nature portfolio

Peer Review File

Neural activations and representations during episodic versus semantic memory retrieval

Corresponding Author: Dr Roni Tibon

Version 0:

Decision Letter:

20th July 2020

Dear Dr Tibon,

Thank you for submitting your Registered Report proposal entitled, "Do activations and representations differ for episodic versus semantic memory?", and for your patience while awaiting an initial decision.

As you may know, we decline a substantial proportion of manuscripts without sending them to referees. After careful consideration, I regret that we cannot offer to send this Registered Report proposal out to review at this time.

We agree that your your study addresses an important question, and we appreciate your novel task design, which we believe is likely to yield new insights on this question. However, we are not able to consider this Registered Report further because we believe that the proposed sample size (Bayesian sequential sampling maximum of N=60) is insufficient to support strong interpretation of null results.

If I understand your manuscript correctly, your modelling indicates that if the null hypothesis were true, your study would have a 75.2% chance to yield a BF10 of 1/6 or lower. In our view, this unfortunately is not sufficient statistical power. We believe that a Registered Report needs to be able to yield interpretable results both in support of H1 and H0.

We would be willing to reconsider this decision if you were able to increase the statistical power. Please contact me if you wish to discuss this decision.

I am sorry that we cannot respond more positively on this occasion, and hope that the negative outcome in this instance will not deter you from submitting future work to Nature Human Behaviour.

Sincerely,

[redacted]

[redacted]

Nature Human Behaviour

****

You might wish to consider transfer of your manuscript to Nature Communications - who have recently started to consider Registered Reports. Please use our manuscript transfer portal to initiate the transfer to this journal (or to another journal of your choice in the Nature Research portfolio). If you transfer to Nature-branded journals or to the Communications journals, you will not have to re-supply manuscript metadata and files. This link can only be used once and remains active until used.
All Nature Research journals are editorially independent, and the decision to consider your manuscript will be taken by their own editorial staff. For more information, please see our manuscript transfer FAQ page.

Version 1:

Decision Letter:

Dear Dr Tibon,

Thank you for your appeal asking us to reconsider our decision on your Registered Report "Do activations and representations differ for episodic versus semantic memory?". Please accept my apologies for the delayed response - as I mentioned, we are currently receiving a very high volume of manuscripts, and first time submissions must take a higher priority over appeals on previous decisions.

Now that I have had a chance to discuss the matter carefully with my colleagues, I am sorry to have to tell you that we do not feel able to reverse our original decision.

You have suggested increasing the maximum sample size to N=80. We appreciate this change and recognize that this would increase the chance of obtaining strong evidence; however, for a Registered Report we require 95% statistical power or 95% chance to achieve a BF of 10 or 1/10 ("strong evidence" according to Lee & Wagenmakers 2014 and others).

I am very sorry that our decision must remain negative. Unfortunately, the number of papers submitted to Nature Human Behaviour vastly exceeds the number that we can publish each month, and we are therefore frequently forced to make difficult decisions. Once again, I hope that you rapidly receive a more favourable response elsewhere.

Best regards,

[redacted]

[redacted]
Nature Human Behaviour

Version 2:

Decision Letter:

24th March 2021

Dear Dr Tibon,

Thank you once again for your manuscript, entitled "Do activations and representations differ for episodic versus semantic memory?," and for your patience during the peer review process. Once again, please accept my apologies for the delay in 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 protocol to be of interest, they also raise some important concerns. We are interested in the possibility of proceeding further with your submission 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 in principle acceptance and Stage 2 submission.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study.

In this case, the major point (raised by Reviewer #1 and #2) concerns the design of the experiment and the degree to which the episodic and semantic memory conditions are matched, and whether these two memory processes can be distinguished fully. It will be very important to respond to these concerns in such a way that the referees are satisfied in the next round of review. Referee #2 also queries the potential contribution of this study, arguing that the distinction between semantic and episodic memory systems in the brain is already established.

In sum, we invite you to revise your Stage 1 Registered Report taking into account reviewer and editor comments. Please highlight all changes in the manuscript text file. Please do not hesitate to get in touch if you would like to discuss these issues further.

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.

When revising your manuscript:

* Include a “Response to reviewers” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the reviewers along with the revised manuscript.

* Ensure that you use our template for Stage 1 Registered Reports to prepare your revised manuscript: https://www.nature.com/documents/NHB_Template_RR_Stage1.docx. Failure to ensure that your revised Stage 1 submission meets our requirements as specified in the template will result in your submission being returned to you, which will delay its consideration.

* In your cover letter, please include the following information:
--An anticipated timeline for completing the study if your Stage 1 submission is accepted in principle.
--A statement confirming that you agree to share your raw data, any digital study materials, computer code (if relevant), and

laboratory log for all eventually published results.
--A statement confirming that, following Stage 1 in principle acceptance, you agree to register your approved protocol on the Open Science Framework (https://osf.io/) or other recognised repository, either publicly or under private embargo, until submission of the Stage 2 manuscript.
--A statement confirming that if you later withdraw your paper, you agree to the Journal publishing a short summary of the pre-registered study under a section Withdrawn Registrations.

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 hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Human Behaviour is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
[redacted]

[redacted]
Nature Human Behaviour

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
This is a timely and interesting project on the overlap in the neural correlates of semantic and episodic memory. The design is definitely interesting but might need to be tweaked or clarified.

Some brand logo pairs will be familiar while others will not. It is mentioned that it is expected that only half of the logo-word pairing will be known to participants. Is this based on the pilot data already collected?

In the episodic task, it is indicated that pairing will be established in the study phase. Will all logo-word pairings be unfamiliar/new to participants? This would create a mismatch with the semantic task (where half of the logo-word are assumed to be familiar). Or, if already known, will some of the pairing conflict with prior semantic knowledge (e.g., Nasa matched with the word Google, as cited page 9)? If the later, this would also be an issue (to measure episodic recollection effects for mismatches/violations of semantic knowledge) and is likely to produce some distinct activations for the semantic and the episodic conditions. I understand that participants' instructions would be to make a narrative linking each logo with a different brand name, but this may still be hard to compare with the semantic task including half congruent semantic knowledge presumably acquired over a long time and new unfamiliar associations. One possibility might be to focus on pairs unknown to the participant, and to compare the initially unknown pairs of the semantic task with unknown pairs used in the episodic task (i.e., logo and brands unknown to participants), but I am not sure that the encoding phase in the semantic task allows to obtain this information and then the tasks would become presumably identical or very similar.

How will task difficulty be matched? Some brain regions that have been proposed to be involved in both semantic and episodic processing like the angular gyrus also appear sensitive to task difficulty.

Why do testing on 2 different days? Is it not likely to maximise differences in behaviour and in brain activity?

Reviewer #2:
Remarks to the Author:
In this project proposal, Tibon and colleagues are seeking to investigate whether dissociable brain representations exist in the successful retrieval of semantic and episodic memory. For that purpose, the authors propose a within-subject fMRI study in which they aim to employ two separate tasks that are designed to tap into the two memory systems in a relatively large group of subjects.

The data will then be analyzed employing both conventional task-based activation approaches as well as multivariate representational similarity analysis.

This is a well-written project proposal and a highly interesting topic that will undoubtedly be of interest to the scientific community at large. However, my initial enthusiasm is somewhat dampened by the research question and the hypothesis put forward as well as the experimental design that is aimed to test this hypothesis. I provide detailed points below in the hopes that they further improve the study design.

To begin with, I have three major concerns that I would like to emphasize: First, the authors state their research question as “Can we distinguish successful episodic retrieval from semantic retrieval?” As the authors highlighted themselves, there is ample amount of evidence in the literature to suggest dissociable brain-based representations of these two memory types, generally centered on the ATL and MTL for semantic and episodic memory, respectively. As such, if this is the authors’ aim, then the results of this study will only confirm what is already known for the past two decades of neuroimaging-based research. Hence, the only novelty lies in the multi-session aspect of this experiment in which the two memory systems are probed within the same individuals (traditionally they are studied in isolation).

Moreover, the authors state their hypothesis as “The brain regions supporting successful episodic retrieval are distinct from those supporting successful semantic retrieval.” Testing this hypothesis requires complete behavioral isolation of the two functionally (and mostly experientially) distinct memory systems. Unfortunately, I do not believe that the proposed project achieves this level of separation. Furthermore, as per the authors predictions, finding that “greater recall success effect in the episodic task in episodic ROIs, but in the semantic task in semantic ROIs” would not allow the authors to refute the stated null hypothesis that the brain regions supporting successful episodic and semantic retrieval are shared, but rather show that there are both common and distinct neural features, which is exactly what is already suggested in the existing literature. The same issue goes for the representational similarity analysis, but here the authors should also consider what they are defining as “representation” versus “retrieval” (I believe the question is framed around successful retrieval and not representation).

In addition, the authors aim to focus their analysis on successful versus unsuccessful retrieval, which I believe will introduce a confound to the same hypothesis that the authors are trying to test, especially if levels of successful retrieval are different across the two memory systems.

