81.

Note that if they prefer other negatively accelerating functions, I believe the exponential is:

```
ft = fittype('a*exp(-b*x)');
```

and the power law is:

```
ft = fittype('a*x^-b');
```

Whether or not the authors decide to report any of these results is of course up to them, but I wanted to add this correction.

#3
To be clear, were the exploratory slow and fast spindle results significantly different from each other? I also don't know if there's enough information from the new sentence ('interacted with Time') to glean how slow spindles predicted memory. Please make these effects a bit clearer.

Reviewer #3 (Remarks to the Author):

The authors addressed all my concerns. I have nothing more to add.

Version 2:

Decision Letter:

Dear Dr Levine,

We are pleased to inform you that your Article "Sleep selectively and durably enhances memory for the sequence of real-world experiences", has now been accepted for publication in Nature Human Behaviour.

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.springernature.com/gp/open-

research/policies/journal-policies">self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,

Giacomo Ariani, PhD
Senior Editor
Nature Human Behaviour

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian
http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Response to the reviewers

Reviewer #1:
The authors investigated the differential consolidation of featural versus sequential aspects of an episodic memory formed during a real-world experience (an art tour). Most studies in humans have investigated episodic memory formation in the laboratory environment, often with the use of static stimuli such as word-picture associations. On the contrary, the experimental task used in this study allows participants to navigate in a real-world environment as the encoding unfolds, rendering the experience multimodal and extremely ecological compared to the standard laboratory procedures. Moreover, sleep's benefit on memory after a retention interval is usually observed in the form of decreased forgetting in the sleep group compared to the wake counterpart. In this work, instead, the authors show a selective sleep effect on the sequential aspects of an episodic memory in the form of memory enhancement. Finally, few studies so far have been investigating episodic memory across multiple delayed time points reaching up to 15 months as in the current work. This is a great research design to study how real life episodic memory is formed and retained, and what role sleep has in the process.

The findings reported in this work are the result of multiple studies. The experiments have been cleverly designed and planned, and the main effects were replicated across the different studies, making the observed findings robust and overall convincing. This work contributes to narrowing the current gap between laboratory and naturalistic setting in the investigation of episodic memory formation, as well as the role sleep might play in selectively enhancing some specific features of episodic memory. I am very excited about the findings!

We thank Reviewer 1 for these positive comments.

Major considerations:
Retrieval test modality:
1) Not being familiar with this type of memory task, I was wondering whether the mode of inquiry (true vs. false statements) could inadvertently favour the formation of false memories and create unwanted confounds in the data. Is there evidence for/against this in the data?

It is certainly the case that memory test stimuli – whether true or false – can influence memory at the time of retrieval (for example, see literature on retrieval-mediated learning and output interference). That said, it is standard practice to incorporate both target (old/true) and lure (new/false) stimuli in recognition memory tests to account for response bias (i.e., someone responding 'true' to all items would score perfectly if lures were not taken into consideration). Thus, recognition memory performance is typically operationalized – as it is in our study – as *discrimination* between old (or true) vs. new (or false). In our pilot testing and preliminary item analysis, we selected true and false memory cues from a large pool of statements that reflected all tour components equally and were balanced for difficulty.

3560 Bathurst St. T: 416 785 2500
Toronto, Ontario W: www.baycrest.org
Canada M6A 2E1

Baycrest is fully affiliated with the University of Toronto

Our design included controls for potential confounds. Within participants, recognition responses for featural and sequence items pertained to the same set of stimuli (i.e. artwork in the tour). Any confound shared by those stimuli—such as prior exposure to false statements—would apply equally to featural and sequence judgements. Our randomization of participants to sleep vs. wake groups in Study 2a ensured that participants in each group were exposed to identical encoding and recognition conditions such that potential confounds in those conditions could not account for group differences.

Finally, we statistically assessed the differential contribution of item type on overall statement accuracy across participants (see Supplementary Information and Figure S1) and we included statement validity (true vs. false) in our main model of sequence vs. featural memory change. As reported on p. 9 (results), statement validity did not emerge as significant (alone or in combination with other factors) in these analyses, neither in item-level nor participant-level analyses (see SI p. 8) providing further support for our contention that trial accuracy is not affected by whether a statement is true or false.

2) Did you look at the memory change over time and across retrieval types also for hits alone?

As mentioned above, we assessed whether memory performance (and its interaction with time and retrieval type) were different for true (i.e. hits / misses) vs. false (i.e. correct rejections / false alarms) items – it was not. Given this result, and that there is no *a priori* theoretical reason to explore true vs. false items separately, we did not probe this further in the paper.

3) Did you correct for the unequal number of featural vs sequence item statements presented at the retrieval test, especially at T5?

Our measures of featural and sequence performance were averaged over items, correcting for the unequal number of featural and sequence items. In the case of Tests 1-4, imbalances were minimal and identical across tests (i.e. 34 featural vs. 35 sequence items). T5 was designed to include 100 items (50 featural and 50 sequence) from the original four test forms in equal proportion. However, as discussed on p. 18 of the manuscript, 49 featural and 42 sequence remained after exclusions for performance and a coding error. In any case, any conclusions concerning the effect of sleep (vs. wake) at T5 could not be accounted for by this difference, as both sleep and wake groups got the same test.

Most importantly, as reported in the Results (p. 7), featural and sequence memory cues were balanced for difficulty overall (i.e. no main effect of memory type).

Sleep group:
4) It wasn't immediately clear to me, while reading the provided material, to what extent the 10 additional participants of the sleep group in study 2B contributed to the

results displayed. The text states that they underwent overnight PSG recordings in the sleep lab between T1 and T2. Were they only tested behaviorally at T1 and T2? If so, this should be stated more clearly, and it should be stated clearly where behavioral data of these participants is included in analyses.

We have now clarified on p. 7 and 11 of the main manuscript and p. 4 of the Supplementary Information that the 10 additional participants in Study 2b underwent the same procedures as the participants in the sleep group in Study 2a, including behavioral testing across all time points. Their behavioral data were included in the analyses of PSG data that were focused on T1-T2. We did not include their data in the Study 2a behavioral analysis because they were not randomized to sleep vs. wake, as was the case for all of the Study 2a participants.

Figure caption, p. 7

> ***Study 2b*** *(not depicted) comprised the Study 2a Sleep group plus 10 additional participants who underwent the same procedures as the sleep group (total N = 49) to increase power for individual differences analyses linking sleep neurophysiology to memory change.*

Results, p. 11

> *Study 2b participants – including the Sleep group from Study 2a and additional participants recruited for individual differences analyses – underwent overnight polysomnography recordings in the sleep laboratory between T1 and T2 (and testing at T3 and T4; N = 49 after exclusions, see Supplementary Methods).*

Supplementary Information, p. 4

> *Eighteen additional participants were recruited to increase the sample size of the sleep group for testing individual differences, prior to halting of the study due to the onset of the COVID-19 pandemic. Although these participants underwent the same procedures as the Study 2a Sleep group, they were not included in behavioral analyses of T1-T4 as they were not randomly assigned to the sleep vs. wake groups.*

Seven of those 10 participants returned for T5 testing. In preparing this response, we recognized that we should have confined the T1-T5 analysis to the Study 2a participants, as this analysis primarily concerned the comparison of sleep vs. wake. We therefore reran the T1-T5 analyses excluding those participants (now with *N*'s of 26 and 23 for the Sleep and Wake groups, respectively), which had no appreciable effect on the findings. The new results are reported on p. 10:

> *As expected, overall memory performance declined over time ($\chi^2(1) = 169.73, p$*

< .001), but it remained above chance at T5, M = 0.63, SD = 0.03). As reported for the Study 2a Sleep group, a sleep-related advantage for sequence over featural memory performance from T1 to T4 was evident among those who completed T5 (see Supplementary Methods). This mnemonic enhancement conferred by one night of sleep (versus wake) held at 15 months post-encoding (Group X Time [T1, T5] interaction; $\chi^2(1) = 5.57$, p = .018). Although the three-way Group X Time X Retrieval Type interaction was not significant ($\chi^2(1) = 0.67$, p = .411), hypothesis-driven planned comparisons confirmed sequence memory performance from T1 to T5 was preserved by sleep relative to wake (i.e., the T1-T5 decline in sequence memory was greater in the Wake versus Sleep group; z = 2.24, p = .025, d = 0.638). There were no group differences in T1-T5 change in featural memory performance (z = 1.09, p = .274, d = 0.432).

