The role of positive and negative emotions on multiple components of episodic memory (“what”, “when”, “in which context”) in older compared to younger adults: a pre-registered study

Pierrick Laulan, Ulrike Rimmele

Romain Yang, who appeared as second author in the first version of the manuscript, has been removed from the list of authors in this new version, as he will be working on another research project within our team.

Response to reviews

Dear Dr. Wonnacott, dear Reviewers,

Thank you very much for your insightful and constructive comments on the first version of our manuscript. We have carefully addressed all comments and provided as complete a response as possible in order to substantially improve the manuscript we are resubmitting to you for evaluation.

You'll find a point-by-point response to your comments and questions: reviewers' comments are in bold, our responses are in italics and changes made in the body of the manuscript are highlighted in yellow (all relevant changes have also been made in the design table). Changes are also indicated in yellow in the manuscript itself. We hope you will consider the manuscript suitable for acceptance after these changes have been made. We hope you will consider the manuscript suitable for acceptance after these changes have been made.

Revision invited - Liz Wonnacott

Thank you for submitting your work as a registered report to PCI RR. Your paper has received two reviews from experts in the area. Both are generally positive and consider the study overall well designed. However, they do each raise some (different) points about aspects of the design which you should consider. The first reviewer also has some important comments on the literature review, with places where you could be a little more nuanced in your discussion. Consideration of these points may also impact on design choices.

I have some further comments/questions concerning analyses plan. First, I would like to note that the models are generally well described with a good level of detail, and the outcomes of analyses are specifically linked to the hypotheses, which is great. However, I would like to ask for a little more detail in places and to further probe the rational for the decisions you made, to make sure you have the best plan in place. (NB - I would like to acknowledge that I haven't used GEE's myself -I mostly use mixed effect models- however from a quick look at some online resources it seems like the fixed effect factors are set up in the same way as in lmer/glmer. I have therefore made suggestions in the same way as I would for those models, but do feel free to educate me if any of my points are misplaced in this context).

Response: We thank the editor for inviting us to revise our manuscript. We thank the editor and the reviewers for their valuable and constructive feedback. We have carefully considered all of their comments and have made significant revisions to the manuscript accordingly. The changes are highlighted in yellow throughout the revised manuscript and our response letter. We have also provided point-to-point responses addressing each of the reviewers' comments.

We firmly believe that the review process has greatly enhanced the quality of our manuscript. We have taken all the necessary steps to address the reviewers' concerns and improve the overall clarity and scientific rigor of the work. We look forward to hearing back from you regarding our revised manuscript.

1) Where you say you are going to have age and valance as factors in each model I think you should add "and their interaction". In one of the models you also have target-type- how will this be coded? Will you include any interactions with this factor? If so, if any are significant how will you deal with them- will they be interpreted but treated as exploratory?

Response: We thank the editor on the comments/questions concerning to depict our analysis plan in more detail. In line with your propositions and for greater precision, we have added that the interaction between valence and age will be introduced whenever there is an age factor in a model. We had now added the following sentence in yellow to the manuscript:

See p. 26: "The data acquired in the item memory, temporal judgment and associative recognition tasks will be analyzed three times consecutively to provide precise answers to our six hypotheses. Thus, in the first stage, and for each dependent variable, we will introduce

valence (positive vs. negative vs. neutral images) and age (younger vs. older adults) and the interaction between valence and age as factors; in the second stage, we will replicate the analyses, focusing only on younger adults; in the third stage, we will replicate the analyses, focusing only on older adults."

Response: In addition, we have indicated in the manuscript that deviation coding will be applied to the "target type" factor. This factor will be introduced in interaction with emotional valence and age in the analysis of data from the item recognition task. All significant effects will be interpreted, although we have no hypothesis as to whether they emerge at the level of response bias and/or discriminability.

See p. 26: "Deviation coding will be applied to age and type of target (age, young adults = -1, older adults = 1; type of target: previously seen images = 1, distractors = -1)."

2) For your inferences, will you be interpreting the specific coefficients that are output from the model which are relevant to each hypothesis? (I.e. using the "Robust z" output for that coefficient to get the p-value, rather than getting the p-value from model comparison of a model with and without the effect in question)?

Response: In order to maintain the statistical power of our tests, and to directly address our hypotheses (i.e., whether a factor or combination of factors significantly predicts memory performance), we will interpret the coefficients from our analyses that are relevant to our set of hypotheses (i.e., valence effect in younger and older adults or valence x age effect).

3) You plan to use a quadratic contrast to see if neutral is worse than both positive and negative and a linear contrast to see if positive is worse than neutral which is worse than negative. I can see the logic here, but my understanding was that polynomial contrasts could only be applied to ordinal data with equally spaced levels (e.g. see https://stats.oarc.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/#ORTHOGONAL). Can you provide some further justification of this approach?