Taken together, although I wholeheartedly appreciate the difficulty in designing an experiment that perfectly isolates these two functionally distinct memory systems, given the authors question of interest and their hypothesis, I believe this is necessary and is only partially achieved by the current design. Below are further concerns and suggestions that provide details on these major points:

- For the inclusion criteria, I would suggest that the authors keep a narrower age range. Although I understand the difficulty in data collection when setting such restricted criteria, in this particular instance it might be beneficial to select participants who all have had a similar level of experience with the world and exposure to different brands throughout their lifetime. For example, a 35-year-old will only have started using Google in the second half of their life, whereas an 18-year-old will have grown up with it. Similarly, a 35-year-old might be more familiar with household cleaning products than an 18-year-old. As such, trying to homogenize prior experience would help reduce age related biases in experience and subsequent retrieval and might help with the post-hoc division of successful/unsuccessful retrieval.
- As for the exclusion criteria, specifically for motion, it might be best to define a predetermined threshold for replicability instead of one that is dependent on data collection (e.g. movement greater than 0.1 mm in any direction etc.)
- Given that the authors are trying to reach 50% accuracy in line with their design, I believe that a 10/90% threshold is too low i.e. if they end up with lots of participants performing at 15% then this would refrain the authors from carrying out a meaningful analysis of the data (at least in the manner that the authors intend to pursue).
- I was not sure why the authors have chosen to have a different number of trials for the semantic versus episodic tasks? Would this not affect the degrees of freedom for the subsequent analyses (in addition to the differences that will exist due to post-hoc partition of trial types)?
- Although I generally understand the logic behind the control task, I am not sure if it is fit for purpose in this instance. Although, as per the author’s statement, this task might reveal general success effects, this still does not eliminate the possibility for an interaction between success and memory retrieval. In other words, what is considered “success” might be different between semantic and episodic memory retrieval, which might also be slightly different than what is being measured in the control task (which I believe does not require memory retrieval).
- Related to the above point, the logos/brands are divided into familiar and unfamiliar versions. Thus, one would assume that the unfamiliar logos will be more “unsuccessfully” retrieved. As such, the successful/unsuccessful division will be confounded by familiarity and strength of association. This is particularly problematic if there is a systematic difference in how familiar the participants are with semantic versus episodic paired associates (i.e. one is dependent on lifelong experience, whereas the other is based on experimental encoding with no manipulation in this regard).
- Given the experimental design, I am concerned about whether the authors will be able to achieve the level of isolation between the two memory systems that is required. For example, how will the authors ensure that the participants are not retrieving episodic information (e.g. last time they had a pop) when they see a Coca Cola logo. I believe, it will be difficult to assess in the episodic task whether they are indeed retrieving the episodic association that they created, personal semantic/episodic association to the brand, or just semantic information. Similarly, would it not be possible that semantic information (e.g. cars) is retrieved when a Totoya logo is shown in the episodic task?
- In the behavioral piloting, were 5 seconds enough to generate meaningful stories? It might be good for the participants to verbally generate the stories, which can be parsed for detail and richness, and used to confirm later assessments in the test phase.
- Although I understand the logic behind the debriefing session, I am concerned that one needs to know sufficient information about the logo/brand in order to be able to create a story associated with it in the first place. This might be particularly problematic if there are unfamiliar logos/brands.
- For replicability, I would urge the authors to also state what type of head-coil will be used.

- For the multi-echo, multi-band sequence, the slice thickness is given as 3 mm, yet the voxel size is indicated as 2 x 2 x 2 mm.
- Although I am familiar with the use of multi-echo sequences to better detect motion-related artifacts, I am not sure if it can better recover signal in problematic regions such as the ATL. Are there any prior publications on this that the authors can cite? Another reason why I am asking this question is that although ME sequences do have certain advantages, the analysis pipelines (including me-ica and tedana) are not as well-documented, tested and updated as conventional analysis pipelines that employ data from gradient echo EPI sequences. Furthermore, would TE based removal of signal not influence degrees of freedom?
- Related to the above point, do the authors have any strategies in testing tSNR, at least across the ROIs that they intend to compare?
- Will there be a special prescription of slices that align with the temporal lobes? This tends to recover the signal better from ATL but might lead to loss in other regions.
- In the preprocessing steps the authors indicate that they will use an 8 mm FHWM Gaussian kernel for smoothing. Would this not affect the subsequent representational similarity analyses?
- For the ROI definition, I am not sure about the logic behind this approach. The authors state that they would like to assess if brain-based representations of semantic and episodic memory differ from one another. But then they already pre-define different ROIs for each memory type that will be maximally separate (with some potential overlap). Would this not bias the results in finding different representations if the regions are pre-defined as different in the first place? In other words, I am not sure why the authors have not opted for using a whole-brain analysis.

Overall, this is a highly interesting topic, but I would strongly urge the authors to re-think their experimental design and analysis strategies in order to be able to test their question of interest and the accompanying hypothesis.

Reviewer #3:
Remarks to the Author:
I was asked to specifically focus on the fMRI analysis and sequential approach to sample design.
In general these seem sophisticated and very appropriate analyses, particularly the sequential stopping approach to sample size. However, there I have a few comments and there are a few points where the analysis plan is underspecified, as detailed below.

The Bayesian linear mixed-models proposed for behaviour and neuroimaging data are well justified. What about specific stimuli related item effects, could/should these be modelled (e.g., as random effects) as well?

Regarding the sequential sample size procedure, if I have understood it correctly, this is a really nice approach. However, there are a few points I was not clear on: what is the batch size after which a decision will be taken as to whether sufficient subjects have been collected (e.g., 20 or 10 participants)? What is the total N, at one point it says 60 but in the simulations it appears to be 100. Most critically, I wasn't very sure what hypothesis the effect of the stopping criteria is on the hypotheses being tested: particularly how the multiple tests are aggregated in the simulation and how the BF takes this into account. Isn't it possible that one ROI will have BF>10 or <1/6 but all the other aren't? Equally, the similarity analysis could be BF>10 and the univariate not etc.

Version 3:

Decision Letter:

23rd July 2021

Dear Dr Tibon,

Thank you once again for your manuscript, entitled "Do activations and representations differ for episodic versus semantic memory?," and for your patience during the peer review process.

Your manuscript has now been evaluated again by the 3 reviewers, and I'm pleased to say that they are all positive about the project and have no further concerns. However, in reviewing your manuscript we have noted some editorial concerns.

1) Although you correctly set BF = 10 as the threshold for accepting the alternative hypothesis, you specify BF = 1/6 for accepting the null hypothesis. At NHB we require BF = 1/10 in this case. You should set this to 1/10 and rerun the simulations.

2) Better justification is needed for the estimated effect size (medium: 0.5). At NHB we require that RR's provide estimated effect sizes with reference to the existing literature or the smallest theoretically meaningful effect size. If based on the prior literature, since publication bias overinflates published estimates of effect size, power analysis must be based on the lowest available or meaningful estimate of the effect size. You will need to provide compelling justification that 0.5 is the smallest theoretically or practically meaningful effect size.

3) The minimum number of participants is stated as N=20. With a target BF10 = 10, we are concerned that completing the study with N=20 would lead to an unacceptable level of potential false positives (see Schönbrodt & Wagenmakers (ref 50)). You would need to commit to a minimum N of 40.

4) The hypotheses should be stated in the intro.

5) In the design table, the last column must contain specified interpretations both for results that support the alternative hypothesis and results that support the null hypothesis. Please also be careful of terminology, e.g. results themselves can't be null in the context of Bayesian statistics, although they can support the null hypothesis with a certain BF.

I apologize that these issues were not raised earlier. As soon as these issues have been addressed, we will be ready to accept this manuscript. Please do not hesitate to contact me if you have any questions. I would be happy to discuss these requests further.

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 hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere.

Nature Human Behaviour is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
[redacted]

[redacted]
Nature Human Behaviour

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I feel that the points raised in the previous round of review have been satisfactorily addressed. I have no other comments. I am looking forward to hearing about the results.

Reviewer #2:
Remarks to the Author:
I thank Tibon and colleagues for their thorough reply to my prior concerns, which have now been diligently addressed. I wish them all the very best during data acquisition and analyses, and look forward to reading about the results of this highly interesting study in due course.

Kind regards,

Deniz Vatansever

Reviewer #3:
Remarks to the Author:
The authors have addressed my queries comprehensively and I support acceptance.

Version 4:

Decision Letter:

15th September 2021

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Dr Tibon,

Thank you once again for submitting your revised Stage 1 Registered Report, entitled "Do activations and representations differ for episodic versus semantic memory?."

I apologize for the delay in sending this decision. Everything is in order and I am delighted to say that we can offer acceptance in principle. You may progress to Stage 2 and complete the study as approved.

However, we would like you to make one final change to the protocol, to address the issue of the Bayes Factor threshold for the null hypothesis. While we appreciate your arguments in favor of BF=1/6 for this threshold, our policy does require BF=1/10 to accept the null. However, we appreciate that this would require an unrealistic max N.

Therefore, we ask you to do the following as a solution:

1) State in the text that BF=1/10 is the threshold to accept the null hypothesis, and re-run the power simulations accordingly.
2) State that the max N was chosen as the largest practical value possible (i.e. you do not need to change your current max N, but explain it)
3) State that if the BF does not reach 1/10, you will not be able to interpret it as evidence for the null hypothesis.

As you know, a condition of in-principle-acceptance is that the authors agree to deposit their Stage 1 accepted protocol in a repository, either publicly or under embargo until Stage 2 acceptance and publication. We are very keen to showcase our in-principle accepted protocols, so that our readers, reviewers, and potential authors can gain insight into the requirements of the format as well as an idea of the types of projects that are suitable for publication in Nature Human Behaviour. We have set up a space on figshare (https://springernature.figshare.com/registered-reports_NHB) to host all of our in-principle accepted protocols, which can either be made public or kept under embargo until Stage 2 acceptance (depending on author preference). This gives you the opportunity to have your work publicly associated with Nature Human Behaviour, and of course we will be very pleased to showcase your report if you agree to share it publicly.

Depositing the work on our figshare space does not preclude deposition of your Stage 1 protocol on other depositories – your protocol can also be posted on OSF, Dataverse, Dryad or any other public repository of your choice. You also do not need to do anything – if you agree with posting your protocol on our figshare space, we will upload your protocol on your behalf and either set it public or place it under embargo, depending on your choice. Your protocol will be licensed under a CC BY license (Creative Commons Attribution 4.0 International License). The CC BY license allows for maximum dissemination and re-use of open access materials and is preferred by many research funding bodies. Under this license users are free to share (copy, distribute and transmit) and remix (adapt) the contribution including for commercial purposes, providing they attribute the contribution in the manner specified by the author or licensor (read full legal code: http://creativecommons.org/licenses/by/4.0/legalcode) Please note that any use of https://springernature.figshare.com will be subject to the Figshare terms of use. Figshare has the right to enforce these terms and conditions where applicable. Use of third party services and sites will be subject to the relevant terms of use and will apply if we act on your behalf in this regard. Do let me know if you would like to take up this option or if you have any questions regarding the protocol deposition requirement.