We have also modified Figure 3 with a new caption and sleep panel to reflect this reduced sample (p. 11):

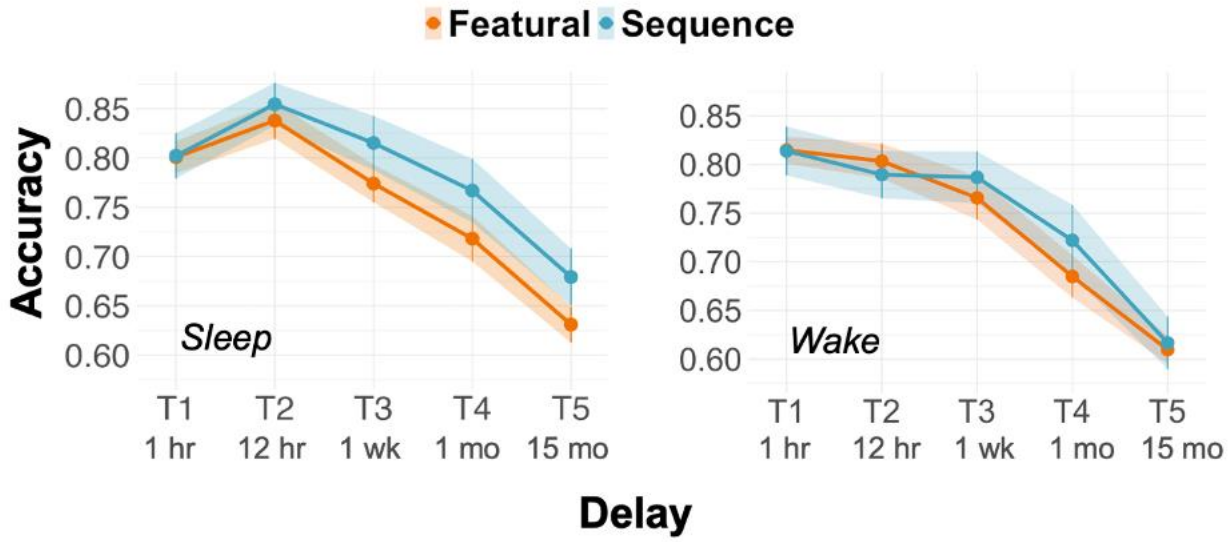

Figure 3. *Sequence versus featural memory accuracy from T1 to T5 in a subset of the Study 2a participants who completed the ancillary T5 testing (N's = 26 and N = 23 for Sleep and Wake, respectively)."*

5) Correlations between memory change and SWS hallmarks are reported for this sample and are both significantly positive for sequence and featural memory change. Did you also test the statistical difference in correlations' strength for coupled spindles count for sequence vs featural memory change?

The correlations between coupled spindles and featural / sequence performance are not statistically different from each other. This is now added to the Figure 4 caption:

> ***Figure 4.*** *Study 2b. Sequence memory change (T2 – T1) is significantly associated with coupled (r (47) = 0.42, p = .003) but not uncoupled spindles (r (47) = 0.01, p = .921) in SWS. These correlations are statistically different from each other (Steiger's z = 2.24, p = .029). Similarly, featural memory change (T2 – T1) is related to coupled (r (47) = 0.34, p = .018) but not uncoupled spindles (r (47) = 0.11, p = .451) in SWS. These correlations are not statistically different from each other (Steiger's z = 1.21, p = .231). Considering coupled spindles alone, the correlations for sequence vs. featural memory change are not statistically different from each other (Steiger's z = 0.52, p = .61).*

Behaviorally, what was the T1-T2 change for both retrieval types alone, as well as the difference in overnight retention between retrieval types for this specific sample (N = 49)? If I read tables S3-S5 correctly, the behavioral finding does not fully hold in this subsample? It would be great if the authors briefly discussed/qualified this somewhere in the paper.

The mean overnight change in featural (-0.002) and sequence (0.026) memory for this sample are now reported in Table S2. These scores are comparable to those observed for the Study 2a sleep group (which includes 39 of the 49 participants included in Study 2b).

As noted by the reviewer, the Time (T1-T2) X Retrieval Type interaction was not significant in the models including polysomnography data, although the Study 2b subsample (N=49) does in fact replicate the main behavioural findings reported for Study 2a, using the same methods: considering all four time points, the Time X Retrieval Type interaction was significant (as in Study 1 and Study 2A), and the estimated marginal means from this model show that sequence memory significantly increased overnight whereas featural memory remained flat. This is now reported in the results on p. 11.

> *Study 2b participants – including the Sleep group from Study 2a and additional participants recruited for individual differences analyses – underwent overnight polysomnography recordings in the sleep laboratory between T1 and T2 (and testing at T3 and T4;N = 49 after exclusions, see Supplementary Methods). As in Study 1 and 2A, the Time X Retrieval type interaction - with memory for sequence preserved relative to features from T1-T4—was significant, (χ^2(1) = 23.41, p < .001), with sequence memory significantly improving overnight (z = 2.20, p = .026, d = 0.36**);** as in Study 2A, featural memory remained flat overnight, z = 0.09, p = .927, d = 0.03).*

We addressed the overnight preservation of featural memory in Study 2 relative to Study 1 in the original discussion, now with added emphasis on the interaction involving T1-T4 (p. 15):

This unexpected overnight protection of featural memory may be attributable to the fact that Study 2a participants slept shortly after encoding (i.e., in less than 2 hours versus up to 15 hours in Study 1), suggesting that sleep may play a passive, interference-reducing role in protecting atemporal associations in contrast to an active strengthening of sequential associations. While the Study 2a and 2b sample sizes were large relative to related studies (62, 68-71–82) and adequately powered to replicate the effects of Study 1, the better-than-expected featural memory performance at T2 meant that the interactions between Retrieval Type and Group (Study 2a; Sleep versus Wake) and neurophysiology (Study 2b) were attenuated; effects were instead observed for overall memory enhancement, although the critical Time X Retrieval Type interaction—with sequence and featural forgetting functions crossing after a night's sleep— held in all three groups (between T1 and T2 in Study 1 and Study 2a Sleep, and between T2 and T3 in Study 2a Wake).

6) Likewise, in the Sleep group subsample with T5 (N = 33), it descriptively seems like also featural memory increases between T1 and T2. Was this statistically tested?

The reviewer is correct that featural (and sequence) memory increased overnight in this subsample (not initially tested, but now reported; see below). However, T5 analysis was specifically focused on the evolution of the featural/sequence dissociation over long timescales, rather than the T1-T2 effect. Accordingly, the Time (T1-T4) X Retrieval Type interaction for this sample is now referenced on p. 10 and reported in the Supplementary Methods.