Response: Although the two a priori contrasts selected to test our hypotheses (i.e., a quadratic contrast and a linear contrast) might suggest polynomial contrasts, our approach is in fact slightly different, since it is based on specific contrasts. Indeed, our aim is not to determine whether there is a general trend in the data (e.g., a U-shaped relationship between

valence and performance in the item memory task), but rather to test specific hypotheses concerning the relationship between the three modalities of the emotional valence variable in our study (i.e., positive, negative and neutral). Thus, specific weights were applied to the modalities of the emotional valence variable in each of the two contrasts we selected: (1) quadratic contrast: positive images = 1; negative images = 1; neutral images = -2 ; (2) linear contrast: positive images = 1; negative images = -1; neutral images = 0. In this situation, it is not necessary for the different levels of the variable of interest to be equally spaced (see, for example, Brauer & McClelland, 2005).

Brauer, M., & McClelland, G. (2005). L'utilisation des contrastes dans l'analyse des données : Comment tester les hypothèses spécifiques dans la recherche en psychologie ? [The use of contrasts in data analysis: How to test specific hypotheses in psychological research]. L'Année Psychologique, 105(2),273–305.

4) Regardless of coding implementation, I also want to check that the two tests you have planned here (to be repeated three times each for each DV) are well designed in terms of the patterns that might turn up in your data. Take the first hypothesis: as stated, you are going to test whether the average of positive and negative leads to higher DV than neutral, averaged across both age-groups. The benefit of looking across age groups is that you have more power, but it does mean that you can't necessarily conclude that the effect holds in each age group separately (and NB even if you had an NS interaction with age you couldn't do that, as you can't establish a null interaction from a p value). It could be that actually your effect only really holds in one group, but you still get a significant effect main effect across the groups. It also might be that you miss seeing the effect because the main effect is NS, but if you were to have looked at one of the groups in isolation, you would have found evidence that the effect was present in that group. An alternative would be to look at the effect separately in each age group (e.g. by running the model twice with age dummy coded but with a different reference level each time and inspecting the coefficient for the contrast in each case). That would seem to me to be more in line with your hypotheses as stated (i.e., that the effect is present "in both groups"). Of course, you would have to ensure power for this.

Response: We totally agree with the reviewer's point. Indeed, as it stands, the proposed statistical tests do not allow us to confirm or refute hypotheses 1a, 2a and 3a, as we are

unable to indicate whether or not there is an emotional bias in both younger and older adults. We therefore plan to add separate tests for younger and older adults. More concretely, for each group of hypotheses (i.e., effect of age and emotion on 1) item memory, 2) temporal memory, 3) associative recognition), we will test emotional effects (i.e., quadratic contrast and linear contrast applied to valence) in young vs. older adults, then in young adults, and finally in older adults. With this approach, we'll be able to confirm or refute each of our hypotheses, and we'll also obtain precisions about our effects of interest (e.g., in addition to knowing whether there is an emotional bias in younger and older adults, we'll have information on a possible difference concerning this bias as a function of age). We now clarified this in the text on p. 26.

See p. 26: "The data acquired in the item memory, temporal judgment and associative recognition tasks will be analyzed three times consecutively to provide precise answers to our six hypotheses. Thus, in the first stage, and for each dependent variable, we will introduce valence (positive vs. negative vs. neutral images) and age (younger vs. older adults) and the interaction between valence and age as factors; in the second stage, we will replicate the analyses, focusing only on younger adults; in the third stage, we will replicate the analyses, focusing only on older adults."

5) Power analyses: I appreciate why you base your power analyses on RM-ANOVA given the lack of equivalent software for GEE. However, I am not clear that you are computing power specifically for the equivalent tests. You have two hypotheses, one relating to a main effect and one an interaction. The required sample for 90% power for each of these will be different, however, you report one sample size requirement. I am not fully familiar with the software you are using, but what does it mean to say you have power "for the ANOVA"? Is this specifically the power for the interaction? (I presume you would have chosen that rather than the main effect as it will need more power?). But then would this be power for the omnibus F test for age by valence, as is typically reported for ANOVA? But you are actually interested in the power for the interaction of age by the linear contrast. Do we know that the omnibus F is a reasonable proxy here? Please consider these points and be specific about what you are doing and the justification.

Response: We thank the reviewer for raising these important methodological points concerning power analyses. We have indeed sought to determine the number of participants

needed to show a valence x age interaction effect in each of the three experimental tasks of our paradigm with sufficient power. We initially set the target power for each interaction effect at 0.90, but then realized that since our study includes six independent hypotheses, with this approach we would in fact only have a power of 0.53 to test simultaneously all 6 hypotheses (i.e., 0.90^6). Consequently, we ran our power analysis again, this time with a target power of 0.964 (i.e., the sixth root of 0.80), and the results showed that a sample size of 212 participants is required in our study (vs. 160 in the old analysis).

Given the greater statistical power of a single-degree-of-freedom test vs. an omnibus test (e.g., Schad et al, 2020), our power analysis is deliberately conservative (we are aware of techniques for more accurately estimating the number of participants needed in the context of analyses based on a priori contrasts, but these seem more suitable when there are more precise predictions regarding expected effects; see Perugini et al, 2018).

The following modifications have been made in the manuscript to make the power analyses clearer.