Following completion of your study, we invite you to resubmit your paper for peer review as a Stage 2 Registered Report. Please note that your manuscript can still be rejected for publication at Stage 2 if the Editors consider any of the following to hold:

• The results were unable to test the authors' proposed hypotheses by failing to meet the approved outcome-neutral criteria
• The authors altered the Introduction, rationale, or hypotheses, as approved in the Stage 1 submission
• The authors failed to adhere closely to the registered experimental procedures
• Any post hoc (unregistered) analyses were either unjustified, insufficiently caveated, or overly dominant in shaping the authors' conclusions
• The authors' conclusions were not justified given the data obtained

We encourage you to read the complete guidelines for authors concerning Stage 2 submissions at https://www.nature.com/nathumbehav/registeredreports. Please especially note the requirements for protocol deposition, data sharing, and that withdrawing your manuscript will result in publication of a Retracted Registration.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Do activations and representations differ for episodic versus semantic memory?". For those reviewers who give their assent, we will be publishing their names alongside the published article.

When you are ready, please use the following link to access your home page and submit your Stage 2 Registered Report:

Link Redacted

*This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first.

We expect your Stage 2 Registered Report to be submitted by the date specified in your latest cover letter. If unforeseen circumstances prevent submission by that date, please contact us as soon as possible to discuss any changes to the submission time-frame.

Thank you again for offering us this work and we look forward to receiving your Stage 2 Registered Report.

Yours sincerely,
[redacted]

[redacted]
Nature Human Behaviour

Version 5:

Decision Letter:

19th December 2023

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Dr Tibon,

Thank you for submitting your Stage 2 Registered Report, entitled "Do activations and representations differ during successful retrieval from episodic versus semantic memory?". Congratulations on the completion of the project!

I'm afraid that we will require some changes to the Stage 2 report before we can send it back to the reviewers for Stage 2 review.

We noticed that you have made a number of deviations from the protocol in terms of the data analysis. For Registered Reports, the expectation is that authors carry out their work following exactly the procedures specified unless this is not technically possible. It appears that at least some of these deviations were introduced to improve the analysis, rather than because the original analysis would not be possible.

Therefore, we ask that you carry out the analysis exactly as preregistered. This should be the main analysis. You can report the results of the modified analyses, but only as exploratory results.

If it would not be technically possible to carry out the preregistered analysis as described, or if you believe the original analysis is fatally flawed (not merely suboptimal), you should still carry out the original analysis as far as possible with only the minimum necessary deviations, before reporting the results of your preferred analysis. We will ask reviewers to comment on all deviations from the original protocol and the impact they have on the project.

Please let me know if you have any questions about these requirements.

Please use the following link for uploading these materials:
Link Redacted

If you have any further questions, please feel free to contact me.

With best regards,
[redacted]

[redacted]
Nature Human Behaviour

Version 6:

Decision Letter:

26th February 2025

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Dr Tibon,

Thank you once again for submitting your Stage 2 Registered Report, entitled "Do activations and representations differ during successful retrieval from episodic versus semantic memory?," and for your patience during the re-review process.

Your manuscript has now been evaluated by Reviewers #1 and #2, from the previous rounds of review, whose comments are included at the end of this letter. In the light of our reviewers' advice, we are pleased to inform you that we will be able accept your Stage 2 manuscript, pending revisions to address reviewer comments and editorial requests. We will share our detailed editorial requests within two weeks.

Nature Human Behaviour offers a transparent peer review option for new original research manuscripts submitted from 1st December 2019. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

To submit your revised manuscript (**once we have shared our detailed guidance**), you will need to provide the following:
• Cover letter
• Point-by-point response to the reviewers (if applicable)
• Manuscript text (not including the figures) in .docx or .tex format
• Individual figure files (one figure per file)
• Extended Data & Supplementary Information, as instructed
• Reporting summary
• Editorial policy checklist
• Third-party rights table (if applicable)
• Suggestions for cover illustrations (if desired)

Consortia authorship:
For papers containing one or more consortia, all members of the consortium who contributed to the paper must be listed in the paper (i.e., print/online PDF). If necessary, individual authors can be listed in both the main author list and as a member of a consortium listed at the end of the paper. When submitting your revised manuscript via the online submission system, the consortium name should be entered as an author, together with the contact details of a nominated consortium representative. See https://www.nature.com/authors/policies/authorship.html for our authorship policy and https://www.nature.com/documents/nr-consortia-formatting.pdf for further consortia formatting guidelines, which should be adhered to prior to acceptance.

Forms:
Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Once your paper is accepted, you will receive an email in approximately 10 business days providing you with a link to complete the grant of rights. If you choose to publish Open Access, our Author Services team will also be in touch at that time regarding any additional information that may be required to arrange payment for your article.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
Link Redacted

If you have any further questions, please feel free to contact me.

With best regards,
[redacted]

[redacted]
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

This is a timely and interesting manuscript on the overlap in the neural correlates of semantic and episodic memory. The experimental design is clever and original and the manuscript well-written and well-presented. The deviations from the registered protocol were minimal and well-justified. The pilot study to determine whether logo names were known (nameable) was somehow underpowered (N=8), but part of the original submission. I have a number of clarification questions for the authors.

Could the use of multi-echo fMRI have impaired hippocampal activation detection and therefore some specificity in episodic memory activations?

Could the authors comment on excluded ROI? (supplementary table 1) I was not sure about the reasons for this. The main

manuscript mentions excluding overlapping voxels between the core recollection network and the semantic network, but I would not have thought that temporal pole would have been part of this for example?

Comments on ATL findings would be useful (was it analysed or excluded) I thought the motivation of using multi echo fMRI was in part to capture ATL (the authors also rightly commented in the introduction on the poor coverage of typical episodic memory studies in this region)

Some clarifications about the semantic measures would be useful. Was success retrieval for semantic memory simply logo name retrieval or actual semantic knowledge about the logo? Did the authors check the details provided by participants were actually accurate? ("detailedness" analysis) Did the authors compare semantic and episodic memory retrieval activations when matched in number of details retrieved?

Any reasons why right hippocampus (observed with less severe correction) would be more robust than left?

Since stage 1, new literature relevant to continuum models of declarative memory has been published:
Addis & Szpunar, 2024, Philosophical Transactions, B
Curot et al., 2024, Cortex.
Tanguay et al., 2023, eLife

Reviewer #1 (Remarks on code availability):

Relevant data appears to be accessible on OSF.

Reviewer #2 (Remarks to the Author):

In this study, Tibon and colleagues investigated potential differences in the activations and neural representations of episodic and semantic memory. Having previously reviewed the Stage 1 submission of this registered report, I am delighted to see the research results in this Stage 2 submission. Overall, this is a carefully thought-out experiment on a topic that will undoubtedly interest the scientific community and beyond.

The introduction section, including the study rationale and stated hypotheses, remains the same as the approved Stage 1 submission. The authors clearly state a few minor deviations from the registered protocol, which I believe would not substantially influence the reported results.

The authors adhered to the registered experimental procedures and clearly outlined their exploratory analyses in the Results section. These unregistered analyses are justified, methodologically sound, and informative, especially given the null finding.

The behavioral analyses demonstrate that the stimuli selection and paradigm design procedures worked successfully, with a relatively equal number of successful trials across conditions. This yielded sufficient trial-wise exemplars for the authors to test their hypothesis. Therefore, I believe their conclusions are justified by the data.

However, given the authors' findings I would have expected a discussion of the null results in relation to prior studies that investigated this research question, albeit using other experimental paradigms (e.g., Burianova et al, Vatansever et al, Humphreys et al). Instead, these works appear only briefly in the introduction section, which understates the research conducted before this registered report. While "absence of evidence does not constitute evidence of absence," theoretical suggestions for a unitary retrieval system warrant discussion in the context of these findings. The authors devoted considerable attention to a small number of observed dissociations based on exploratory rather than pre-registered analyses. I urge the authors to revise their discussion to emphasize the null results from pre-registered analyses rather than findings from exploratory analyses.

Regarding style, the manuscript contains numerous parenthetical asides within long sentences. To improve readability, I suggest reducing this practice. While I appreciate the detailed information provided, certain sections, such as the lengthy description of sample size determination procedures, could move to the supplementary notes, as they are not central to the experimental work and conclusions. Finally, I recommend improving text visibility in certain figures (e.g., Figure 1).

Reviewer #2 (Remarks on code availability):

The authors have made their code publicly available which is extensively commented and includes clear instructions for usage in the README file.

Version 7:

Decision Letter:

Dear Dr Tibon,

We are pleased to inform you that your Stage 2 Registered Report "Neural activations and representations during episodic versus semantic memory retrieval", has now been accepted for publication in Nature Human Behaviour. I sincerely apologize again for the delays in processing this manuscript.

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 <a

href="https://www.springernature.com/gp/open-science/plan-s-compliance"> Plan S principles or the NIH public access policy) then you should select the gold OA route, and we will direct you to the compliant route where possible. Because authors warrant under our subscription licensing terms that they haven't committed to licensing any version of their article under a licence inconsistent with the terms of our agreement – including the applicable embargo period – publication under the subscription model isn't suitable for authors whose funders require no embargo.

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,

[redacted]

[redacted]

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.**

Response to Reviewers

Reviewer #1:

This is a timely and interesting project on the overlap in the neural correlates of semantic and episodic memory.

We thank the reviewer for their positive remarks, and have addressed their concerns in this revised submission.

The design is definitely interesting but might need to be tweaked or clarified.
Some brand logo pairs will be familiar while others will not. It is mentioned that it is expected that only half of the logo-word pairing will be known to participants. Is this based on the pilot data already collected?

Yes. The initial selection of stimuli was based on familiarity ratings collected in a pilot study. The raw data from this pilot are available on the project's OSF page (https://osf.io/dm47y/?view_only=aabac55776f9426684aa1cf5885aba25; Pilot Data/StimPilot.xlsx). The averaged familiarity ratings and recall scores were matched across the various experimental lists, and are also available on OSF (Stimuli/LogoStimLists.xlsx). Notably, due to our trial-based (rather than averaged-based) statistical modelling, it is not crucial to have a 50/50 distribution (i.e., 50% success rate). The main aim of this procedure was to ensure that (1) the episodic and semantic tasks are roughly equated for task difficulty, and (2) that we have enough trials is each bin to allow our proposed analyses. This was further confirmed by a second pilot, in which the paradigm was tested (data summarised in PilotData_Summary.xlsx), which indicated that although accuracy rates in the semantic task fell slightly below 50% ($M_{success}$ = 39.32, $M_{failure}$ = 58.6, M_{false_alarms} = 1.67), they were roughly matched with those of the episodic task ($M_{success}$ = 31.37, $M_{failure}$ = 57.22, M_{false_alarms} = 11.57), and were sufficiently high to allow the analysis of success and failure trials (false alarm trials are to be excluded).

