Dear Editor,

We would like to thank the editor and the reviewer for their constructive comments. We address all the points below and we report two additional control analyses. We have updated the manuscript and expanded the discussion to acknowledge the points made by the editor and the reviewer. We feel that the manuscript is improved as a result of these changes and hope that it is now considered acceptable for publication.

REVIEWER COMMENTS (italics) and our responses (normal text)

Dear Dr. Mojescik,

Thank you for your submission to Collabra. After reading the original reviews and decision letter, I was overall positive about the manuscript but did feel the need to consult one additional expert reviewer. As you will read, this reviewer is also mainly positive, but also raises a few issues. The two comments that stood to me (and that I hadn't picked up on myself) are about the potential difference in reliability between your gist and detail measures, which might translate to differences in the correlation with the subjective ratings, and the fixed order of questioning. Please address these points in the revision. In addition, please carefully consider the other points raised by the reviewer, although I see them as less crucial and thus leave it up to you whether or not you want to act on them.

In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal. Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all necessary copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,
Sebastiaan Mathôt

We thank the editor for the positive evaluation of our work, and we address the reviewer's points highlighted by the editor in the responses directly to the reviewer.

REVIEWER 1:

SUMMARY

This manuscript reports the results of a study on the relationships between objective and subjective measures of episodic memory, and tested whether these relationships differ between older (68-75 years old) and younger adults (20-25 years old). Episodic memories were induced using an event-imagination paradigm in which participants imagined events comprising a theme, a famous person, and object, and a location, the latter three being presented as images and words. The objective measures of episodic memory comprised performance on a 4-alterative test of gist memory (i.e., select the appropriate person / object / location for an event, based on its theme), and a 2-alternative test of memory for perceptual details (i.e., for faces: select which of two pictures of the same person was shown in the imagination task; for objects and locations: select which of two pictures of objects or locations belonging to the same basic level category was associated with the event).

The main finding of interest was that for both age groups, memory vividness and memory confidence were associated with accurate recognition of gist, but not with memory for perceptual details, and that this pattern of associations was similar for the two age groups.

EVALUATION

This study presents a highly powered replication and extension of an earlier study by Cooper and Ritchley (2022), using an exact replication of their methods and analyses. My evaluation of the study is that it makes a nice empirical contribution to the field by replicating earlier work and adding the comparison of the relationship between objective and subjective memory measures for younger and older adults. Moreover, the authors did an exemplary job in their adherence to the recommendations for open science. That said, I do have some points of critique and suggestions to improve the paper before it can be published.

We thank the reviewer for their positive evaluation of our work.

My main concern about the study is whether the comparison of the correlations with the two objective memory tests is fair, given that there are several differences between these two memory tests that might have influenced the correlations.

*To start, the two measures probably differed in reliability as a 4-AFC test would be expected to generate a more reliable measure of memory contents than a 2-AFC (e.g, Rodriguez, 2005). Importantly, a lower reliability would entail that the correlation with another measure is also bound to be lower (e.g., Vul et al., 2009), thus offering a potential explanation for why the correlation with perceptual details was non-significant and lower than that observed for gist. The authors could address this issue by computing the reliability (e.g., Cronbach's alpha) of each of the two memory tests, and then using these reliability estimates to normalize the correlation: rc = ru / sqrt (r11 * r22), where rc is the corrected correlation, ru is the observed (uncorrected) correlation, and r11 and r22 are the reliabilities of the two measures (formula copied from*

http://www.med.uottawa.ca/courses/CMED6203/Index_notes/Reliability & Correlations.htm#:~:text=rc %3D ru %2F sqrt,0.8 gives rc %3D 0.44.).

We understand and appreciate the reviewer's concern and we address it below. Before we do so, it is worth noting that the primary aim of our study was to compare the performance of a younger and older population of healthy adults, using a newly developed paradigm. Our key finding is that the relationships between objective and subjective measures of memory were very similar in both age groups, despite differences in overall level of performance. This result holds, even if there are concerns about the differential reliabilities of the tests for gist- and detail-level memory.