See p. 14-15: "A power analysis performed with MorePower (version 6.0.4; Campbell & Thompson, 2012) indicated that for a repeated-measures ANOVA (RM-ANOVA), a total of 212 participants is required to show a within-between interaction effect (i.e., 3 [Valence: positive vs. neutral vs. negative] x2 [Age: young vs. older adults]; see hypotheses 1b, 2b and 3b) of size $d = 0.40$ with an alpha of .05 and a power of .964."

See p. 15 (footnote): "Given that we aim to test 6 independent hypotheses divided into 3 blocks (respectively hypotheses 1a and 1b, 2a and 2b, 3a and 3b) with a power of 0.80, we have chosen a targeted power of 0.964 for the interaction hypotheses, i.e. the sixth root of 0.80."

6) As one of the reviewers points out, when planning for sequential analyses you need to think about the fact that you have six hypotheses you are testing (I am assuming you won't be looking at the two additional follow up analyses at the interim points) rather than one. From what I understand, you are only going to stop if all of your 6 tests are significant at the required alpha level. I further presume that if a particular test A is significant at the first "look", but you have to keep going to all three time points (i.e. because other tests were not significant), you would nevertheless still eventually report A as in the final sample at time 3, and using the time 3 alpha boundary? This all seems

sensible to me, but it would be good to see if there is precedent/discussion in the literature.

Response: The response below was also submitted to another reviewer (Reviewer 2, Comment 11) who asked for clarification of our approach to sequential analyses. We would also like to add that we have not found any previous study that has used sequential analyses in an experimental design comparable to ours (i.e., with a large number of hypotheses).

*Response (also submitted to Reviewer 2, Comment 11): It seems unlikely that we would be able to stop the inclusion of participants at interim analyses 1 or 2 (i.e., after having included 50 or 75% of the final sample determined with a power analysis) having validated the 6 hypotheses we wish to test in our study. Indeed, a sensitivity analysis performed on G*Power 3.1.9 (Faul et al., 2007) revealed that: at time 1, with 106 participants, a target power of 0.964 and an alpha of .003, and using an RM-ANOVA, we will be able to highlight within-between interaction effects of size greater than or equal to d = 0. 75; at time 2, with 160 participants, a target power of 0.964 and an alpha of .018, and using an RM-ANOVA, we will be able to demonstrate within-between interaction effects of size greater than or equal to d = 0.58. Consequently, in order to increase the probability of stopping the experimental runs earlier and thus reduce the costs associated with our study (time, money), we decided to add stop conditions for futility, i.e. thresholds for which we consider that the chances of obtaining a significant effect of size d = 0.4 by adding participants will be null or too low to be relevant (see Lakens et al., 2021). These futility thresholds were determined via an O'Brien and Fleming-type beta spending function, on the model of the procedure used to determine efficacy thresholds. Thus, we will stop the analyses at Time 1 if, for each hypothesis test, p < .003; we will stop the analyses at Time 2 if, for each hypothesis test, p < .018 or p > .298; finally, at Time 3, we will consider that an effect is significant if p < .044. We added this to the manuscript on p. 15.*

See p.15: "By performing sequential analyses, it is possible to reduce the experimental sample size and thus minimize the cost of the experiment (i.e., time, money) thanks to the opportunity of rejecting the null hypothesis or stopping the study for futility at an interim look while controlling both Type 1 and Type 2 error rates. Efficacy and futility boundaries for interim analyses were calculated with an online Shiny application using functions from the R *rpact* package (Wassmer & Pahlke, 2020). Analyses indicated that with a two-tailed test with an alpha of .05, a power of 0.964 and three analyses using O'Brien and Fleming-type alpha

and beta spending functions, the efficacy boundary for the first analysis (time = 0.50) is .003; the efficacy and futility boundaries for the second analysis (time = 0.75) are respectively .018 and .298; the efficacy boundary for the third analysis (time = 1.00) is .044[1]. Thus, if the hypothesis tests all return a $p < .003$ after the first interim analyses (i.e., after 106 participants, including 53 young adults and 53 older adults), data collection will be interrupted; if the hypothesis tests all return a $p < .018$ or a $p > .298$ after the second interim analyses (i.e., after 160 participants, including 80 young adults and 80 older adults), data collection will be interrupted."

7) A suggestion: not obligatory, but it could be useful to create your analyses script in R, possibility with a dataframe with some dummy data. It helps to make the analyses plan really clear (and its of course helpful for you later).

Following your recommendations, we have prepared a complete script in R to analyze the data to be collected during the experimental runs, respecting all the steps and elements described in the manuscript.

The script will uploaded to OSF together with the pre-registration and can be accessed via the following link: https://osf.io/heu7c

8) Another suggestion: Even with high power, it is difficult to draw strong conclusions about any null results you might obtain. You could consider adding in either Bayes Factors of equivalency tests which can allow you to make inferences about null effects (Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. The Journals of Gerontology: Series B, 75(1), 45-57.) Note that the former requires an estimate of effect size and the latter requires an estimate of smallest effect of interest so which you use may depend on which of these you find easiest to estimate. Again, this isn't a requirement for stage 1 acceptance, just something you might like to consider.