Results, p. 10

As reported for the Study 2a Sleep group, a sleep-related advantage for sequence over featural memory performance from T1 to T4 was evident among those who completed T5 (see Supplementary Methods). This mnemonic enhancement conferred by one night of sleep (versus wake) held at 15 months post-encoding (Group Time [T1, T5] interaction; $\square^2(1) = 5.57$, $p = .018$. Although the three-way Group Time Retrieval Type interaction was not significant ($\square^2(1) = 0.67$, $p = .411$), hypothesis-driven planned comparisons confirmed sequence memory performance from T1 to T5 was preserved by sleep relative to wake ($z = 2.24$, $p = .025$, $d = 0.638$). There were no group differences in T1-T5 change in featural memory performance ($z = 1.09$, $p = .274$, $d = 0.432$). While the Study 2a long-term (T5) follow-up participants were composed of a subset of the full sample, they were indistinguishable from the main sample based on either demographics or memory performance (see Supplementary Information) and there was no a priori reason to expect that these individuals would show a disproportionate advantage for sequence over featural memory.

We included a new section in the Supplementary Information (p. 3) devoted to the comparison of this subsample to the original Study 2a sample, including the overnight changes in sequence and featural memory. We elaborate on this matter below in our response to Reviewer 2 (point 2).

> ***Study 2a participants who completed the T5 assessment were representative of the full Study 2a sample***
> *The ancillary T5 assessment to assess remote memory for the Baycrest tour was completed by 49 of the original 77 Study 2a participants (26 Sleep and 23 Wake). We compared those who completed this assessment to the original Study 2a sample on demographics and memory performance. Those who completed the T5 assessment did not statistically differ in age ($t(65) = 0.41$, $p = .685$) or years of education ($t(61) = 0.30$, $p = .768$) from the 28 Study 2a participants who did not participate in this long-term follow-up test. Considering memory performance, we compared participants who completed T5 to those who did not (i.e. Attrition Group) in an ANOVA that included effects of Group (sleep vs. wake), Time (T1-T4), and Memory Type (featural vs. sequence). There was no significant effect of Attrition Group ($\chi^2(1) = 0.03$, $p = .862$), nor were there any significant interactions between Attrition Group and the other factors ($\chi^2(1)$s > 0.72, p's $> .093$). Finally, the critical Time (T1-T4) X Retrieval Type (featural vs. sequence) interaction observed in Studies 1 and 2a held for the Sleep group who completed the T5 test, even though both sequence and featural memory improved from T1 to T2 (sequence: $z = 3.03$, $p = .002$, $d = 0.986$; featural: $z = 2.08$, $p = .038$, $d = 0.545$). In sum, participants who completed the T5 test were representative of the full sample in terms of demographics, memory performance, and the Sleep group's relative preservation of sequence memory over time.*

7) Although the long-term benefit for sequence memory after 15 months is strongly visible and specific to the sleep group, the difference in slope direction at T2 for featural memory seems to vary depending on the sample taken into consideration for the behavioral analyses of the sleep group. A note of caution for interpretation/a note that additional research would help clarify these findings would be a valuable addition to the paper.

We agree with this point, and had addressed the variation in T1-T2 featural memory change across samples in the discussion, p. 15:

> *We speculate that design differences mitigated the dissociation of sequence and featural memory as assessed at T2 in Studies 2a and 2b…This unexpected overnight protection of featural memory may be attributable to the fact that Study 2a participants slept shortly after encoding…the better-than-expected featural memory performance at T2 meant that the interactions between Retrieval Type and Group (Study 2a; Sleep versus Wake) and neurophysiology (Study 2b) were attenuated.*

We have now included a note about the need for future research on this point in the discussion on p. 17.

> *The persistence of sequence memory for a single experience up to a year raises the question of whether it comes to rely on cortical mechanisms, or rather reflects an enduring role for the hippocampus in remote episodic memory access and sequential organization (54, 88). Moreover, further research is needed to assess the effects of arousal or subjective engagement with the memoranda, the timing of sleep relative to encoding (discussed above), the temporal evolution of memory transformation over multiple sleep-wake cycles, the retrieval cue methodology, and the interactions of these factors.*

Minor comments:
1) How different were the items to be learned from each other in terms of features and art modality (painting, sculpture, miniatures etc…)? Did you control for subjective emotional reactions to the art as a confound?

The Baycrest tour items were very diverse in modality, colour, theme, etc. (see Figure 1 and the list of test true/false statements and our OSF page at https://osf.io/bxm5w/). Although we controlled arousal by excluding provocative items, we did not assess subjective reactions, which would necessarily be governed by one's personal history or interest in the material, among other factors. Such variability is inherent to any study, particularly one involving diverse, naturalistic stimuli. Although investigating or controlling for such factors was beyond the scope of our study, we made note of the importance of arousal and subjective engagement in our edits to the discussion (p. 17, see above).

2) From the methods section, it seems that each art piece item was accompanied by the audioguide providing semantic information about the item, prior to visual inspection by participants. However, in the audioguide example that I listened to (103_n, the following ones weren't working for me), visual examination for encoding the item's features immediately followed the title and was instead preceding further information about the art piece provided by the audioguide. This is a very minor detail, but I would consider inverting the order in the explanation of the tour to make sure the reader understands the tour's procedure as accurately as possible/clarifying why this may have been the case for some items but not others (I was not able to verify).

We thank the reviewer for highlighting this discrepancy. The audio guide directed participants' attention to a given item, prompted visual inspection in a set interval, then additional description or semantic information. This has now been clarified in the General Methods section on p. 18:

Each track directed participants to a given item (e.g., painting, sculpture, photograph), instructed participants to examine the item (followed by 10 seconds of silence), and then provided information about that item (e.g., the artist's name, the medium). The audio guide then directed participants to walk to the next item…

Thank you also for flagging the OSF playback issue for some of the audio files - we are not yet able to rectify this problem (it's specific to OSF), but have now included in the OSF repository a separate Dropbox link to all the audio files.

3) The reported N does not match for the same studies in different parts of the manuscript: Study 1 (N = 53 in the intro and supplementary methods but N = 57 in figure 2a note), Study 2a (NSleep = 39 and NWake = 38 in the general intro but then Nsleep = 44 and Nwake = 46 in the Supplementary Methods and Figure 2c and 2d). The N in figure 2 should probably report the final sample included in the analysis and not the overall sample before data exclusion? Please clarify.

We thank the reviewer for pointing out this discrepancy. We have changed the caption in Figure 2 (p. 8) to refer to the final sample after data exclusions.

Figure 2. *Sequence (blue) versus featural (orange) memory accuracy across T1-T4, approximately logarithmically spaced across time to fit a negatively-accelerated forgetting function. Dots depict means, vertical lines and shaded regions depict between-subjects standard errors.* ***A.*** *Study 1 (N = 53).* ***B.*** *Accuracy change from T1-T2 in Study 1: thin lines depict individual Study 1 participants' memory, and circles and thick lines depict group averages; sequence change and featural change were not significantly correlated (r(51) = .002, p = .986).* ***C.*** *Study 2a Sleep Group (N = 39), who slept in the sleep lab between T1 and T2.* ***D.*** *Study 2a Wake Group (N = 38), who did not sleep between T1 and T2.*

4) It would facilitate getting a quick overview of the main findings if there were asterisks showing where the change within, or at least the difference between retrieval types, was significant. For example, in figure 2C Sleep at T4 or 2A at T3 and T4 between retrieval types.

We have now added in the asterisks as requested to Figure 2A and C, and adjusted the figure caption accordingly.

5) Second page of the discussion, second to last line, Study 2a Wake and Study 2a Sleep seem to be inverted regarding the switch in direction across time points?

This error has been fixed.

…between T1 and T2 in Study 1 and Study 2a Sleep, and between T2 and T3 in Study 2a Wake.

Reviewer #2:
Remarks to the Author:
The authors report results from human participants recalling a real-world art tour while varying the retention interval and the presence of sleep. Intriguingly, they found that memory for the sequential order of experiences in the tour improved over short delays, which they show relies on sleep. Memory for individual features - which were matched for overall difficulty - were maintained but did not show a raw improvement over sleep. These results show some of the first – and arguably the most convincing – evidence that sleep can outright enhance memory, rather than simply reducing forgetting. The authors also show that sequence memory remains better for those in the sleep group compared with the wake group at much longer intervals (up to 1 month initially and up to 15 months in a follow up sample), showing that differences in initial sleep vs. wake have consequences well after both groups have gained equivalent amount of sleep. Overall, I found the paper exceptionally well-written, and the data are mostly convincing. That said, I have a few comments and questions I hope the authors can clarify.