To address the reviewer's concern, we conducted additional control analyses. Nevertheless, we note that our main analyses (see Figure 2) were performed at a trial-level, separately for older and young adults, rather than looking at the overall correlations based on the results of all participants. We are not aware of any methods for correcting these analyses for potential differences in the reliabilities of the measures used. In the individual differences section of the study, we did compute the correlations at a mean performance and mean rating level; however, they were still split across the three types of content elements (place/person/object) and across the age groups. For these reasons, it was not possible to use the formula suggested by the reviewer to correct for potential differences in the reliability of the gist- and detail-level performance measures directly on the correlations reported in the manuscript. However, we have conducted additional analyses that investigate the overall correlations between the two performance measures and vividness ratings and here we were able to use the formula to calculate corrected correlations. We then compared these two corrected correlations and found that gist was still significantly more strongly correlated with vividness than detail. Below we provide more detail on how we addressed the point.

To address the reviewer's concern, we conducted a new analysis based on overall performance on the gist- and detail-level scores (collapsed across the three elements of the events – place, person and object). On both measures, the participants could score minimum of 0 – none out of the three gist or detail questions answered correctly, and a maximum of 3 – all gist or detail questions answered correctly. We then obtained a mean average of these scores across the 24 trials for each participant. These gist and detail memory scores were correlated with the participant's mean average vividness score. These provided us with the uncorrected correlations between the objective memory measures and the subjective memory measures. We found a moderate correlation between gist memory and vividness ratings, $r(191) = 0.52$, $p < 0.001$, and low correlation between detail memory and vividness ratings, $r(191) = 0.35$, $p < 0.001$. There was also a low correlation between the gist memory scores and detail memory scores, $r(191) = 0.44$, $p < 0.001$.

Next, we performed a split-half reliability by splitting the data into odd and even trials to obtain a pair of performance scores per participant. These paired scores were correlated

across the whole sample to calculate split-half reliability scores for the gist, detail and vividness measures. The split-half reliabilities were: gist memory ($r(191) = 0.80$, $p < 0.001$), detail memory ($r(191) = 0.62$, $p < 0.001$), and vividness rating ($r(191) = 0.87$, $p < 0.001$). Interestingly, this did reveal that the 2AFC detail measure was indeed less reliable than the 4AFC gist measure. However, it is important to note that performance was also closer to chance for the detail memory test, so a greater proportion of participants' responses will have been guesses, which will also lead to more variability.

We then used the formula *rc = ru / sqrt (r11 * r22),* to calculate the corrected correlations, separately for gist memory and detail memory. We divided the uncorrected correlations with mean vividness ratings (gist memory: 0.52; detail memory: 0.35) by the square root of the objective memory measure reliability (gist memory: 0.80; detail memory: 0.62), multiplied by the reliability of the vividness measure (0.87). We compared the obtained corrected correlations of 0.62 for gist memory and vividness ratings and 0.48 for detail memory and vividness ratings using the Steiger's Z test for dependent correlations. We found that there was still a remaining significant difference between the two correlations: $z = 2.34$, $p = 0.01$.

We first conclude that the gist- and detail-level measures do indeed differ in their reliability. Moreover, correcting for the reliability of the measures reduces the difference in strength between how both measures correlate with vividness scores (uncorrected difference in correlations = 0.17, difference after correction = 0.14). However, the difference remains significant and therefore the main findings of the study are unchanged.

To address the reviewer's concern within our revised manuscript, we have included this additional analysis within the supplementary material. However, we have not mentioned it within the manuscript itself. This is due to a number of reasons; (1) we pre-registered our intended analyses which aimed to replicate the procedure and analyses used by Cooper and Ritchey (2022), (2) the novel aspect of this study is the comparison between a younger, and an older cohort, and any significant age-related differences should be observable, irrespective of the specific analyses used, (3) the results of our study are unchanged by performing the correction suggested by the reviewer. We felt that the inclusion of this analysis was likely to confuse, rather than enlighten the reader. However, we are happy to include details of this analysis within the main Results section if the editor feels that it is important to demonstrate the robustness of our findings.

Another potentially problematic aspect of the chosen methodology is that memory for gist was always probed before memory for perceptual details. Arguably, this could lead to relatively inflated performance for the test of perceptual details, as previous work has shown that repeatedly probing the same engram leads to an increase in performance (i.e., if you test the different pairwise associations for an engram comprised of three elements, performance

increases across the sequence of tests; e.g., Horner & Burgess, 2013). This potential benefit of retrieval practice could impact the observed correlations as inflated performance means there will be less variance.