Response: We thank the reviewer for this intrinsically very pertinent suggestion. However, in view of the large number of analyses planned for our study and their relative complexity, we

[1] A simulation study showed that the cumulative error probabilities for ANOVAs are similar to the probabilities of t-tests, especially when the number of looks of the data is small (i.e., $\leqslant 3$; Lang, 2017), so it is valid to use the computational methods of the R package *rpact* (Wassmer & Pahlke, 2020) in our situation.

prefer to confine ourselves to the analyses indicated in the first version of the manuscript (i.e., NHST) in order to maintain sufficient clarity for the reader.

One final point: Apologies if I somehow missed it, but I couldn't see where you stated the age boundaries for your groups, which seems critical.

Response: Age limits for both groups of participants (i.e., younger and older) were added to the inclusion criteria.

See p. 17: "In order to be included in our study, participants will have to meet the following criteria: be between 18 and 30 years old or between 60 and 80 years old; be native French speakers or fluent in French; have normal or corrected-to-normal vision; have no history of major neurological (e.g., stroke, head injury) or psychiatric disorders (e.g., depression, bipolar disorder)."

Reviewer 1 – Mara Mather

As outlined in this study preregistration, researchers will examine how positive vs. negative emotion modulate item memory, associative memory and temporal memory for younger vs. older adults.

Participants will be shown short videos depicting everyday life activities with positive, negative and neutral pictures inserted into these videos to be shown for 2 s each at a random time. For each combination, participants will be asked to rate the compatibility between the video and the picture. After completing this phase and completing questionnaires during a 45-min delay, participants will be shown old and new images and tested on their item recognition. They will also be tested on their memory for where in each 20-s video the temporal placement of the picture occured, and their memory for which screenshot (from 5 different videos) was from the same video context as the one associated with the picture.

This is a well-designed study. My main concern was whether the temporal memory question was being tested appropriately given prior findings.

The authors previously discussed discrepancies between the findings from Ceccato et al. (2022) and Palumbo et al. (2018) and suggested that the 48-delay in one study but not the other might have contributed to the differences. A critical aspect of this long a delay is that it includes sleep. Jones, Schultz, Adams, Baran, & Spencer (2016) found that the emotional bias of sleep-dependent processing shifts from negative to positive in aging. Thus, increasing the planned delay from 45 min to a day or two, in order to include sleep in the delay period, may increase the likelihood of detecting age differences.

The current study also uses the Ceccato et al. (2022) results to motivate the investigation of the memory for temporal context. However, the Ceccato et al. study "exposed young and older adults to an experimental task in which they saw negative, neutral, and positive images in three sessions that occurred 48 hours apart." This is quite different than the current study's memory test where participants will be asked to indicate where in each 20-s short video context each picture was inserted. This is temporal context relative to each video, not relative to the participant's sense of how long it has been since they viewed the pictures, which could have been the driving influence in Ceccato et al.

These prior results suggest that to more effectively test their hypotheses, the authors should consider including at least two encoding sessions that are each separated by overnight sleep.

The passage of time between different encoding contexts and the memory tests is relevant not only for the question about temporal memory, but also for the age-related positivity effect in item memory. "The second modification to this paradigm will be to increase the delay between encoding and retrieval phases from 10 min in the Palombo et al. (2021) paradigm to 45 min, as the age-related positivity effect in item memory has been shown to increase with time (e.g., Kalenzaga et al., 2016; Laulan et al., 2020)." It would of course be interesting to manipulate these time variables, but given the already many measures in the study, the key goal to focus on would be to select parameters that would increase power to detect the proposed outcomes. The prior research cited in the paper indicating that longer delays may increase the positivity effect and the Jones et al (2016) findings that sleep may differentially enhance negative memories in younger vs. older adults suggests that having 2 or more encoding sessions should not reduce power to detect the age-related positivity effect.

Response: We thank the reviewer for raising this important point of discussion concerning the methodology of the experiment. The main reason that we identified for not adding a longer delay between encoding and retrieval in our study stemmed from the results obtained by Waring and Kensinger (2009). To our knowledge, these are the only researchers who tested how age and the delay between encoding and retrieval influence the effect of emotion on memory for the extrinsic context of an item. The results of their study revealed that after a 24-hour delay, there was no further decrease in memory for the background associated with an emotional item in older adults. The authors explained that this lack of effect in older adults after a delay could stem from differences in the way young and older adults encode and hold representations of items and backgrounds in memory. Indeed, if older adults have difficulty associating an item and its background within a memory representation, it's possible that over time the processing of the item no longer influences the processing of the background. Thus, following the example of other studies that have examined the effect of emotions on associative memory in aging (see Nashiro & Mather, 2010, 2011), we consider it prudent in this study to maintain a short delay between encoding and retrieval phases.

Nevertheless, we agree that the question of the impact of sleep on emotional biases in associative memory is very interesting. Consequently, we plan to launch a parallel study in which we will investigate the effect of sleep (i.e., 12 h delay between encoding and retrieval without sleep vs. 12 h delay between encoding and retrieval with sleep) in a paradigm similar to that used in the present study. This study will initially be conducted in young adults and potentially extended to older adults depending on the results obtained.