In the episodic task, it is indicated that pairing will be established in the study phase. Will all logo-word pairings be unfamiliar/new to participants? This would create a mismatch with the semantic task (where half of the logo-word are assumed to be familiar).

*The same pool of logos and brands will be used in the semantic task and the episodic task. Experimental lists will be counterbalanced across participants in order to match the items used in these tasks. Importantly, in the episodic task, all the **pairings** will be new, and participants will be asked about the novel (episodic) pairing, rather than about the original (semantic) pairing. Probing novel vs. established pairings is the experimental manipulation rather than a confounding mismatch, as it taps into episodic and semantic processes respectively.*

Or, if already known, will some of the pairing conflict with prior semantic knowledge (e.g., Nasa matched with the word Google, as cited page 9)? If the later, this would also be an issue (to measure episodic recollection effects for mismatches/violations of semantic knowledge) and is likely to produce some distinct activations for the semantic and the episodic conditions. I understand that participants' instructions would be to make a narrative linking each logo with a different brand name, but this may still be hard to compare with the semantic task including half

congruent semantic knowledge presumably acquired over a long time and new unfamiliar associations. One possibility might be to focus on pairs unknown to the participant, and to compare the initially unknown pairs of the semantic task with unknown pairs used in the episodic task (i.e., logo and brands unknown to participants), but I am not sure that the encoding phase in the semantic task allows to obtain this information and then the tasks would become presumably identical or very similar.

We are thankful for the reviewer for bringing up this important point, which we have considered thoroughly in our initial design. However, given that this concern was raised by both Reviewers 1 and 2, we realise now that we did not provide sufficient details on this issue in our original submission.

As noted by the reviewer, in the episodic tasks, for some logos participants are likely to know the original (semantic) pairing which might cause semantic interference when creating new pairing, whereas for others the original pairing will be unfamiliar. Therefore, to minimize semantic processing / interference in the episodic task, we will utilize the data obtained in the debriefing session, and will exclude trials corresponding to logos that were classified as "can name" during debriefing. We do note, however, that in order to increase statistical power, we will run and report all analyses both with and without these trials, and if numerical trends remain the same, these trials will be included in our main analyses (which determine the stopping criteria for the experiment). We now clarify this on p11 of our revised report:

> *Data from the debriefing phase will be used to inform our analyses. Namely, in the episodic task, participants are likely to know the original pairing for some logos, which might cause semantic interference when the new (episodic) pairings are formed. To minimize this potential confound, trials that were classified as "can name" during the debriefing session will be initially excluded from the analyses (although they will be reinstated if they conform to the same numerical trends, see Analysis Plan section below). To allow sufficient statistical power after the potential exclusion of these trials, the episodic task will include more trials than the semantic task (144 vs. 96 trials), with the number of trials-per-block (48) remaining constant across tasks.*

And on p15:

> *As was mentioned above, in order to minimize semantic processing in the episodic task, trials corresponding to logos that were classified as "can name" in the debriefing session (regardless of whether or not the correct response was provided) will also be excluded from the analyses. Nevertheless, to increase statistical power, we will run and report all analyses both with and without these trials, and if numerical trends remain the same, these trials will be included in our main analyses (which determine the stopping criteria for the experiment, see below).*

How will task difficulty be matched? Some brain regions that have been proposed to be involved in both semantic and episodic processing like the angular gyrus also appear sensitive to task difficulty.

Pilot data (now reported on p13 of the revised manuscript, and also available in more details on the project's OSF page) were collected in order to ensure that the tasks are (roughly) matched for difficulty. On top of that, two aspects of our proposed analytical approach were designed to minimize this issue. First, our proposed contrasts are all based on success effects (i.e., the difference between

"success" and "failure" trials). Although differences in difficulty levels would affect the distribution of trials (i.e., potentially more successful trials in one task than the other), these effects would be orthogonal to our comparisons. In other words, success trials should be similarly easy across tasks, and failure trials should be similarly difficult. Therefore, the trials themselves will be matched for difficulty, regardless overall performance in the task. Second, we will test for overall task effects by further comparing activation for success and failure trials in each task against baseline periods in our ROIs. This analysis (mentioned on p18) would capture any remaining effects that are due to task difficulty, and would facilitate the interpretation of the results.

Taken together, the combination of our design (based on pilot data indicating roughly matched success) and our analytical approach (contrasting success versus failure, and comparison against a baseline period), should address the major concerns about task difficulty.

Why do testing on 2 different days? Is it not likely to maximise differences in behaviour and in brain activity?

Together with all the instructions and preparations, the total running time for the experiment is ~4.5h (including ~2.5h of scanning sessions that need to be separated into two parts, to allow for the un-scanned study/familiarity phases). Therefore, performing the whole experiment in one day is unrealistic, and might arguably result in increased behavioural effects of fatigue. As noted above, however, all comparisons are based on success effects (contrasted within the same session) and against a within-session baseline period, thus minimizing between-session effects on brain activity.

Reviewer #2:

In this project proposal, Tibon and colleagues are seeking to investigate whether dissociable brain representations exist in the successful retrieval of semantic and episodic memory. For that purpose, the authors propose a within-subject fMRI study in which they aim to employ two separate tasks that are designed to tap into the two memory systems in a relatively large group of subjects. The data will then be analyzed employing both conventional task-based activation approaches as well as multivariate representational similarity analysis.

This is a well-written project proposal and a highly interesting topic that will undoubtedly be of interest to the scientific community at large. However, my initial enthusiasm is somewhat dampened by the research question and the hypothesis put forward as well as the experimental design that is aimed to test this hypothesis. I provide detailed points below in the hopes that they further improve the study design.

We are grateful for the positive comments and thorough feedback, and have addressed the Reviewer's concerns in this revision.

To begin with, I have three major concerns that I would like to emphasize: First, the authors state their research question as "Can we distinguish successful episodic retrieval from semantic retrieval?" As the authors highlighted themselves, there is ample amount of evidence in the literature to suggest dissociable brain-based representations of these two memory types, generally centered on the ATL and MTL for semantic and episodic memory, respectively. As such, if

this is the authors' aim, then the results of this study will only confirm what is already known for the past two decades of neuroimaging-based research. Hence, the only novelty lies in the multi-session aspect of this experiment in which the two memory systems are probed within the same individuals (traditionally they are studied in isolation).

We have thoroughly revised the introduction to address this issue. On p3-4 of the revised manuscript (additions are highlighted below), we now explain how the current study goes beyond what is already known, and why within-study comparisons are necessary to establish the distinction between episodic and semantic memories:

> *"In many ways, this distinction has dominated research on the cognitive neuroscience of memory, and remains of central importance to the field. It is supported by a large corpus of neuropsychological studies, including studies of neurodegenerative disorders and selective brain lesions. For example, early stages of Alzheimer's disease (AD), which are characterized mainly by medial temporal lobe (MTL) degeneration, typically produce deficits in episodic memory, whereas semantic memory remains intact until later stages of the disease[3,4] (see also findings of intact ability to acquire new semantic knowledge in developmental amnesia[5]). Conversely, semantic dementia (SD), a younger-onset neurodegenerative disorder that is associated with degeneration that starts in the anterior temporal lobe (ATL), produces a multi-modal loss of semantic knowledge, while episodic memory appears relatively spared until later in the disease[6–9].*
>
> *Despite this historical distinction between episodic and semantic processing, established via neuropsychological and clinical studies, the picture arising from neuroimaging studies is much less conclusive. Indeed, neuroimaging studies with healthy human participants seem to challenge the episodic-semantic dissociation, by revealing considerable overlap in the brain regions involved in semantic and episodic processing[10]. Episodic memory has been associated with enhanced activity in a consistent set of brain regions (the "core recollection network"), regardless the nature of the recollected content[11,12]. Similarly, semantic memory has been associated with a core-network that is responsive to tasks that require semantic memory retrieval, e.g. word vs. non-word processing (the "general semantic network"[13]). Importantly, these two networks have extensive overlap, specifically in parahippocampal, middle temporal, inferior parietal, posterior cingulate/precuneus, and midline frontal regions[10]. These findings inspired other alternative perspectives, suggesting that episodic and semantic memories lie along a continuum of contextualization within time, space and valence that is processed by a unitary memory system[10], or that the dichotomy between the systems should be replaced with a continuum spanning from concrete to abstract representations[14]. Such proposals, which have received increasing attention in recent years, further prompt the reconsideration of evidence from clinical populations, suggesting that the boundaries between the two memory systems might be unclear[15].*
>
> *Because, for many years, episodic and semantic memories were considered to be distinct entities, a research tradition had developed in which they are explored separately. The consequence of this is a lack of within-study designs that tap into both systems, which, ironically, prevents adequate rebuttal to the unifying theories described above. Reviews and meta-analyses cannot provide conclusive evidence either: a spatial overlap might lead to the conclusion that there is a functional overlap, but might also be the result of blurring small-scale variability across studies. Similarly, a spatial segregation might lead to the conclusion that the systems are distinct, but may also result from experimental or procedural variations (e.g., different scanning parameters designed to maximize ATL or MTL signal for semantic vs.*

> *episodic tasks). Therefore, in order to provide compelling evidence supporting one view or the other, within-study (and preferably, within-participant) comparisons are crucial. The few studies that have compared episodic and semantic memory within the same experimental design have tended to observe extensive overlap of functional networks/brain regions…"*

Moreover, the authors state their hypothesis as "The brain regions supporting successful episodic retrieval are distinct from those supporting successful semantic retrieval." Testing this hypothesis requires complete behavioral isolation of the two functionally (and mostly experientially) distinct memory systems. Unfortunately, I do not believe that the proposed project achieves this level of separation.