We appreciate these comments.

Larger issues:

1. **P. 12: "Given our prior results concerning the behavioural dissociation of sequence and featural memory change overnight, these measures are visualized separately in relation to coupled and uncoupled spindles (Figure 4). Overnight sequence memory enhancement was uniquely associated with coupled spindles, and significantly more so than uncoupled spindles. Featural memory showed a similar pattern, although the correlations between coupled versus uncoupled spindles were not significantly different (see Figure 4)." The issue generally with single-night sleep physiology correlations is that there is no baseline comparison, so one cannot say how memory specifically correlated with physiology as a results of learning (and in support of post-sleep memory). Therefore, correlations between memory and spindles / higher SO-spindle coupling, etc. could simply result from an underlying factor. For example, let's say individuals who have better memory, or higher IQ, simply have more spindles or higher SO-spindle coupling, which is something that has been found (though debated) (Ujma, 2021, Sleep Spindles & Cortical Up States). This would mean they have better memory and also more spindles, but the spindles may not be contributing at all to the memory performance. Overall, the physiology results nicely complement their behavioral findings, but I believe the authors should point out this important caveat in their discussion.**

We agree with the reviewer that the SO-spindle coupling likely reflects pre-existing individual differences rather than a consequence of tour encoding, per se. We have rewritten the discussion to emphasize that—whatever the underlying cause of the relationship—it pertains to sequence memory to a greater degree than featural memory (p. 14).

> *In accordance with active, as opposed to passive or 'permissive' theories of sleep-dependent transformation or consolidation (12, 19), we found that the duration of SWS (but not other sleep stages) and SWS's neural hallmarks, including slow waves and spindles, were associated with overnight memory change. Specifically, spindles coupled to slow waves uniquely predicted overnight memory enhancement, aligning with contemporary models of hippocampal-neocortical communication in sleep-related mnemonic processing (17, 26) (for review, see (19)). This study was not designed to distinguish the directionality of the relationship between the learning episode itself and inter-individual characteristics in sleep physiology. On the one hand, learning manipulations can induce changes in sleep spindles and slow oscillations (61–63). On the other hand, the Baycrest Tour was—by design—just one event in participants' day prior to sleep. There is evidence that inter-individual differences in learning, cognition, and intelligence are related to sleep physiology (28, 63–65), particularly spindle-slow oscillation coupling (67). Whatever the direction of the relationship, our results demonstrate that the interplay of spindles and slow oscillations supports memory transformation, particularly for sequences of everyday events.*

2. P. 10: "While the Study 2a long-term (T5) follow-up participants were composed of a subset of the full sample, they were demographically similar to the main sample (see Supplementary Materials) and there was no a priori reason to expect that these individuals would show a disproportionate advantage for sequence over featural memory." Did this subset differ from the original groups in memory performance at any of the intervals? This might suggest they are a biased sample. It does appear visually that featural memory in the Study 2a Sleep group increases across the 12 hours (Fig 3), which it does not appear to do as strongly in the main group (Fig 2), but it isn't clear if this is significant.

In response to this comment, we ran a new analysis assessing the effect of Attrition (i.e., those who did T5 versus those who did not) to assess effects of bias on performance. There was neither a significant main effect of Attrition, nor was there any significant interaction involving this effect, further supporting the contention that the participants who returned for T5 were representative of the original Study 2a sample. This analysis is now reported on p. 3 of the Supplementary Information.

> ***Study 2a participants who completed the T5 assessment were representative of the full Study 2a sample***

The ancillary T5 assessment to assess remote memory for the Baycrest tour was completed by 49 of the original 77 Study 2a participants (26 Sleep and 23 Wake). We compared those who completed this assessment to the original Study 2a sample on demographics and memory performance. Those who completed the T5 assessment did not statistically differ in age ($t(65) = 0.41$, $p = .685$) or years of education ($t(61) = 0.30$, $p = .768$) from the 28 Study 2a participants who did not participate in this long-term follow-up test. Considering memory performance, we compared participants who completed T5 to those who did not (i.e. Attrition Group) in an ANOVA that included effects of Group (sleep vs. wake), Time (T1-T4), and Memory Type (featural vs. sequence). There was no significant effect of Attrition Group ($\chi^2(1) = 0.03$, $p = .862$), nor were there any significant interactions between Attrition Group and the other factors ($\chi^2(1)$s > 0.72, p's $> .093$). Finally, the critical Time (T1-T4) X Retrieval Type (featural vs. sequence) interaction observed in Studies 1 and 2a held for the Sleep group who completed the T5 test, even though both and sequence and featural memory improved from T1 to T2 (sequence: $z = 3.03$, $p = .002$, $d = 0.986$; featural: $z = 2.08$, $p = .038$, $d = 0.545$). In sum, participants who completed the T5 test were representative of the full sample in terms of demographics, memory performance, and the Sleep group's relative preservation of sequence memory over time.

As noted in our response to Reviewer 1 (point 6), the analysis at T5 was focused on the remote assessments and not T1-T2. Accordingly, the critical Time (T1-T4) X Retrieval Type interaction observed in Sample 1 and Sample 2a was also evident in this subsample (Sleep Group only), as reported in the results (p. 10).

As reported for the Study 2a sleep group, a sleep-related advantage for sequence over featural memory performance from T1 to T4 was evident among those who completed T5 (see Supplementary Materials). This mnemonic enhancement conferred by one night of sleep (versus wake) held at 15 months post-encoding (Group × Time [T1, T5] interaction; $\chi^2(1) = 5.57$, $p = .018$. Although the three-way Group × Time × Retrieval Type interaction was not significant ($\chi^2(1) = 0.67$, $p = .411$), hypothesis-driven planned comparisons confirmed sequence memory performance from T1 to T5 was preserved by sleep relative to wake ($z = -2.24$, $p= .025$, $d = -0.638$). There were no group differences in T1-T5 change in featural memory performance ($z = -1.09$, $p = .274$, $d = 0.432$). While the Study 2a long-term (T5) follow-up participants were composed of a subset of the full sample, they were indistinguishable from the main sample based on either demographics or memory performance (see Supplementary Materials) and there was no a priori reason to expect that these individuals would show a disproportionate advantage for sequence over featural memory.

While it is the case that overnight change in featural memory significantly increased in this subsample (now reported in the Supplementary Information, p. 3; see above), the fact that

this effect crossed the significance threshold in one sample and not in another should not be interpreted as a true difference across samples; this effect was formally tested in the Attrition Group analysis, and it was not significant.

3. **P. 18: In the T5 test, did the sequence and/or feature questions have the EXACT same question (comparing the same elements) or were they mixed and recombined to ask about different parts of the tour (using the same parts from before)? That is, was it the case that, e.g., a sculpture called "Sails" was compared in time to the shuffleboard game, but that at the T5 test, the sculpture called "Sails" was compared to the sculpture called "One Nine North" (a different part of the tour), or was the EXACT same comparison made (e.g., "Sails" vs. shuffleboard again)? If the exact, there could be testing effects (e.g., Rowland, 2014) that should be mentioned as a caveat here, but if not, that is less likely.**

As Test 5 was not part of the original study design, we re-used items from T1-T4 to create a new test. The items were repeated verbatim; none were mixed or recombined, as now clarified in the manuscript on p. 19:

> *As this was not part of the original test design, 100 items (50 featural and 50 sequence) were selected from the original four test forms in equal proportion for the T5 test. These items were used in their exact same format as in previous tests.*

Any effect owing to the reuse of items would have been minimized due to the amount of time (i.e., more than one year) that passed between T4 and T5. This conjecture is supported by the fact that the forgetting curves (Figure 3) followed the expected trajectory considering the data observed for T1-T4. Most importantly, the critical conclusions from the T1-T5 analysis concerned a comparison of sleep vs. wake groups, who were exposed to the same test items. Therefore, any testing effect – if present – was controlled.