Reviewer 2

In a registered report, Dr. Laulan et al. plan to explore different facets of emotional memory in younger and older adults. The study is modelled after a previous report by Palombo et al., (2021), with methodological differences clearly noted. This RR is a clear, mostly well written registered report and could make for an important contribution to the literature. Many of the methodological choices are well conceived and rigorous. However, I do have some major and minor reservations about the reporting of prior literature, the rationale, hypotheses, and the approaches used. I hope my comments are constructive.

Response: We thank the reviewer for calling our RR clear, mostly well written with the potential of making an important contribution to the literature. We thank the reviewer for the constructive comments, which helped us improve our RR.

Intro:

1. The authors seem confident that emotion will impair associative memory even for positive stimuli. Is the literature so clear on this? Work by Christopher Madan shows that emotion can sometimes enhance associative memory for positive stimuli (e.g., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6612425/). Does such work inform the hypotheses here and if not, why not?

Response: We thank the reviewer for raising this important point for discussion. Concerning the effects of valence on the memory for item extrinsic context in young adults, we based our hypotheses on the results of the meta-analysis by Pereira et al. (2022). These authors showed that negative valence has a detrimental effect on memory for item extrinsic contexts, and that this effect is generalizable to positive valence, albeit of lesser magnitude (d = 0.21 vs. d = 0.12). We are aware of the results obtained by Madan et al. (2019) but feel that they do not challenge our hypotheses given that the beneficial effect of positive emotions on associative memory emerged in their paradigm only when both items in a pair to be memorized were positive but not when one item was positive and the other neutral, as will be the case in our study. However, in order to better convey to readers the subtlety of emotional effects on memory for item extrinsic context and on associative memory, we have added to the introduction a brief description of the results obtained by Madan and colleagues (Madan et

al., 2019; Madan et al., 2020) that may qualify the conclusions presented in the first version of our manuscript.

See p. 7: "It is also important to note that these effects remain subtle and sensitive to the methodology used to test them. Thus, in a series of studies using a cued recall paradigm to disentangle item- and association-memory contributions, Madan and colleagues have reported that: 1) when individuals are asked to learn pairs of neutral or positive words, positive emotions enhance associative memory, but only when both words in a pair are positive (Madan et al., 2019); 2) when scenes consisting of a neutral vs. negative object and a neutral background are presented to individuals, negative emotions have a beneficial effect on item memory, deleterious on background memory, and beneficial on associative memory, and this pattern of results could stem from the fact that, contrary to the majority of studies, the associated stimuli were congruent (Madan et al., 2020).

In summary, in young adults, the effect of emotion on memory for intrinsic item features and item extrinsic context are very sensitive to the choices made by researchers when setting up their experimental paradigms (e.g., selection of material and operationalization of concepts such as time and location), and the data in the literature suggest that: 1) memory for intrinsic item features is enhanced for emotional *vs.* neutral items, and in the specific case of temporal memory, the magnitude of this effect would be greater for negative *vs.* positive items; 2) memory for extrinsic item context is disrupted for emotional *vs.* neutral items, with no difference between positive and negative items.

2. I do not think enough granularity when describing Petrucci and Palombo (2021)'s conclusions; the author of the current RR state that: "Petrucci and Palombo (2021) explained that the majority of studies on the subject have found that emotion enhances the memory of temporal information." But it seems that Petrucci and Palombo (2021) highlight that temporal memory is not 'one thing' and the effects of emotion on this form of remembering look different depending on whether one is considering e.g., item order, source memory, or duration. It seems there are a lot of mixed findings and the literature is too nascent to draw conclusions. The footnote in the RR is helpful but I do not believe it goes far enough. The same concern applies to the statement: "the beneficial effect of emotion on memory for temporal information appears to be robust (Petrucci & Palombo, 2021)". Can the authors tone this down a bit?

Response: We thank the reviewer for noting this lack of nuance in the presentation of the findings of Petrucci and Palombo's (2021) literature review concerning the effect of emotion on temporal memory. We have reformulated the parts of our manuscript that deal with this topic to better account for the multiple ways in which temporal memory can be operationalized (e.g., temporal order, source memory, duration memory) and consequently for the variable aspect of the effect of emotion on this facet of memory.

See p. 5: For example, in a review of the literature, Petrucci and Palombo (2021) explained that the effect of emotion on temporal memory depends directly on how the latter is operationalized, i.e., either as memory of the moment when an event took place, or "when" memory, or as memory of the duration of an event, or "how long" memory. If we focus on "when" memory, and more specifically on source memory (which can be distinguished from temporal order memory), Petrucci and Palombo (2021) concluded that most studies have found that emotion enhances memory for temporal information.

See p. 15: The use of this paradigm is also relevant because, in young adults, the beneficial effect of emotion on memory for temporal information seems to be robust when it concerns the moment when an event occurred.

3. Elsewhere, it seems that the authors are trying to argue that when you parse different types of intrinsic features (e.g., temporal versus non temporal), one can better explain inconsistencies in the literature. But the evidence provided in the intro is pretty thin on that front. Related, at times the study dances around the idea of comparing the effects of agingXemotion on intrinsic versus extrinsic features of an experience but the current study design is not optimized for that particular comparison as it potentially conflates temporal versus non temporal memory with intrinsic versus extrinsic—if we see different patterns in the different subtasks, would we be able to conclude it is due to the "intrinsicisity" of the task?