We assume that the Reviewer's comment regarding the behavioural isolation refers to the interference of semantic processes (due to familiarity with the original pairing) as detailed in their additional comments below. We are thankful to the reviewer for bringing up this important point. A similar point was raised by Reviewer 1, arguing that semantic processes might interfere with the episodic task, thereby precluding behavioural isolation. This point was considered thoroughly in our initial design, but we realise now that we did not provide sufficient details on this issue in our original submission. To clarify this issue, in the episodic task, for some logos participants are likely to know the original (semantic) pairing which might cause semantic interference when creating new pairing, whereas for others the original pairing will be unfamiliar. Therefore, to minimize semantic processing / interference in the episodic task, we will utilize the data obtained in the debriefing session, and will exclude trials corresponding to logos that were classified as "can name" during debriefing. We do note, however, that in order to increase statistical power, we will run and report all analyses both with and without these trials, and if numerical trends remain the same, these trials will be included in our main analyses (which determine the stopping criteria for the experiment). We now clarify this on p11 of our revised report:

> *Data from the debriefing phase will be used to inform our analyses. Namely, in the episodic task, participants are likely to know the original pairing for some logos, which might cause semantic interference when the new (episodic) pairings are formed. To minimize this potential confound, trials that were classified as "can name" during the debriefing session will be initially excluded from the analyses (although they will be reinstated if they conform to the same numerical trends, see Analysis Plan section below). To allow sufficient statistical power after the potential exclusion of these trials, the episodic task will include more trials than the semantic task (144 vs. 96 trials), with the number of trials-per-block (48) remaining constant across tasks.*

And on p15:

> *As was mentioned above, in order to minimize semantic processing in the episodic task, trials corresponding to logos that were classified as "can name" in the debriefing session (regardless of whether or not the correct response was provided) will also be excluded from the analyses. Nevertheless, to increase statistical power, we will run and report all analyses both with and without these trials, and if numerical trends remain the same, these trials will be included in our main analyses (which determine the stopping criteria for the experiment, see below).*

Furthermore, as per the authors predictions, finding that "greater recall success effect in the episodic task in episodic ROIs, but in the semantic task in semantic ROIs" would not allow the authors to refute the stated null hypothesis that the brain regions supporting successful episodic and semantic retrieval are shared, but rather show that there are both common and distinct neural features, which is exactly what is already suggested in the existing literature.

In the current study, the null hypothesis is that there is no qualitative difference in activation between the brain regions involved in semantic and episodic memories. The alternative is that there is some difference, even if some regions do show an overlap. Indeed, even if assuming that semantic and episodic processes rely on entirely different brain regions, we would still expect some level of shared activation due to additional supporting processes (e.g., visual processing). To be able to argue that the processes are distinct, even though some shared neural activation is expected, we utilize a forward inference procedure (i.e., the use of qualitatively different patterns of activity over the brain to distinguish between competing cognitive theories; in this case, single vs. dual process theory), by first contrasting success effects in the critical task with those obtained in the control task, and then contrasting success effects obtained in the two critical tasks. In this revised version of the manuscript, we also included ROI as a factor in our statistical models, to support our ability to infer qualitative differences (see Henson, 2006: https://doi.org/10.1016/j.tics.2005.12.005 [Figure 1] for details). As was explained above, establishing a dissociation within the same study, rather than based on two parallel bodies of work, is essential and goes beyond what is already suggested in the existing literature.

The same issue goes for the representational similarity analysis, but here the authors should also consider what they are defining as "representation" versus "retrieval" (I believe the question is framed around successful retrieval and not representation).

We reworded the text to indicate that the similarity analysis is conducted for representation of retrieved associates, see p4.

> *"To our knowledge, no previous study has directly compared multivoxel patterns across various regions within these core-systems (one study examined multivoxel pattern of episodic and sematic memories, but focused on the angular gyrus and sensory areas[23]), which can potentially capture differences in the representations of retrieved episodic versus semantic memories."*

In addition, the authors aim to focus their analysis on successful versus unsuccessful retrieval, which I believe will introduce a confound to the same hypothesis that the authors are trying to test, especially if levels of successful retrieval are different across the two memory systems.

Pilot data (now reported on p13 of the revised manuscript, and also available in more details on the project's OSF page) were collected in order to ensure that the tasks are (roughly) matched for difficulty. Furthermore, an important reason for using a recall success contrast (success-failure) is that it minimizes (rather than increases) potential confounds associated with task difficulty. Namely, although differences in difficulty levels would affect the distribution of trials (i.e., potentially more successful trials in one task than the other), these effects would be orthogonal to our comparisons. In other words, success trials would be similarly easy across tasks, and failure trials would be similarly difficult. Therefore, the trials themselves will be matched for difficulty, regardless overall performance in the task. In addition, we will test for overall task effects by further comparing

activation for success and failure trials in each task against baseline periods in our ROIs. This analysis (mentioned on p18) would capture any remaining effects that are due to task difficulty, and would facilitate the interpretation of the univariate results. Taken together, the combination of our design (based on pilot data indicating roughly matched success) and our analytical approach (contrasting success effects and comparison against a baseline period), minimizes any potential confounds that are due to task difficulty.

Taken together, although I wholeheartedly appreciate the difficulty in designing an experiment that perfectly isolates these two functionally distinct memory systems, given the authors question of interest and their hypothesis, I believe this is necessary and is only partially achieved by the current design.

We thank the reviewer again for bringing these important issues to our attention, and hope that our responses above clarify how the current design goes beyond what is already known, and has the potential to test definitively whether any pair of ROIs differ reliably for episodic and semantic retrieval, while controlling for as much as possible (and certainly controlling more than previous studies).

Below are further concerns and suggestions that provide details on these major points:

For the inclusion criteria, I would suggest that the authors keep a narrower age range. Although I understand the difficulty in data collection when setting such restricted criteria, in this particular instance it might be beneficial to select participants who all have had a similar level of experience with the world and exposure to different brands throughout their lifetime. For example, a 35-year-old will only have started using Google in the second half of their life, whereas an 18-year-old will have grown up with it. Similarly, a 35-year-old might be more familiar with household cleaning products than an 18-year-old. As such, trying to homogenize prior experience would help reduce age related biases in experience and subsequent retrieval and might help with the post-hoc division of successful/unsuccessful retrieval.

We concur with the reviewer that there might be some differences between the stimuli, owing to the level of experience one has with a particular brand. Notably, such differences in the semantic representation of the brands do not introduce a confound to the semantic/episodic distinction, but rather adds some "noise" to the semantic task (that is, if a distinction still emerges, that will be despite these differences, rather than due to them). In terms of individual differences between participants, the post-hoc division of trials into experimental bins (success / failure), which is done on a per-subject basis (that is, for each participant we contrast their own individual successes and failures), should minimize concerns related to individual differences. Nevertheless, in order to further alleviate any potential interactions between age and specific items, we will pseudo-randomly assign participants into the various counterbalancing conditions, so that the mean age across the various experimental lists is roughly equated. This is now mentioned on p6:

> *"Participants will be pseudo-randomly assigned into the various counterbalancing conditions, in order to keep the mean age across the various experimental lists roughly equated."*

As for the exclusion criteria, specifically for motion, it might be best to define a predetermined threshold for replicability instead of one that is dependent on data collection (e.g. movement greater than 0.1 mm in any direction etc.)

Different types of motion have different consequences for fMRI artifacts (e.g., within vs between session, gradual vs sudden, etc.), and importantly, the amount of motion depends on the paradigm (task length, number of sessions, participant age, experience, etc.). Therefore, in our previous studies, we did not specify absolute motion criteria which we believe to be inappropriate. Instead, we have always adopted a relative measure of excluding participants whose motion (according to multiple measures, e.g., mean, variance, etc.) is atypical for the sample tested on that specific paradigm. Nevertheless, in addition to the relative criteria already specified, we have added the absolute criterion that motion should not be so great that key ROIs are missed (owing to movement outside field of view). We now mention this on p7:

> *"Reasons for post-analysis exclusion are: lack of trials in one or more experimental conditions (due to floor/ceiling performance), and/or increased movements (2 SDs or more than the mean motion across participants), and/or motion that resulted in reduced coverage of key ROIs (anatomical ROIs, as detailed below)."*

Given that the authors are trying to reach 50% accuracy in line with their design, I believe that a 10/90% threshold is too low i.e. if they end up with lots of participants performing at 15% then this would refrain the authors from carrying out a meaningful analysis of the data (at least in the manner that the authors intend to pursue).

*We would like to clarify that while we **expect** roughly 50% accuracy (based on pilot data), and indeed this would be the ideal distribution (as it would allow the maximal number of trials in each experimental bin), our design does not **require** that we have a 50/50 distribution (i.e., 50% success rate) within each participant. In fact, our statistical modelling, which is based on single trials rather than mean data, is particularly suitable for this kind of imbalanced designs (i.e., when the number of trials in the various experimental conditions differs across participants; see Tibon & Levy 2015, Tibon et al. 2019).*

I was not sure why the authors have chosen to have a different number of trials for the semantic versus episodic tasks? Would this not affect the degrees of freedom for the subsequent analyses (in addition to the differences that will exist due to post-hoc partition of trial types)?

This was done in order to allow sufficient statistical power even after the potential exclusion of trials for which the original (semantic) pairing is known. This is now explained on p11 of the manuscript (see also our detailed response to the Reviewer's 2nd point above).

Although I generally understand the logic behind the control task, I am not sure if it is fit for purpose in this instance. Although, as per the author's statement, this task might reveal general success effects, this still does not eliminate the possibility for an interaction between success and memory retrieval. In other words, what is considered "success" might be different between semantic and episodic memory retrieval, which might also be slightly different than what is being measured in the control task (which I believe does not require memory retrieval).

Differences between success and failure can reflect multiple factors such as emotional reaction, self-regulation, and more. Contrasting the control task with the two mnemonic tasks would control for these potential non-mnemonic effects. It is therefore crucial that the control task does not have a mnemonic component. Indeed, designing a declarative memory task that is neither semantic nor episodic is not only unfeasible (as declarative memories would be episodic or semantic by conventional definitions), but will also preclude isolation of the mnemonic component. We concur with the reviewer that there might still be differences between semantic and episodic success effects, but arguably, these are the exactly the differences of interest. This is a subtle point that we are happy to consider in the discussion of the manuscript (i.e., in Stage 2).

Related to the above point, the logos/brands are divided into familiar and unfamiliar versions. Thus, one would assume that the unfamiliar logos will be more "unsuccessfully" retrieved. As such, the successful/unsuccessful division will be confounded by familiarity and strength of association. This is particularly problematic if there is a systematic difference in how familiar the participants are with semantic versus episodic paired associates (i.e. one is dependent on lifelong experience, whereas the other is based on experimental encoding with no manipulation in this regard).