4. **P. 21: "This procedure involved building a time window of 4 seconds around the negative peaks of the detected slow waves. The spindles were identified as coupled slow wave-spindle (SW-SP) complexes if the spindle peak latency fell within a 4-second time window of the slow wave peak detected in the corresponding slow wave channel. Alternatively, spindles were marked as uncoupled if the spindle peak latency occurred outside the 4-second time window." In my opinion, a 4-second window is too wide, as most slow waves in SWS (and outside of deep anesthesia) occur at a higher frequency that would be more amenable to a shorter window. Even a 3-second window (~ a 0.66 Hz slow wave cutoff) would capture some spindles that start in the next slow wave. The distribution even appears in their Figure S6 – almost all coupled spindles are within the first +/- 1 second. The danger here is that one could be frequently**

capturing spindles that occur after the slow wave, which may or may not meet the definition of a slow wave. Personally, I think the better choice would be to use a method where they find the time of troughs between verified slow oscillations and the timing of the spindle start and only consider it to be coupled if it occurs between these troughs. This avoids the issue of needing to place any frequency cutoff to begin with, and it also avoids the issue of spindles occurring within the next slow wave. I see that they use an established method (from the Baena Perez studies, who is a co-author), so I am curious if there is anything from those studies that would allay this concern. Please either explain why this concern is misplaced or consider changing the method for coupling detection.

We share the concern regarding the fixed window length detection method. However it is a well-established method initially developed and used by Mölle et al. (Mölle et al., 2002, 2011), and has been replicated in many subsequent studies (e.g., Baena et al., 2020; Clemens et al., 2011; Denis et al., 2021; Staresina et al., 2015). Figure S6B shows that this approach yields very tight average coupling between spindles and SW. Figure S6A shows the expected distribution of spindles coupled predominantly in the upstate – as is consistent with the literature.

Further studies using the same approach have reduced this window to 1 second to improve precision (Denis et al., 2021). However, a 1-second window would be overly restrictive given our defined frequency range of slow waves (0.5-4.0 Hz), where the period of a slow wave at the lower end (0.5 Hz) is 2 seconds. A smaller window would exclude certain coupled events. Accordingly, we employed a 2-second window around the slow wave to capture spindles that might occur during the entirety of the slow wave cycle that is physiologically possible. Furthermore, our detection method uses the spindle peak latency, ensuring that even in the most extreme cases, at least half of the spindle event is contained within the slow oscillation window.

Nonetheless, for the initial analyses of this manuscript, we had in fact applied a very similar (if not identical) refined approach. This new approach is currently in development by our group, and we are in the process of validating the method, which is similar to that suggested by the reviewer. We have not been able to use this approach in our previous studies (Baena et al., 2022, 2023) due to the shorter duration of sleep recordings in those studies. This new approach considers the individual duration of the SOs' negative and positive peaks, with the spindle peak latency contained within the duration of each detected SO, making it more precise.

We decided to use the fixed window approach for its use of well-established methods, and developing a novel method for this particular study (already involving various novel methodologies beyond sleep physiology) would be premature. However, in response to this comment, we report the results using this novel methodology in the supplementary materials on p. 16, referred to on p. 12 and 23 of the main manuscript.

Main manuscript, p. 12

> *An alternative method for identifying spindle–slow wave coupling, flexibly identifying spindles occurring during the upstate vs. downstate of individual slow waves, replicated this finding, and highlight the particular importance of coupling in the slow wave upstate (see Supplementary Results) .*

Main manuscript, p. 22

> *Using the slow wave negative peak latencies from F3 and F4 and the spindle peak latencies from C3 and C4, we performed coupling detection procedures using the approach originally developed and validated by Mölle and colleagues (20, 66, 93–96) employing EEGlab-compatible software (92) written for MATLAB R2019b (The MathWorks Inc.). To ensure that spindles would only be coupled to slow waves detected on the same hemisphere, electrode pairings F3-C3 and F4-C4 were made. This procedure involved building a time window of 4 seconds (i.e., ±2 seconds) around the negative peaks of the detected slow waves. The spindles were identified as coupled slow wave-spindle (SW-SP) complexes if the spindle peak latency fell within the identified 4-second time window of the slow wave peak detected in the corresponding slow wave channel (Supplementary Figure 6A). Alternatively, spindles were marked as uncoupled if the spindle peak latency occurred outside the 4-second time window. Results of an alternative method that flexibly identifies spindles occurring during the peak (upstate) vs. trough (downstate) of each individual SO event did not affect interpretation (see Supplementary Results).*

Supplementary Information, p. 16:

> **Study 2B: Spindle-SO coupling results using a refined method targeting SO peak (upstate) vs. trough (downstate)**
>
> *To augment our spindle-SO coupling results reported in the main manuscript, we developed a refined method to detect spindles that are specifically coupled to either the upstate or downstate of the slow wave. Instead of relying on a fixed time window of ±2 seconds around the slow wave's negative peak, we created individualized time windows for each positive half-wave (upstate) and negative half-wave (downstate)—based on their specific durations. This approach ensures that spindles occurring within these windows are tightly coupled to the corresponding part of the slow wave. As with the fixed window length approach, spindle peak latency served as the reference point. In rare cases where the spindle peak latency occurred during the transition from the down to upstate, the event was considered coupled to both segments. On average, 12 spindle events (SD = 7) were classified as coupled to both parts of the slow wave. The same electrode pairings, F3-C3 and F4-C4, were used for this refined approach. This method offers an advantage over the fixed ±2 seconds time window approach by either reclassifying events located at the tails of the slow*

wave ±2 seconds time window (indicated by the grey bars in Figure S6A) into more appropriate half-wave segments or removing them altogether, thus enhancing detection accuracy.

As reported in the main manuscript, coupled, but not uncoupled spindles were related to overnight change in overall memory performance. Considering the SO positive half-waves, spindles coupled to slow waves' upstate—but not uncoupled spindles—were associated with overall memory enhancement (interactions with Time: F(1, 138) = 8.14, p = .005, partial eta squared = 0.06; F(1, 138) = 0.49, p = .484, eta squared ~ 0.00, for coupled and uncoupled spindles respectively. Similarly, spindles coupled to slow waves' downstate—but not uncoupled spindles—were associated with overall memory enhancement (interactions with Time: F(1, 138) = 4.07, p = .046, partial eta squared = 0.03; F(1, 138) = 0.41, p = .521, eta squared ~ 0.00, for coupled and uncoupled spindles respectively). When spindles coupled to the upstate and downstate were considered in relation to overnight change in memory performance in a single model, only the effect for coupled spindles in the upstate was significant (F(1, 138) = 5.04, p = .026, partial eta squared = 0.04, vs F(1, 138) = 1.04, p = .311, partial eta squared ~ 0.00 for coupled spindles in the downstate), suggesting coupled spindles in the upstate uniquely relate to overnight change in memory performance.

Moderate issues:

Fig 2 caption and in manuscript: The authors report what qualitatively looks to be a negatively accelerating forgetting function, but I think they could actually fit it. They mention a linear decline against log spacing, so they could test the fit for y = log(x) +b, where y is memory and x is retention interval, and report the R-squared value. I imagine this will break for the sequential memory case, but I believe that is fine, as it is clear sequential memory differs from the typical forgetting curve, and they aren't making an argument that sequential memory fits this pattern anyway.