Response: We thank the reviewer for this comment. After re-reading, we agree with the reviewer that evidence is thin for explaining the inconsistencies in the literature by parsing different types of intrinsic item features.

We now state this issue in the introduction:

See p. 5-6: “However, Pereira et al. (2022) reported in their meta-analysis that, overall, these effects are not robust, and that emotion does not significantly influence memory for the intrinsic features of items, either when comparing negative *vs.* neutral items (color: $p = .393$; location: $p = .910$; time: $p = .204$) or positive *vs.* neutral items (color: $p = .538$; location: $p = .730$; time: $p = .571$). One reason for the lack of reliability of these effects is that they can vary significantly depending on how each type of intrinsic feature is implemented in experimental paradigms.”

Furthermore, concerning the reviewer’s statement that “at times the study dances around the idea of comparing the effects of agingXemotion on intrinsic versus extrinsic features of an experience but the current study design is not optimized for that particular comparison as it potentially conflates temporal versus non temporal memory with intrinsic versus extrinsic”, we established our hypotheses based on the object-based binding account proposed by Mather (2007), given that this is the interpretative framework generally used to distinguish the effect of emotions on memory for time or location (considered intrinsic features of the items) vs. memory for background or peripheral details such as color (considered contextual elements extrinsic to the items) (see e.g. Rimmele et al., 2011, 2012). Also, in the discussion of their study, Palombo et al. (2021) suggested that their paradigm could potentially increase the centrality with which temporal features are processed, which could explain the emergence of a beneficial effect of negative vs. neutral valence observed in temporal judgment task. Thus, in our study, we have recourse to the theory proposed by Mather (2007) but we do not aim to test it as was the case, for example, by MacKenzie et al. (2015) who made the same type of information (color) alternately an intrinsic or extrinsic element of the items to be memorized. Consequently, the intrinsic vs. extrinsic distinction will be one possible explanation if different effects emerge in our study between temporal judgment and associative recognition tasks, but it will not be the only one (e.g., we may dwell in the discussion on the particularities of time vs. other types of information that can be associated with an event).

4. More critical, the findings of Palombo et al., 2021 appear a bit more nuanced than the authors allude to in the introduction. In that paper, the authors examine the bias versus sensitivity of temporal memory and find that the effects of emotion on temporal memory are driven by bias, not by precision. Indeed, to quote Palombo et al: “Critically, we note that in our study, negative emotion did not affect precision per se,

but it did affect participants' responding; in the neutral condition, there was a shift to later temporal estimates. In other words, when participants made temporal judgments in the neutral condition, they tended to judge the events as having happened later. By contrast, in the negative condition, participants' responses were not consistently biased to be either early or late. If timing was encoded as an intrinsic feature that was enhanced by negative emotion, we might have expected to see enhanced precision in the negative condition, but instead, we observed comparable precision for both conditions. It is not clear what mechanism can account for the present results." How does this nuance inform the author's current hypotheses? Is this the right paradigm, to test the authors' hypotheses? It seems that the authors are interested in temporal precision?

Response: We thank the reviewer for this crucial question. In our study, we would like to assess temporal judgment accuracy based on descriptions of this measurement found in the literature, which allow us to distinguish it from the measure of temporal judgment bias. (e.g., see Schönenkorb et al., 2023):

1) *temporal judgment accuracy: a measure of how close participants' estimates of the timing of images are to the actual moment when these images appeared in the videos (formula: /Estimated Time-Actual Time/). That is, this calculation of accuracy scores for each trial is* ***based on the absolute value*** *of the difference between each participant's response along the timeline and the actual moment when the image was placed in the video*
2) *temporal judgment bias: a measure of the tendency of participants to estimate that images appear on average earlier or later in the video than they actually do (formula: Estimated Time-Actual Time). That is, this calculation of accuracy scores for each trial is* ***based on the relative value*** *of the difference between each participant's response along the timeline and the actual moment when the image was placed in the video.*

n the study by Palombo et al. (2021), the authors indicated how they measured accuracy in their pre-registered analyses: "Accuracy was assessed by first subtracting each participant's response along the timeline from the actual time when the image was placed in the video for each trial (i.e., the amount of error in their estimate) and then averaging across trials for each participant". It should be noted that this computation corresponds to the formula

provided by Schönenkorb et al. (2023) for assessing temporal judgment bias, not the formula for assessing temporal judgment accuracy. Using this formula, it is possible to obtain scores of differences between the participant's response along the timeline and the actual moment when the image was placed in the video, which are sometimes positive and sometimes negative. These scores may offset each other when averaged. This approach is therefore unsuitable for assessing the accuracy of temporal judgments, as it may lead to an underestimation of this variable. Thus, it seems preferable to use the formula for the accuracy of temporal judgments as defined by Schönenkorb, as this is based on the absolute value of the difference scores between the participant's response along the time line and the actual moment when the image was placed in the video, rather than on the raw value.