As mentioned in our response to the reviewer's 2nd comment above, episodic trials for which the logo is classified as "nameable" during the debriefing session will be initially excluded from the analysis (see our full response and edits above).

Given the experimental design, I am concerned about whether the authors will be able to achieve the level of isolation between the two memory systems that is required. For example, how will the authors ensure that the participants are not retrieving episodic information (e.g. last time they had a pop) when they see a Coca Cola logo. I believe, it will be difficult to assess in the episodic task whether they are indeed retrieving the episodic association that they created, personal semantic/episodic association to the brand, or just semantic information. Similarly, would it not be possible that semantic information (e.g. cars) is retrieved when a Totoya logo is shown in the episodic task?

*In the healthy brain, the systems operate together and one informs the other. In the context of the current study, it is important to make a distinction between item memory and associative memory of episodic/semantic information; with only the latter being probed in the current study (i.e., the task requires **associative** semantic/episodic memory). With our proposed design, episodic information might help retrieve the semantic associate, but not without actual semantic knowledge of the pairing. Similarly, semantic knowledge might facilitate episodic retrieval, but not without actual episodic knowledge. In both cases, the alternative process (i.e., semantic in the episodic task and episodic in the semantic task) can provide scaffolding (in the same way that attention or visual processing can support memory), but would not determine success. We chose to use cued-recall to preclude guessing and ensure that even if the alternative process can support memory, a correct response cannot be provided without information from the probed process. We now explain this point on p5 of the manuscript:*

> *"This cued-recall procedure will allow us to prevent guessing. Moreover, with our proposed design, episodic information might help retrieve the semantic associate, but not without actual semantic knowledge of the pairing. Similarly, semantic knowledge might facilitate episodic retrieval, but not without actual episodic knowledge. Thus, even if the alternative*

process (i.e., semantic memory in the episodic task and episodic memory in the semantic task) might provide some memory scaffolding (in the same way that attention or visual processing can support memory), it would not determine success. In other words, a correct response cannot be provided without information from the probed process. "

In the behavioral piloting, were 5 seconds enough to generate meaningful stories? It might be good for the participants to verbally generate the stories, which can be parsed for detail and richness, and used to confirm later assessments in the test phase.

The overall time allowed for story generation was 10 sec, as the participants are instructed to start generating their stories as soon as the stimuli appear. We now clarify this on p10. Our pilots confirmed that this timeframe was sufficient, as during practice trials, participants had to verbally generate their stories.

Although I understand the logic behind the debriefing session, I am concerned that one needs to know sufficient information about the logo/brand in order to be able to create a story associated with it in the first place. This might be particularly problematic if there are unfamiliar logos/brands.

The reason for the debriefing session is to allow us to exclude trials with semantic knowledge. As explained above, this is essential to ensure a behavioural separation of semantic and episodic processing. Our pilot data indicated that participants were able to generate episodic links even for unfamiliar logos. On average, out of the successfully retrieved episodic pairings, 61% were logos for which the semantically associated brand name was not recallable. Pilot data are available on the OSF project's page (https://osf.io/dm47y/?view_only=aabac55776f9426684aa1cf5885aba25; PilotData_Summary.xlsx).

For replicability, I would urge the authors to also state what type of head-coil will be used.

As we now specify on p13, we will use a 32 channel head-coil.

For the multi-echo, multi-band sequence, the slice thickness is given as 3 mm, yet the voxel size is indicated as 2 x 2 x 2 mm.

Thank you, voxel size should have been 3 x 3 x 3 mm (now corrected).

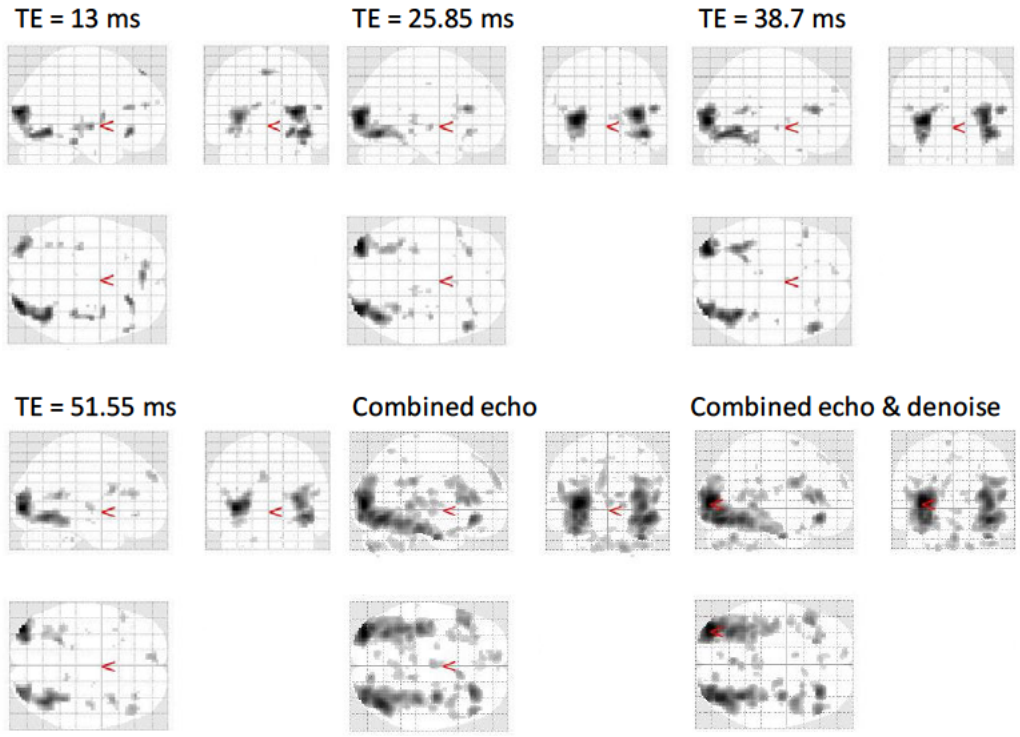
Although I am familiar with the use of multi-echo sequences to better detect motion-related artifacts, I am not sure if it can better recover signal in problematic regions such as the ATL. Are there any prior publications on this that the authors can cite? Another reason why I am asking this question is that although ME sequences do have certain advantages, the analysis pipelines (including me-ica and tedana) are not as well-documented, tested and updated as conventional analysis pipelines that employ data from gradient echo EPI sequences. Furthermore, would TE based removal of signal not influence degrees of freedom?

Several previous studies have shown that multi-echo fMRI offers advantages for imaging brain regions such as the anterior temporal lobes and the orbitofrontal gyrus, which are prone to susceptibility distortions and signal dropouts (Halai et al., 2014; Halai et al., 2015; reviewed in

Caballero-Gaudesa & Reynolds, 2017). We now provide these references on p14 of the manuscript. Indeed, this was the primary motivation of multi-echo sequences, before it was realized that a linear fit to the TEs would also enable separation of BOLD signal versus noise (e.g., due to motion) via ICA.

The tedana pipeline is reasonably well documented (see https://tedana.readthedocs.io/en/stable/) and the combination of TEs (using T_2^* estimates to combine the signal across echoes via a weighted average) is fairly non-controversial in the field. The method and threshold for distinguishing signal (TE-dependent) from noise (non-TE-dependent) ICs is less established, but as explained below, we will compare de-noising results on the first batch before deciding whether to use this separate option (yes, removing ICs does affect df's in data, but SPM estimates these during prewhitening anyway).

Moreover, in a recent study conducted in our lab, Lee, Cooper, & Henson evaluated the contribution of a multi-echo sequence on the ability to recover ATL signal. These results are not yet published, but are included here for your consideration. 4 single-echo images (TEs = 13 ms, 25.85 ms, 38.7 ms, and 51.55 ms), were compared with a combined-echo image, and a combined-echo image which was further subjected to ICA-denoising. As shown in the figure below, group level statistics (N=28) for the main effect (in that case - repetition suppression; collapsed across 6 experimental conditions) showed more suprathreshold voxels for the combined-echo image (thresholded at p < .001, uncorrected) vs. any of the single-echo images.



Furthermore, when extracting the number of ATL "brain" voxels (using SPM's definition: voxels with averaged raw signal > 80% of the averaged signal across the entire image), the signal was recovered in ~60% of the voxels when the echoes were combined (both with and without ICA-denoising; see table below), compared to ~50% voxels that would have been recovered with a TE=30ms (the traditionally used "optimal" TE). Moreover, when testing for repetition suppression effects within these "brain" ATL voxels, T-values were greater for combined vs. single echoes. However, following ICA-denoising, the combined echoes resulted in reduced T-values, possibly due to the removal of some brain signal as well as noise.

	TE = 13ms	TE = 25.85ms	TE = 38.7ms	TE = 51.55ms	Combined echo	Combined + ICA-denoise
% of voxels (out of 333)	80.6	58.3	38.6	25.8	58.2	58.2
Mean T-value	-0.01	0.05	0.02	0.03	0.28	0.19

We therefore conclude that while there are clear benefits to combining echoes, the benefits of the ICA-denoising are more debatable and should be explored further. Therefore, we will analyse the data from the first batch of participants using both methods (combined and combined + ICA-denoise), and will formally compare them as described above. The most beneficial approach will then be used in the current study. We now address this on p14 of the manuscript:

> *"The functional MEMB data will be pre-processed using a combination of tools in FSL, AFNI (v18.3.03)[34] and a python package to perform TE-dependent analysis[35–37]. Despiked (3dDespike), slice-time corrected (3dTshift, to the middle slice), and realigned (3dvolreg) images will be submitted to the "tedana" toolbox (max iterations = 100, max restarts = 10), which takes the time series from all the collected TEs, decomposes the resulting data into components that can be classified as BOLD or non-BOLD based on their TE-dependence, and then combines the echoes for each component, weighted by the estimated T2* in each voxel, and projects the noise components from the data[36,38]. We note that the benefits of the denoising procedure require further validation for task designs. Therefore, we will analyze the data from the first batch of participants (n=20) using both methods (combined and combined + ICA-denoise), and will formally compare the results using a contrast of overall task effects (for success and failure trials against rest), which is orthogonal to our key contrast of interest. The most beneficial approach will then be used in the current study."*

Related to the above point, do the authors have any strategies in testing tSNR, at least across the ROIs that they intend to compare?