The reason why the order of the memory tests was fixed is because we wanted to obtain both gist and detail memory measurements for every trial, permitting within-subjects analyses of how they are mutually related to vividness. We note that the perceptual detail questions are always tested on the correct conceptual-level item (for example, if "Tom Cruise" was the correct concept-level person element within an event, then the perceptual detail question will always show two similar photos of Tom Cruise, irrespective of whether the participant correctly identified Tom Cruise as being the person present in the event). Consequently, it would not be possible to test gist memory after revealing the correct answer in the perceptual detail memory test. Thus, the gist memory questions had to be presented before detail memory questions. We discuss the limitations concerning the order of the memory tests in this paradigm in the last two paragraphs of discussion before conclusion. For the revised manuscript, one of these paragraphs was added to explain why the gist questions necessarily had to be presented before perceptual detail questions in this paradigm and suggesting possible solutions for the future research.

Relatedly, a more minor suggestion would be for the authors to explain and motivate their choice of how the objective memory measures were computed. Based on the means reported for the analyses of memory performance (p. 12-13), it seems the authors computed the average total number of elements recognized correctly per engram. What should be clarified here or in the methods section is whether these scores were corrected for the different guess rates in the 2- and 4-alternative choice tests that were used for to test memory for perceptual details and gist.

It is correct, the objective memory measures were computed using the average total number of elements recognized correctly per engram. Regarding correcting for difference guess rates, we did not do this. This was because we never compare performance on the gist- and detail-level questions, so different guess rates are not an issue. The only comparisons made were the difference between older and young adults on gist memory questions (thus having the same guess rate as they were answering the same questions), and the difference between older and young adults on detail memory questions (again, having the same guess rate as they were answering the same questions).

Another aspect that should be clarified about the objective memory tests is the order in which the tests of the three types of elements (faces / objects / locations) were presented: Was the order of these tests randomized, counterbalanced, or fixed? If it was fixed, this could again introduce a retrieval practice effect that would lead to differences in performance for the three elements, potentially contributing to a difference in correlations.

We thank the reviewer for spotting the omission in the methods section. The order of questions for person/object/place was counterbalanced across trials but was fixed for gist and detail questions within the same trial. This information has now been added to the method section. Thus, the results should not be affected by a retrieval practice effect.

We also note that the choice of example trials accessed via the link provided in the Method section could have been misleading, as the three selected trials happened to test the elements in the same order. The choice of example trials has been now updated to reflect that the order in which the elements were tested was not kept constant across all trials.

Lastly, the authors could discuss whether the memory tests for perceptual detail were equivalent for the three elements. Here, it seems that the test for faces would require more detailed information than the tests for objects and locations, as the latter involved a lure that was a different exemplar from the same basic level category while the former involved a lure of the same exemplar (i.e., another picture of the same person).

We appreciate this point from the reviewer. Nevertheless, the difficulty of the test depended on the perceptual similarity between the target and lure items. We aimed to choose items that would balance the difficulty of the test across all element types.

To address the specific concern of the reviewer, we conducted an additional analysis to compare the memory performance for the three elements on the perceptual detail memory test. We found that there were no significant differences in performance on the three elements amongst young adults, $F(2, 97) = 1.44$, $p = 0.24$. However, amongst older adults, we did find a significant main effect of content on perceptual detail memory performance, $F(2, 97) = 4.45$, $p = 0.01$. Post hoc tests using Bonferroni correction for multiple comparisons have shown that older adults performed significantly poorer on the place perceptual detail memory ($M = 0.67$, $SD = 0.14$) than the object perceptual detail memory ($M = 0.71$, $SD = 0.14$). There were no significant differences between the person element ($M = 0.69$, $SD = 0.13$) and neither the object nor the place elements. Thus, there was no indication of young or older adults finding the person element more challenging than the place or object elements on the perceptual detail memory test.

Taken together, I think that the methods used to measure of memory for gist and perceptual details have some suboptimalities that could be detrimental to the interpretation of any differences in correlation with these two measures. For future studies, a better approach might be to assess memory for gist and detail independently across different engrams, using a choice task with the same number of choice options for each type of memory, much like

Brady et al. (2008) assessed and compared memory for pictures of objects at the level of gist, exemplar, and state.

We thank the reviewer for very useful comments on the manuscript – and in particular, their suggestions for improvements to the paradigm for future studies.