We do indeed conclude that featural memory forgetting (roughly) fits the classic, Ebbinghausian, negatively accelerating forgetting curve, whereas sequence memory forgetting does not. As noted by the reviewer, the suggested model fit is approximately captured in the approximate log-spacing of the intervals visualized and analyzed as a linear variable (e.g., Figure 2).

We originally opted not to formally fit a forgetting function given that the precise mathematical form of forgetting—including for naturalistic events and timescales—has been much debated and discussed for decades (e.g., logarithmic vs. power vs. exponential power functions, etc; e.g., Rubin & Wenzel, 1996; Wixted, 2004). In response this suggestion, we converted the Time variable back to its 'raw' expression in hours (1, 24, 168, 672), and tested the fit for y=log(x)+b (for featural memory only, Study 1), which produced an R-squared value of .197 (and for sequence memory: < .001, indicating a bad fit as expected). This supports the reviewer's prediction. While the suggestion is warranted, we feel that the

additional analysis reporting does not add much value over and above the existing analysis, given that our study was not designed to contribute to this debate (beyond the observation that the roughly Ebbinghausian function may apply to featural but not sequence information from the same experience).

P. 14: "Specifically, spindles coupled to slow waves uniquely predicted overnight memory enhancement, aligning with contemporary models of hippocampal-neocortical communication in sleep-related mnemonic processing (17, 24) (for review, see (19)). Contrary to our hypothesis and behavioural evidence for sleep's selective effect on sequence memory, SWS duration and neurophysiology predicted both sequence and featural memory change – although there was evidence for greater specificity to spindle-slow oscillation coupling for sequence relative to featural memory." Despite my largest comment on needing to weaken the physiology interpretations, the authors' own work may be useful in speculating about something else here. SO-spindle coupling changes in aging (Helfrich et al., 2018, Neuron), and given their prior work on aging in the current paradigm (Diamond & Levine, 2020, Psych Sci), there could be room for meaningful speculation about performance on this task in aging and how it relates to changes in sleep physiology.

We thank the reviewer for this insight and have now incorporated this idea into the discussion (p. 16)

> *Given prior observations of altered spindle-SO coupling during sleep in older adults (76–78) – and disorganized neural replay in aged rodents (79) – dysfunctional sleep oscillatory coordination could be a candidate mechanism for the age-related loss of temporal organization in real-world memory retrieval, as previously demonstrated with the Baycrest Tour (46). Future investigations could inform compensatory techniques or therapeutic interventions such as oscillatory entrainment (80) and/or targeted memory reactivation during sleep (81, 82) or wakefulness (e.g., 83).*

P.21: "Consideration of fast and slow spindles separately did not add new information to hypothesis-driven analyses." Is this a statistical model comparison or something else? Please clarify.

Our exploratory analyses of fast vs. slow spindles is now described on p. 22 of the main manuscript:

> *Exploratory analyses of slow (11–13.5 Hz), fast (13.5–16 Hz) spindle counts in relation to memory performance indicated that slow spindles interacted with Time ($F(1, 141) = 10.26$, $p = .002$) in a similar manner to total spindle counts; there were no significant findings involving fast spindles.*

Small issues:

P. 4: "Sleep may in some cases improve non-episodic memory tasks like motor sequence learning or statistical learning (28, 29) (but see (30))…" I recognize the authors may be trying to avoid too strong a claim (in attempts to avoid offending anyone), but I would state the lack of raw procedural learning improvements across sleep more strongly. The only studies to find improvements (e.g., Walker et al., 2002) have clear fatigue confounds (from the context of today, after the methods measuring fatigue effects have become clearer). Not that they would need this to be the case, but I think this ultimately makes their findings of raw improvement more novel.

We appreciate that there have been contentious claims in this literature as evidence has accumulated. We attempted to succinctly characterize the literature on above-baseline sleep-related enhancement in learning and memory—with appropriate caveats so as not to overstate any claims. We made some changes to the citations to better characterize the consensus on this topic.

> *To our knowledge, however, there is little or no evidence for the predicted sleep-induced above-baseline enhancement of episodic memory, owing perhaps to insensitive behavioural measurement. Sleep may in some cases improve non-episodic memory tasks like motor sequence learning or statistical learning (30, 31) (but see (32)), but in studies of episodic memory, sleep at best reduces the amount of forgetting relative to wakefulness (24) (but see (33)).*

P. 4: "One idea sleep serves as "memory triage" system, tagging only the most relevant pieces of information to persist over time" Please fix this sentence.

This sentence is fixed:

> *These discrepancies may be attributable in part to the differential forgetting of heterogeneous episodic representations over time (33–36), and perhaps sleep. One idea is that sleep serves as "memory triage" system, tagging only the most relevant pieces of information to persist over time (24, 37), such as information that is emotional (13, 38–40), congruent with prior knowledge (41), or personally relevant for future utilization (e.g., (42, 43)).*

Paragraph spanning p. 4-5: It isn't 'once-presented', but the sequential memory paradigm in Petzka et al. (2022, Current Biology) seems relevant here.

We have now added this citation specifically in our introduction on pg 3:

> *Evidence in humans is consistent with this proposed mechanism (e.g. (20–22)...*

P. 5: The authors may find some of the following studies helpful from the targeted memory reactivation literature, which examine memory over longer periods of time: Knötzele et al. (2023, Sci Rep); Neumann et al. (2020, Sci Rep).

We have now added these citations specifically in our introduction on pg 5:

> *Furthermore, little is known about sleep's influence on episodic memory for real-world events experienced outside of the laboratory – which is more durable than memory for laboratory stimuli (53) – beyond delays of a few hours or a day (54, 55), despite evidence that memory consolidation or transformation unfolds over weeks to months or longer (5, 54, 55).*

P. 8: "The preferential effect of sleep on sequence versus featural memory could not be…" Up to this point, they've only shown a 24-hr increase, which may or may not be attributable to sleep. Please fix to not (yet) focus on sleep.

We have altered the phrasing here as follows:

> *The improvement of sequence versus featural memory over the 24-hr interval from T1 to T2 could not be explained by item characteristics such as difficulty or true/false phrasing (see Supplementary Figure S1).*

P. 9: "Here, we first replicated the findings observed in Study 1 in the Sleep condition, where..." -> "whereas"

We clarified this statement as follows (removed "where"):

> *Here, we first replicated the findings observed in Study 1 in the Sleep condition: a main effect of Time (T1 – T4) (□(1) = 60.40, p < .001) was once again qualified by a Time X Retrieval Type crossover interaction (□^{2}(1) = 13.85, p < 0.001) (Figure 2C).*

P. 11: "(N = 49 after exclusions, see Supplementary Methods" Please add ')'.

Done.

P. 13: "Sleep's influence on the fate of our memories has been studied for at least a century." The authors should cite a study supporting this claim.

We now cite Jenkins and Dallenbach (1924) in support of this claim.

> *Sleep's influence on the fate of our memories has been studied for at least a century (e.g., 7).*

From the Supplement:
Table S3 – Were the stage correlations corrected for multiple comparisons, or was there an a priori hypothesis predicting SWS duration would lead to behavioral improvements? These are not the featured sleep physiology analyses, so I do not believe these carry that aspect of the paper, but the authors should state either way.

Tables S3-S5 in the Supplementary Information lists the terms from linear mixed effects models assessing the relationship of sleep macrostructure (Table S3) and microstructure (Tables S4-S5) variables to memory performance in T1-T2. For macrostructure, all sleep stages competed for variance in a single model, they were not separate correlations. As noted in the methods, significant model predictors were followed up with bivariate Pearson correlations (Figure S4). As pointed out by the reviewer, these are not the featured sleep physiology measures; they are presented for descriptive purposes and completeness. That said, there is sufficient literature to justify the focus on slow wave sleep and spindles. We now clarify that these correlation coefficients are reported without correction.