When using task designs, unlike when analysing resting-state data, tSNR is not a good measure of SNR. This is because task-related brain signals reflect deviations from the mean throughout the time-series data (i.e., deviations represent actual neural activity, not noise). Contrast-to-noise (CNR) is a better measure. Even so, to account for potential CNR differences across various ROIs, we now include ROI as a factor in our statistical models: significant interaction effects that include this factor cannot be explained by differences to the level of noise across ROIs.

Will there be a special prescription of slices that align with the temporal lobes? This tends to recover the signal better from ATL but might lead to loss in other regions.

For each participant, slice alignment will be done to allow whole brain coverage, based on their position in the scanner. The recovery of ATL signal will be maximized using the ME sequence as described above.

In the preprocessing steps the authors indicate that they will use an 8 mm FHWM Gaussian kernel for smoothing. Would this not affect the subsequent representational similarity analyses?

We thank the reviewer for highlighting this issue. We will only apply smoothing for whole brain analyses (i.e., when extracting our functional ROI definition) but not when analysing ROI data (neither for the univariate, nor for the similarity analysis). This is now corrected on p14.

For the ROI definition, I am not sure about the logic behind this approach. The authors state that they would like to assess if brain-based representations of semantic and episodic memory differ from one another. But then they already pre-define different ROIs for each memory type that will be maximally separate (with some potential overlap). Would this not bias the results in finding different representations if the regions are pre-defined as different in the first place? In other words, I am not sure why the authors have not opted for using a whole-brain analysis.

Our plan is to use two definitions of ROIs – functional based and anatomical based. The contrast that will be used to define the functional-based ROIs (success effects in the critical tasks vs. the control task) is orthogonal to the critical contrast (success effect in the episodic vs. semantic task) and therefore provides an unbiased ROIs definition. For the anatomical-based ROIs, we do indeed plan to use pre-defined ROIs. Importantly, as was mentioned above, the regions were pre-defined, but based on separate bodies of literature. It could be, for example, that the ATL was not implicated in episodic memory tasks and that the hippocampus was not implicated in semantic memory tasks because the experimental procedures were not suitable to detect such patterns of activation (be that because of the paradigm that was used, the acquisition protocol, or the analytical approach). Therefore, only if semantic and episodic memories are indeed separable, we would expect them to differ within these pre-defined regions.

Overall, this is a highly interesting topic, but I would strongly urge the authors to re-think their experimental design and analysis strategies in order to be able to test their question of interest and the accompanying hypothesis.

<u>Reviewer #3:</u>

I was asked to specifically focus on the fMRI analysis and sequential approach to sample design. In general these seem sophisticated and very appropriate analyses, particularly the sequential stopping approach to sample size. However, there I have a few comments and there are a few points where the analysis plan is underspecified, as detailed below.

We thank the reviewer for their positive comments.

The Bayesian linear mixed-models proposed for behaviour and neuroimaging data are well justified. What about specific stimuli related item effects, could/should these be modelled (e.g., as random effects) as well?

From our previous experience with a similar design and a similar statistical approach (Tibon et al., 2019), models that include both subject and item as random factors do not converge. However, we will try to apply this to the current data, and will report the results if the models converge.

Regarding the sequential sample size procedure, if I have understood it correctly, this is a really nice approach. However, there are a few points I was not clear on: what is the batch size after which a decision will be taken as to whether sufficient subjects have been collected (e.g., 20 or 10 participants)?

As we now clarify on p20, the first batch will include 20 participants. Following that, if the stopping criteria were not met, the experiment will continue to run in batches of 10 participants.

What is the total N, at one point it says 60 but in the simulations it appears to be 100.

The maximal N will be 100 (60 was a typo, now corrected; apologies)

Most critically, I wasn't very sure what hypothesis the effect of the stopping criteria is on the hypotheses being tested: particularly how the multiple tests are aggregated in the simulation and how the BF takes this into account. Isn't it possible that one ROI will have BF>10 or <1/6 but all the other aren't? Equally, the similarity analysis could be BF>10 and the univariate not etc.

In order to be able to set clearer criteria, we have made some revisions to our statistical approach. First, our stopping rule will be determined based on the anatomical definition of ROIs. Each anatomical ROI includes several regions of non-contiguous voxels (i.e., a network), and the number of ROIs is predefined (2 – episodic network and semantic network; note also that overlapping voxels will be excluded). The functional definition, for which the number of ROIs is unknown and would depend on the data, will still be used to analyze the data (once data collection is finished), but not as a stopping criteria. Second, the ROI will be included as a factor in our models, and the key result would be the interaction between ROI, task, and response type. Accordingly, we revised the text on p16 to explain how the ROI definitions will be selected and used:

> *"**Selection and definition of networks/regions of interest (ROIs).** We will use two definitions of ROIs. The first will be anatomically-based, for which we will identify two networks of interest: a core-recollection network that is applicable in various episodic tasks, regardless the nature of the recollected content[11,12], and a semantic network, comprised of regions that show consistent activation in tasks that require semantic processing[39], combined with an ROI of the ATL from a recent fMRI study using a paradigm with enhanced ATL signal coverage[40]. Note that each anatomical ROI includes several regions of non-contiguous voxels (i.e., a network). In order to maximize our ability to detect potential differences, we will exclude any overlapping voxels that are shared between the two networks. This ROI definition will be used for setting the stopping criteria for data collection. The second definition will be functionally-based. For this definition, we will run a second-level whole-brain analysis, and will identify regions where the recall success effect (success-failure) in the two critical tasks (episodic and semantic) is greater than the success effect in the control task ($p < 0.05$ FWE cluster-level corrected, with voxel-level threshold at $p < 0.001$ uncorrected). This definition (together with the anatomically-based definition) will be used to analyze the data once data collection is finished."*

When describing our statistical approach, we define the statistical models as:

> *Univariate: betas ~ ROI * task * response_type + (1 + ROI + task + response_type | subject)*
>
> *Similarity: coefficient ~ ROI * task * trial_type + (1 + ROI + task + trial_type | subject)*

And further explain how the "ROI" factor will be handled in these models (p18):

> *The "ROI" factor will always include two ROIs. For the anatomical definition, these will be the two predefined ROIs. For the functional definition, we will run the model multiple times, using different pairs of ROIs each time until all pairs had been contrasted. To account for multiple comparisons, we would adjust the threshold Bayes factor so that the false positive rate (FPR; i.e., the chance of falsely accepting the alternative, assuming that the null is true) remains below 5%.*

Finally, the revised stopping criteria (p21) are set to be:

> *"(1) BF10 > 10 for either of the two main analyses (univariate or similarity) for the anatomical ROI definition; (2) BF01 > 6 for either of the main analyses; or (3) n=100"*

RESPONSE TO REVIEWERS

Reviewer 1

This is a timely and interesting manuscript on the overlap in the neural correlates of semantic and episodic memory. The experimental design is clever and original and the manuscript well-written and well-presented. The deviations from the registered protocol were minimal and well-justified. The pilot study to determine whether logo names were known (nameable) was somehow underpowered (N=8), but part of the original submission. I have a number of clarification questions for the authors.

We are thankful to the reviewer for their positive remarks and have addressed their questions below.

Could the use of multi-echo fMRI have impaired hippocampal activation detection and therefore some specificity in episodic memory activations?

We thank the reviewer for bringing up this issue. We have added this to the discussion:

> *"An additional limitation concerns the fMRI protocol selected for the study. Namely, while the multi-echo protocol used here crucially allowed us to recover ATL-signal, it might have reduced our ability to detect hippocampal activation. A recent study compared the ability of different fMRI protocols to detect activation during a semantic task[42]. In their study, the number of echoes (single-echo [TE = 30ms], vs. multi-echo [TEs = 12ms, 25.85ms, 38.70ms]) and bands (single-band, multiband [2]) were manipulated independently to construct a 2 x 2 factorial design in which the effects of multi-echoes and multi-bands can be estimated. A whole-brain analysis revealed greater magnitude and precision of activation in the hippocampus when single- vs. multi-echo protocols were used (see Figures 5b and 6b, and Supplementary Materials 4). Nevertheless, a study that examined hippocampal activation during an episodic memory task, showed clear hippocampal involvement in the task when multi-echo data were used (TEs = 12.5ms, 27.6ms, and 42.7ms) together with ME-ICA, but not when single-echo data (TE = 26.6ms, albeit acquired within the same multi-echo protocol) were used[73]. This suggests that while the possibility of some reduced ability to detect hippocampal activation remains, the application of ME-ICA in our study was potentially helpful in balancing trade-offs between the ability to detect ATL vs. hippocampal signals."*

Could the authors comment on excluded ROI? (supplementary table 1) I was not sure about the reasons for this. The main manuscript mentions excluding overlapping voxels between the core recollection network and the semantic network, but I would not have thought that temporal pole would have been part of this for example?