Nevertheless, Palombo et al. (2021) carried out additional analyses to distinguish response bias vs. accuracy, using cumulative distribution functions. These analyses revealed no significant effect of emotion on temporal judgment accuracy. However, these analyses had not been pre-registered, which means we have no information on their statistical power. The authors reported that for cohorts A and B, the 95% confidence intervals (CI) of the fitted shape parameter that corresponds to the distribution slope were respectively [3.59, 3.90] and [3.66, 3.96] for the negative condition and respectively [3.85, 4.22] and [3.93, 4.24] for the neutral condition. Possibly, with greater statistical power, these intervals would not have overlapped and could have shown differences in accuracy of temporal judgments for neutral vs. negative images.

In conclusion, the results obtained by Palombo et al. (2021) do not call into question our hypotheses, which were formulated on the basis of an exhaustive review of the literature (e.g., see Nashiro et al., 2011; Palumbo et al., 2018). Rather, our aim is to extend the work of Palombo et al. (2021) by robustly and validly assessing the accuracy of temporal judgments. To this end, we will: 1) calculate accuracy scores for each trial ***based on the absolute value*** *of the difference between each participant's response along the timeline and the actual moment when the image was placed in the video; 2) recruit a sufficient number of participants to be able to* ***identify small to medium effect sizes*** *(i.e., d = 0.40).*

See p. 26: “Furthermore, for the analysis of responses in the item recognition task, we will use a binomial distribution and a probit link function (e.g., DeCarlo, 1998; Hourihan et al., 2013); for the analysis of responses in the temporal judgment task, we will calculate accuracy scores for each trial based on the absolute value of the difference between each participant's response

along the time line and the actual moment when the image was placed in the video (e.g., Schönenkorb et al., 2023) and we will use a Gaussian distribution and an identity link function; for the analysis of responses in the associative recognition task, we will use a binomial distribution and a logit link function (Ballinger, 2004)."

We have also reformulated the presentation of the results of Palombo et al. (2021) in the light of the elements discussed above.

See p. 8: "With this new paradigm, they found: 1) enhanced item memory for negative *vs.* neutral images which had been inserted in videos and then presented to young adults; 2) later time judgment for the moment of appearance of neutral vs. negative images within the videos[2] ; and 3) weaker memory for the association between negative images and videos *vs.* neutral images and videos."

5. With respect to item memory effects, my comments are more minor but the introduction raised a number of questions for me: Does emotion only affect only discrimination of items or does it also affect bias? (See important work by Dougal & Rotello, 2007.) The authors state that the memory advantage for emotional vs neutral is greater for negative than for positive stimuli. How much of this is due to insufficient matching of arousal in prior studies between positive and negative? The authors cite the roles of attention and consolidation. What about the role of semantic relatedness (a la Deborah Talmi's work?). I realize that the introduction is quite long already but it would be nice to include some of this granularity in the introduction.

Response: We have rewritten the entire paragraph on the effect of emotion on item memory, incorporating the reviewer's suggestions to better reflect the complexity and subtlety of the effects discussed.

See p. 4: "This beneficial effect of emotion can be explained by several mechanisms. When the delay between encoding and retrieval phases is not long enough to enable consolidation processes to develop, emotional memory enhancement depends mainly on three cognitive factors related to the characteristics of emotional *vs.* neutral stimuli (e.g., Talmi & McGarry, 2012): (1) greater semantic relatedness for emotional *vs.* neutral stimuli (e.g., Talmi &

[2] Palombo et al. (2021) did not calculate temporal judgment accuracy scores in their pre-registered analyses, so we have no reliable information concerning a potential effect of emotion on this variable.

Moscovitch, 2004; Talmi et al., 2007), which may facilitate the organization of emotional stimuli around a common theme and thus make their encoding deeper and retrieval easier (e.g., Einstein & Hunt, 1980); (2) greater distinctiveness for emotional *vs.* neutral stimuli (i.e., emotional stimuli possess unique attributes that they do not share with neutral stimuli; Talmi et al., 2007), which may facilitate the retrieval of emotional stimuli after memory search (Tomlinson et al., 2009); (3) greater attentional capture in favor of emotional *vs.* neutral stimuli (e.g., Schmidt & Saari, 2007; Vuilleumier, 2005), resulting in deeper processing of emotional stimuli during encoding and better memory retrieval (Talmi et al., 2008). Of these three factors and mechanisms underlying the effect of emotions on memory, increased attentional capture is likely to enhance memory in recognition and free recall tests, while the greater semantic relatedness and distinctiveness of emotional stimuli could predominantly improve recall accuracy, but not recognition accuracy (Schümann et al., 2018). In addition, the greater attentional capture of emotional vs. neutral stimuli may explain the more frequent experience of recollection during the recognition of emotional items (Kensinger & Corkin, 2003), while the greater semantic relatedness of emotional vs. neutral stimuli could account for a more liberal response bias for emotional vs. neutral stimuli (Dougal & Rotello, 2007; see Bennion et al, 2013). Thus, the better performance observed for emotional vs. neutral stimuli in memory tasks may reflect better recall for emotional stimuli but also a response bias in favor of emotional stimuli (Thapar & Rouder, 2009). In young adults, the memory advantage for emotional *vs.* neutral stimuli is greater for negative than for positive stimuli (see Baumeister et al., 2001), although the magnitude of this effect may be overestimated due to a lack of control of arousal between negative *vs.* positive stimuli in many experimental studies, with the arousal of negative stimuli generally higher than that of positive images (see Williams et al., 2022)."