> ***Fig. S4.*** *Correlation coefficients (uncorrected) following up on significant effects from models testing the relationship of (T1 to T2) change in memory for sequence (top panel) and features (bottom panel) to macrostructural and microstructural sleep measures (see Tables S3-S5).*

Reviewer #3:
Remarks to the Author:
In this study Diamond and colleagues investigate to what extent sleep transforms multidimensional real-world experiences (i.e., art tour related memories). They show that sleep selectively enhances sequential aspects of memories, as compared to perceptual details. Importantly, this effect grew stronger over up to a year post-encoding and was not observed after wakefulness. In addition, memory enhancement was associated with slow-wave-sleep duration and SO-spindle coupling. These findings are highly intriguing and shed light on several important features of memory consolidation (which memory aspects preferentially benefit from sleep? How durable are those effects?). While I am generally very impressed with the high quality of the work, I do have some questions ± the EEG analysis.

We appreciate this assessment.

Comment 1: Why is there a significant discrepancy in the number of detected events? The number of SOs detected appears excessively high. As shown in Fig. S4, the mean of detected SOs is approximately 7000. Assuming participants spent around 5 hours in NREM sleep, this would imply that SOs occurred roughly every 2.5 seconds, which seems improbable, especially during sleep stage N2. Conversely, the number of spindles detected seems quite low (mean around 450) compared to what has been reported in the literature. Could you please provide more details on the spindle detection methodology?

The mean number of detected half-SOs (10,634) and spindle events (1505) are reported in Table S2. The SO detection refers to half-waves, which explains the high number reported (double what would be expected for full waves).

The spindle detection methodology follows a well-established and validated (against both multiple expert scorers and crowd-sourced from a large sample of non-experts) approach from Fogel's lab. Detailed descriptions are provided in the Supplementary Information (pp. 5-6). Briefly, the detection method uses a complex demodulation transformation of the EEG signal with a bandwidth of 5 Hz centered around a carrier frequency of 13.5 Hz (i.e., 11–16 Hz). This method allows us to extract the number, duration, amplitude, and frequency of spindles at each central derivation (C3 and C4).

Comment 2: Related to the above, Fig. 4 suggests that approximately half of the spindles were coupled with SOs. This ratio appears quite high compared to previous findings (e.g., Mühlroth et al., 2019; Schreiner et al., 2021). This may be due to the authors' use of a very liberal criterion for identifying coupled events, where spindle peaks within a 4-second window around an SO downstate are considered coupled. I find it unlikely that a spindle occurring 3 seconds after an SO could be considered coupled. I recommend applying a more conservative criterion, such as ±1.5 seconds, to identify coupled events more accurately.

We now clarify on p. 23 that our 4-second window refers to ±2 seconds around the slow wave's negative peak, making it comparable to the ±1.5 seconds suggested by the reviewer and ruling out the scenario of a spindle occurring 3 seconds after an SO.

On average, 37% of the detected spindles were coupled with SOs. While this might be due in part to the detection method, the fixed time window method is well established and yielded tight average coupling between spindles and SW (Fig S6b) with the expected distribution of spindles coupled predominantly in the upstate (Fig S6a), suggesting this is a valid approach. The ±2 second window optimizes the sensitivity of spindle detection given our defined frequency range of slow waves (0.5-4.0 Hz). That said, we also now report the results from a new and more precise approach that considers the individual duration of the SO in the Supplementary Information. (Please see our response to Reviewer 2 (point 4) for more elaboration on these points.)

Main manuscript, p. 22-23:

> *Using the slow wave negative peak latencies from F3 and F4 and the spindle peak latencies from C3 and C4, we performed coupling detection procedures using the approach originally developed and validated by Mölle and colleagues (20, 66, 93–96) employing EEGlab-compatible software (92) written for MATLAB R2019b (The MathWorks Inc.). To ensure that spindles would only be coupled to slow waves detected on the same hemisphere, electrode pairings F3-C3 and F4-C4 were made.*

This procedure involved building a time window of 4 seconds (i.e., ±2 seconds) around the negative peaks of the detected slow waves. The spindles were identified as coupled slow wave-spindle (SW-SP) complexes if the spindle peak latency fell within the identified 4-second time window of the slow wave peak detected in the corresponding slow wave channel (Supplementary Figure 6A). Alternatively, spindles were marked as uncoupled if the spindle peak latency occurred outside the 4-second time window. Results of an alternative method that flexibly identifies spindles occurring during the peak (upstate) vs. trough (downstate) of each individual SO event did not affect interpretation (see Supplementary Results).

Ultimately, we decided to use fixed window approach in our main analyses given that it is well-established. We are currently working on a methods paper comparing this new approach with the fixed method, as well as with commonly used Phase Amplitude Coupling (PAC) approaches in order to validate this novel approach and provide rationale for future studies.

Comment 3: The authors used different electrodes to detect SOs (frontal) and spindles (central), which makes sense when analyzing these oscillations independently. However, for coupled events, I recommend that the authors focus on a single site (see, e.g., Helfrich et al., 2018).

We tailored the electrode location to each type of event detected for optimal accuracy per event type, at the site where they predominate. Slow oscillations (SOs) are best detected using frontal derivations, as they tend to have higher amplitudes in the frontal regions of the brain. For spindles, to encompass both fast and slow variants, we used central derivations (C3, C4). This approach ensures accurate detection for each oscillation type and is based on well-established methodologies. The two scalp locations are in close proximity to one another (about 3-5 cm), and would therefore not contain discrete information (i.e., the information at both sites are highly correlated). Any temporal lag between these two sites would be negligible and have no appreciable impact on the outcomes given the analysis approach.

Comment 4: In the methods section the authors state that the lag variance between SOs and spindles was calculated for each individual as a measure of coupling strength. Could the authors report the results of this analysis? Same goes for uniformity tests (Rayleigh and V-test).

The requested results are now reported in the main manuscript and the Figure S6 caption.

Main manuscript, p. 23

Additionally, the phase of the bandpass-filtered slow-wave signal in radians at the spindle peak latency was computed (see Supplementary Figure S6). The mean direction of the phase angles for all coupled spindle events was determined using the CircStat toolbox (97). Hilbert transform was applied to extract the preferred phase of

SW-SP coupling for each participant averaging all individual events' preferred phases. Then, we performed uniformity tests (Rayleigh test) using positive slow wave peaks as the predefined mean direction (V-test). Individual-level analyses revealed a non-uniform distribution ($p < 0.001$, Rayleigh test) of the preferred phases of SW-SP in all 49 participants. The coupling of spindle events within the slow wave cycle was maximal shortly before the upstate peak in all 49 participants (0° (slow wave positive peak); $p < 0.001$, V-test), suggesting that spindles were coupled to slow waves preferentially adjacent to or prior to the positive slow wave peak (Supplementary Figure 6B). For the purposes of this paper, we analyzed the count of coupled and uncoupled spindles in SWS averaged across channels.

Supplementary Information, p. 6 (Figure S6 caption)

Fig. S6. A) *Coupled slow wave-spindle histogram for all participants. Each bar represents the number of coupled spindles detected in an interval of 125 milliseconds divided by the total number of spindles. Average slow wave oscillation for all participants is superimposed in black.* ***B)*** *Circular plot of preferred phase for each individual (slow wave phase at spindle amplitude maximum). Red dots denote an individual preferred phase (0^o slow wave down-to-up state, $\pm 180^o$ slow wave up-to-down state). All participants showed a non-uniform distribution (Rayleigh test $p < 0.001$ for all 49 participants). The direction of the line indicates the preferred direction of the grand average. All individuals exhibit spindles adjacent to, or immediately preceding, the positive slow wave peak at 0^o (V-test; $p < 0.001$ for all 49 participants).*

Comment 5: was the correlation between SWS duration and memory change corrected for multiple comparisons?