This refers to the exclusion mentioned in p7 which states:

> *"This resulted in 16 clusters that contained at least 60 voxels (15 additional clusters of fewer than 60 voxels were ignored)".*

To clarify this, we have changed the heading of this section in Supplementary Table 1 which now reads: "Excluded ROIs (due to N<60 voxels)"

Comments on ATL findings would be useful (was it analysed or excluded) I thought the motivation of using multi echo fMRI was in part to capture ATL (the authors also rightly commented in the introduction on the poor coverage of typical episodic memory studies in this region)

The ATL was analysed as part of the a-priori defined semantic network and a data-defined cluster, and was also included as part of the 16 ROIs comprising the exploratory analyses that used a revised threshold for data defined locations. However, following the reviewer's comment, we have decided to explore this region further. We now report these exploratory analyses on p23:

> *"**Analysis of ATL involvement in semantic vs. episodic retrieval.** As explained above, one main motivation for using multi-echo fMRI sequence was to better capture mnemonic effects in the ATL. However, in our preregistered analyses, the ATL was included as part of a broader semantic network, and as part of a large data-defined cluster, but not on its own. Therefore, to further examine mnemonic effects within this region, we extracted time-series data from the ATL location using a mask that originated from the a-priori defined semantic network of interest, but only included the ATL. We then applied the same methods used in all other analyses, and fitted the following model: betas ~ task * response type + (1 + task + response type | subject).*
>
> *We reasoned that a conclusive interaction between response type and task, such that a greater effect of response type is observed in the semantic vs. episodic task, would suggest that the ATL is specifically involved in semantic retrieval. Nevertheless, this analysis showed that the null hypothesis of no difference was conclusively preferred (BF01 = 20.51), indicating that the ATL is similarly involved in both tasks."*

And on p24:

> *"**Analysis of ATL involvement in semantic vs. episodic representation.** To explore whether the ATL is specifically involved in semantic representation, we used the same approach described above for the corresponding univariate analysis but using this model instead: coefficient ~ task * trial type + (1 + task + trial type | subject). Like the univariate analysis, the similarity analysis showed conclusive support for the null hypothesis (BF01 = 13.65), providing evidence against specific involvement of the ATL in semantic vs. episodic representation."*

These results are also mentioned in discussion on p26:

> *"… whereas an exploratory analysis performed specifically within the ATL region showed evidence against specific involvement of this region in semantic vs. episodic memory retrieval."*

Some clarifications about the semantic measures would be useful. Was success retrieval for semantic memory simply logo name retrieval or actual semantic knowledge about the logo? Did the authors check the details provided by participants were actually accurate? ("detailedness" analysis) Did the authors compare semantic

and episodic memory retrieval activations when matched in number of details retrieved?

We now clarify on that in the semantic task successful retrieval was logo name retrieval.

In both tasks, the actual details remembered by participants were not recorded due to technical constraints (i.e., extensive verbal responses in the scanner are problematic due to head movements, and in addition, the time that the task took was already very long). Therefore, we are unable to check whether the details provided by the participants were accurate.

As for activation in episodic vs. semantic retrieval when remembered details are matched, please refer to the Exploratory behavioral analyses section: Exploring differences in "detailedness", in which we have included an analysis showing that the tasks did not differ in "detailedness" (i.e., were matched on that aspect).

Any reasons why right hippocampus (observed with less severe correction) would be more robust than left?

As noted by the reviewer, there is potentially some (exploratory and inconclusive) evidence for episodic-semantic dissociation in the right hippocampus when contrasted with the right cerebellum and the left MTG, that are not present when the left hippocampus is contrasted against these regions. Importantly, however, there was no evidence of differential effects in the right vs. left hippocampus when these were contrasted directly. Therefore, and given the inconclusive nature of this finding, we inclined not to report and discuss it in the manuscript.

Since stage 1, new literature relevant to continuum models of declarative memory has been published:

Addis & Szpunar, 2024, Philosophical Transactions, B

Curot et al., 2024, Cortex.

Tanguay et al., 2023, eLife

We are grateful to the reviewer for bringing these interesting and highly relevant papers to our attention. As the introduction cannot be revised during Stage 2, we have included these in our discussion (p30):

> *"As noted, however, a final possibility to consider is that the distinction between semantic and episodic memories is in fact outdated, or more nuanced than previously thought. As mentioned in the introduction, previous neuroimaging studies have already highlighted more similarities than differences when episodic and semantic memories were compared within the same design[16–22]. Recent theoretical perspectives and a growing body of empirical evidence further suggest alternative approaches to the episodic-semantic distinction. These approaches can be broadly referred to as continuum approaches, that view episodic and semantic memories as opposite ends of a continuum, and multidimensional approaches, which position memories in a multidimensional space, allowing them to vary across a range of categorical or continuous dimensions[10,72,73]. A recent example for the latter is the Multidimensional Model of Mental Representations (MMMR) which suggests that mental representations vary along continuous dimensions in how much they are*

temporally specific or general, perceptual or conceptual, and idiosyncratic or shared[73]. As such, this model arguably teases apart dimensions that are often confounded when the episodic–semantic distinction is considered. Two recent studies support such a refined view of episodic and semantic memories. In an fMRI study, brain activity was recorded while participants verified statements concerning general (semantic) facts, autobiographical facts, repeated events, and unique (episodic) events. Semantic and episodic memories involved shared neural correlates, but the degree of engagement of each region within this network varied across different memory types[74]. In another study, Electrical Brain Stimulation (EBS) was used to trigger a transient brain state in epileptic patients while neural activity was recorded by intracranial EEG (iEEG). Such triggers often evoke sudden involuntary reminiscences that can vary in their degree of contextual dependency. The study provided some suggestive evidence of greater functional connectivity for more complex memories, with a decreasing gradient of complexity from episodic memories, to personal semantics, and to familiar memories[75]. Like our current findings, these findings resonate better with theories that assume a continuum between different memory systems, or with a multidimensional perspective of declarative memory where different memory types vary in magnitude of activation within a common network of brain regions, than with the traditional view that separates the systems apart."

Reviewer 2

In this study, Tibon and colleagues investigated potential differences in the activations and neural representations of episodic and semantic memory. Having previously reviewed the Stage 1 submission of this registered report, I am delighted to see the research results in this Stage 2 submission. Overall, this is a carefully thought-out experiment on a topic that will undoubtedly interest the scientific community and beyond.

The introduction section, including the study rationale and stated hypotheses, remains the same as the approved Stage 1 submission. The authors clearly state a few minor deviations from the registered protocol, which I believe would not substantially influence the reported results.

The authors adhered to the registered experimental procedures and clearly outlined their exploratory analyses in the Results section. These unregistered analyses are justified, methodologically sound, and informative, especially given the null finding.

The behavioral analyses demonstrate that the stimuli selection and paradigm design procedures worked successfully, with a relatively equal number of successful trials across conditions. This yielded sufficient trial-wise exemplars for the authors to test their hypothesis. Therefore, I believe their conclusions are justified by the data.

We are grateful to the reviewer for their efforts in reviewing and their helpful suggestions on both stages of our manuscript.

Given the authors' findings I would have expected a discussion of the null results in relation to prior studies that investigated this research question, albeit using other experimental paradigms (e.g., Burianova et al, Vatansever et al, Humphreys et al).

Instead, these works appear only briefly in the introduction section, which understates the research conducted before this registered report. While "absence of evidence does not constitute evidence of absence," theoretical suggestions for a unitary retrieval system warrant discussion in the context of these findings. The authors devoted considerable attention to a small number of observed dissociations based on exploratory rather than pre-registered analyses. I urge the authors to revise their discussion to emphasize the null results from pre-registered analyses rather than findings from exploratory analyses.

We thank the reviewer for encouraging us to further discuss our null results, as we have done now in our discussion (added text in bold):

> *"**Our above attempts to explain the current null results all suggest that the distinction between semantic and episodic retrieval does exist, but might occur, or might be detected, under specific conditions. Nevertheless, an important possibility to consider is that episodic and semantic memories indeed involve the same neural mechanisms. This view might apply specifically to the retrieval phase, or it might broadly encompass declarative memory. If the former, then** the dissociations observed in neuropsychological studies might be attributed to distinct encoding or retention of episodic and semantic memories. For example, it could be that AD patients are less able to encode novel associations compared to healthy controls. This reduced ability to encode information will then also lead to reduced ability to retrieve information, but not because of difficulties in retrieval per se, but because less to-be-retrieved information is available. Notably, however, in neuropsychological studies, it is often impossible to distinguish between different mnemonic stages, as data preceding the lesion/disease is usually unavailable for patients. Neuroimaging studies can more easily address this question of different neural mechanisms during different mnemonic stages, although unfortunately, this distinction cannot be established in the current study. Namely, we only collected neuroimaging data during retrieval, and indeed, obtaining encoding data would have been impossible with the current design in which semantic knowledge is assumed to be attained over prolonged time periods. Given the current results, and the observed discrepancy between patients and neuroimaging studies, the field would benefit from future studies that are designed to explore this proposal, i.e., that whereas encoding varies for episodic and semantic memories, their retrieval involves overlapping neural mechanisms.*
>
> ***As noted, however, a final possibility to consider is that the distinction between semantic and episodic memories is in fact outdated, or more nuanced than previously thought. As mentioned in the introduction, previous neuroimaging studies have already highlighted more similarities than differences when episodic and semantic memories were compared within the same design[16–22]. Recent theoretical perspectives and a growing body of empirical evidence further suggest alternative approaches to the episodic-semantic distinction. These approaches can be broadly referred to as continuum approaches, that view episodic and semantic memories as opposite ends of a continuum, and multidimensional approaches, which position memories in a multidimensional space, allowing them to vary across a range of categorical or continuous dimensions[10,72,73]. A recent example for the latter is the Multidimensional Model of Mental Representations (MMMR) which suggests that mental representations can vary along continuous dimensions in how much they are temporally specific or general, perceptual or conceptual,***

and idiosyncratic or shared[73]. As such, this model arguably teases apart dimensions that are often confounded when the episodic–semantic distinction is considered. Two recent studies support such a refined view of episodic and semantic memories. In an fMRI study, brain activity was recorded while participants verified statements concerning general (semantic) facts, autobiographical facts, repeated events, and unique (episodic) events. Semantic and episodic memories involved shared neural correlates, but the degree of engagement of each region within this network varied across different memory types[74]. In another study, Electrical Brain Stimulation (EBS) was used to trigger a transient brain state in epileptic patients while neural activity was recorded by intracranial EEG (iEEG). Such triggers often evoke sudden involuntary reminiscences that can vary in their degree of contextual dependency. The study provided some suggestive evidence of greater functional connectivity for more complex memories, with a decreasing gradient of complexity from episodic memories, to personal semantics, and to familiar memories[75]. Like our current findings, these findings resonate better with theories that assume a continuum between different memory systems, or with a multidimensional perspective of declarative memory where different memory types vary in magnitude of activation within a common network of brain regions, than with the traditional view that separates the systems apart."

Regarding style, the manuscript contains numerous parenthetical asides within long sentences. To improve readability, I suggest reducing this practice.

We have greatly reduced the usage of parenthesis in the text.

While I appreciate the detailed information provided, certain sections, such as the lengthy description of sample size determination procedures, could move to the supplementary notes, as they are not central to the experimental work and conclusions.

We removed this section from the main text, and included instead in the supplementary materials.

Finally, I recommend improving text visibility in certain figures (e.g., Figure 1).

Revised as requested