As noted above (last comment, Reviewer 2) these correlations were run as follow up analyses based on significant model predictors. As now noted in the figure caption, they were uncorrected for multiple comparisons.

Fig. S4. *Correlation coefficients (uncorrected) following up on significant effects from models testing the relationship of (T1 to T2) change in memory for sequence (top panel) and features (bottom panel) to macrostructural and microstructural sleep measures (see Tables S3-S5).*

Point-by-point response to the reviewers' comments

Reviewer #2 (Remarks to the Author):

First, I want to commend the authors for a very nice revision. I appreciate their thoughtful comments regarding many points, including the new physiology analyses. I think the paper should be published, but I wanted to follow up on a few points from their response letter.

We appreciate these comments.

#1
"Any effect owing to the reuse of items would have been minimized due to the amount of time (i.e., more than one year) that passed between T4 and T5. This conjecture is supported by the fact that the forgetting curves (Figure 3) followed the expected trajectory considering the data observed for T1-T4. Most importantly, the critical conclusions from the T1-T5 analysis concerned a comparison of sleep vs. wake groups, who were exposed to the same test items. Therefore, any testing effect – if present – was controlled."

Here I disagree. It is not at all clear to me that the effect of reuse would be so minimal as to not be a potential factor at T5.

In one respect, it seems the concern is being minimized due to the time scale. Such timescales resemble those in some large spacing effect studies (that involve retrieval tests) such as Cepeda et al. (2009) (Exp 2). That study is not particularly similar to this one in many ways, and Cepeda and colleagues provide feedback at the Session 2 test, but I mention it to show the profound effect testing on something after a delay of days to a few months can have on later memory of similar timescales (~ a year).

The authors presumably did not include feedback for their T1-4 tests (because there would have been no reason for them to do so -- but please state whether this is incorrect). Given that successful retrieval produces strengthening (e.g., Rowland, 2014), this means that any differences in performance due to sleep versus wake -- which indeed they observe and is a highlight of this paper -- would lead to a different number of items getting strengthened. For instance, about 85.5% of sequence items were remembered in the 12-hr sleep condition (T2), whereas about 78% were remembered in the 12-hr wake condition. This means that, of the items that were on that T2 test that were included in the later T5 test, an average of 7% MORE items would be strengthened in the sleep condition the wake condition, which could have an impact on a later test. The same principle applies for any tests with differential performance (e.g., T3 & T4) (and would apply in the other direction if it were better after wake).

Also, although I appreciate their logic, I disagree that simply following the expected downward trajectory following from T1-4 means there's no cause for concern -- there could easily be a confounding effect of taking a prior test alongside forgetting (over the course of another year) that would lead to the observed results.

Therefore, it still seems to me that – because successful retrieval can have a profound influence on memory, whereas unsuccessful retrieval without feedback seems to NOT have this effect (e.g., Kornell et al., 2011) – the T5 results could be influenced by differential performance across sleep versus wake conditions on T1-4. If the authors still disagree or if I have misunderstood something, of course they should state this in the response. Otherwise, it seems worth highlighting as a limitation for this very long interval data point.

First, the reviewer is correct in assuming we did not provide feedback on any of the tests – we have now clarified this on page 15 of the manuscript ("Participants were not given corrective feedback.")

We now agree that the potential influence of testing effects on T5 performance should be mentioned as a limitation. We included this on page 16:

"Although these items were repeated from earlier tests, the effects of prior testing were expected to be minimal given the passage of time (over one year from the prior test) and the absence of corrective feedback. That said, potential testing effects should be considered in interpreting the T5 data (Rowland, 2014)."

We find the reviewer's suggestion that the difference in Sleep vs. Wake group T5 performance could be shaped by differential retrieval success at T2-T4 plausible, but also compatible (i.e. not mutually exclusive) with sleep-related theories that are the focus of this study, which is not well-suited to adjudicate between different theories of the testing effect's influence on long-term retention. In general, we agree it is important to consider – for the T5 results – the interaction of the sleep manipulation and the influence of repeated testing. To this end, we added the following to the discussion (p. 13):

"…further research is needed to assess the effects of arousal or subjective engagement with the memoranda, the timing of sleep relative to encoding (discussed above), the interaction of sleep and testing effects, the temporal evolution of memory transformation over multiple sleep-wake cycles…"

#2
Regarding the Ebbinghaus point, I first wanted to apologize that I misstated the function. I believe the log fit function would be y=a-b*log(x). When I estimate their memory results from the figure and plug them into code to perform curve fitting, I see an R-squared of 0.91 (or adjusted R-squared of 0.86, if the authors prefer that value). Here is the sample MATLAB code I used (requires the Curve Fitting Toolbox):

```
diam_ri=[1;24;168;672]; %define RI and mem values
diam_s1=[83;80.5;74.5;70.5];
ft=fittype('a-b*log(x)'); %define function fit -- change for different functions
options=fitoptions(ft);options.StartPoint=[1 1];options.Lower=[0 0]; % initializations
[fit_mod,gof]=fit(diam_ri,diam_s1,ft,options); %do fit
R_sq=gof.rsquare %print rsquare to terminal
Adj_R_sq=gof.adjrsquare %print adj rsquare to terminal
```

Regardless of the exact memory values (clearly my estimations were not exact), I think this fit should be much stronger than 0.197 – even a linear fit run by substituting the following line above …

ft = fittype('a-b*x');

…produced an R-squared of 0.81.

Note that if they prefer other negatively accelerating functions, I believe the exponential is:

ft = fittype('a*exp(-b*x)');

and the power law is:

ft = fittype('a*x^-b');

Whether or not the authors decide to report any of these results is of course up to them, but I wanted to add this correction.

We thank the reviewer for correcting their earlier suggestion, and for conducting these analyses. We still feel that these curve fitting results are interesting, but that the additional analysis reporting does not add much value over and above the existing analysis, given that our study was not designed to contribute to the debate between the many possible negatively accelerated forgetting functions in long-term naturalistic memory (beyond the observation that the roughly Ebbinghausian function may apply to featural but not sequence information from the same experience).

#3. To be clear, were the exploratory slow and fast spindle results significantly different from each other? I also don't know if there's enough information from the new sentence ('interacted with Time') to glean how slow spindles predicted memory. Please make these effects a bit clearer.

We elaborated on this point to clarify that (A) the exploratory analyses of slow and fast spindles were significantly different from each other, and (B) what is meant by 'interacted with time' (as with all other related analysis in the paper, we tested whether slow/fast spindles significantly interacted with Time [T1 vs. T2] in mixed effects models, followed up with bivariate corelations). We moved this analysis description from the main text Methods section (where it was previously) to the Supplementary Information (final section):

Exploratory analysis of slow vs. fast spindles

Exploratory analyses of slow (11–13.5 Hz) versus fast (13.5–16 Hz) spindle counts in relation to memory performance, in separate mixed effects models, indicated that slow spindles were associated with significant overnight improvements in overall memory performance in a similar manner to total spindle counts (interaction with Time: $F(1, 141) = 10.26$, $p = .002$, partial eta

squared = 0.07; no significant three-way interaction with Time and Retrieval Type: $F(1, 141) = 0.11$, $p = .738$, partial eta squared < 0.01). Fast spindles, on the other hand, were not associated with overnight memory change (no significant interaction with Time ($F(1, 141) = 0.07$, $p = .787$, partial eta squared < 0.01) nor Time and Retrieval Type ($F(1, 141) = 0.66$, $p = .416$, partial eta squared < 0.01)). Considering bivariate correlations, overnight memory change was positively correlated with the average number of slow spindles in SWS ($r(47) = 0.47$, $p < .001$, 95% CI [0.223 – 0.667]) but not fast spindles ($r(47) = -0.04$, $p = .777$, 95% CI [-0.319 – 0.242]). These correlations are significantly different from each other: Steiger's $z = 2.38$, $p = .021$).