Methodology:

1. Do the authors have any exclusionary criteria for data quality that they want to report (e.g., if a participant misses X trials; does not move around the scale during encoding, shows signs of inattention or rushing; ceiling effects)?

Response: In addition to the data exclusion criterion based on score deviation from the median in each of the experimental tasks, we have added two exclusion criteria in the manuscript: 1) a participant's data will be removed from the final analyses if s/he did not correctly complete all of the experimental phases (e.g., lack of attention, computer problem);

2) a participant's data will be removed from the final analyses if we observe that s/he did not respond to 25% or more of the trials in at least one of the experimental phases.

See p. 23-24:
We have defined three criteria according to which a participant's data may be excluded from the final analyses. First, a participant's data will be excluded if s/he explains during the debriefing that s/he did not perform all the experimental phases correctly for reasons beyond her/his control (e.g., s/he was considerably interrupted or disturbed by a person or alarm while performing a phase of the experiment) and/or within her/his control (e.g., s/he was not concentrating because s/he regularly looked at her/his phone or because s/he was tired). This information will be completed and nuanced by the observations noted by the experimenters during the experimental runs. Second, a participant's data will be excluded if s/he fails to provide a response for 25% or more of the trials in at least one of the four experimental phases (see Palombo et al., 2021). Third, for the three dependent variables in our study (i.e., responses to the item recognition task, responses to the temporal judgment task, and responses to the associative recognition task), we will calculate the median and Median Absolute Deviation (MAD) for each age group. If any participants have scores that fall beyond 3 MAD of the median for a given dependent variable, her/his data will be excluded from the analyses. We decided to use MADs rather than standard deviations to identify potential outliers because standard deviations are measures that are sensitive to extreme values and therefore do not robustly detect them (Leys et al., 2013; see Monéger et al., 2022).

2. How do the authors factor in within subject power? For example, by introducing positive stimuli, the authors are using less trials per condition than I have seen in other studies. How does this compare to e.g., Palombo et al., 2021? Can the authors discuss this?

Response: We chose to keep the total number of stimuli equivalent to that used in the study by Palombo et al. (2021) despite the addition of an experimental condition (i.e., positive stimuli) so that the cost of the task would not be too high in older adults and to reduce the risk of observing a floor effect.

To maintain sufficient statistical power, several precautions were taken:
- 1) we chose a SESOI of d = 0.40, more conservative than the a priori effect size retained by Palombo et al. (2021), i.e., d = 0.44 for an effect of emotions in the item memory task and d =

0.52 for an effect of emotions in the temporal judgment task;
- 2) we plan to use more powerful statistical analyses than those used by Palombo et al. (2021): with GEE the data will be processed trial by trial and not averaged as is the case with ANOVAs, and the use of a priori contrasts will reduce the number of degrees of freedom compared with omnibus ANOVAs.

3. I suggest the authors unpack H1-3 with respect to hypothesis b ("There is an age-related positivity effect, such that…" to unburden the reader)

Response: The sub-hypotheses b of each of the 3 main hypotheses have been reformulated to specify the pattern of results expected in each case if an age-related positivity effect is found.

See p. 13-14:
"H1 – b: There is an age-related positivity effect, such that positive images are better remembered than negative images in older adults *vs.* younger adults (e.g., Reed et al., 2014).

H2 – b: There is an age-related positivity effect, such that positive images are positioned more accurately in time than negative images in older adults *vs.* younger adults.

H3 – b: There is an age-related positivity effect, such that associations between positive images and videos are less well remembered than associations between negative images and videos in older adults *vs.* younger adults (Waring & Kensinger, 2009)."

4. For Leclerc & Kensinger (2011) description, please first describe the scales (ranges from X to Y)—for example, is 9 the highest possible score on the scale?

Response: The valence and arousal of the images in Leclerc & Kensinger's (2011) study were indeed rated on scales from 1 to 9.

See p. 17:
"All these images were from the International Affective Picture System (IAPS, Lang et al., 2005) and their valence and arousal were rated on scales from 1 to 9 by 50 young and 50 older adults."

5. The authors did a nice job with stimuli matching. But it would be helpful to show e.g., superimposed histogram/density style plots of the arousal and valence ratings in younger and older adults for the items chosen to demonstrate not just that the mean is

matched but that the distributions are also matched across groups. The SDs give some sense but is not complete. The range would be helpful too. This is not often seen in the literature (unfortunately) but would bolster the notion of tight matching at the manipulation level or at least demonstrate how tight the matching really is.

Response: We thank the reviewer for this comment. We agree that it is very informative to provide superimposed histogram/density style plots of the arousal and valence ratings in younger and older adults. In order to better account for the matchings made regarding the valence and arousal ratings of the images selected for our study as a function of age, we followed the suggestions proposed by the reviewer and 1) added the ranges of valence and arousal ratings as a function of image category (i.e., negative, neutral and positive) and age, 2) added graphs representing the density of valence and arousal ratings as a function of age to the manuscript.

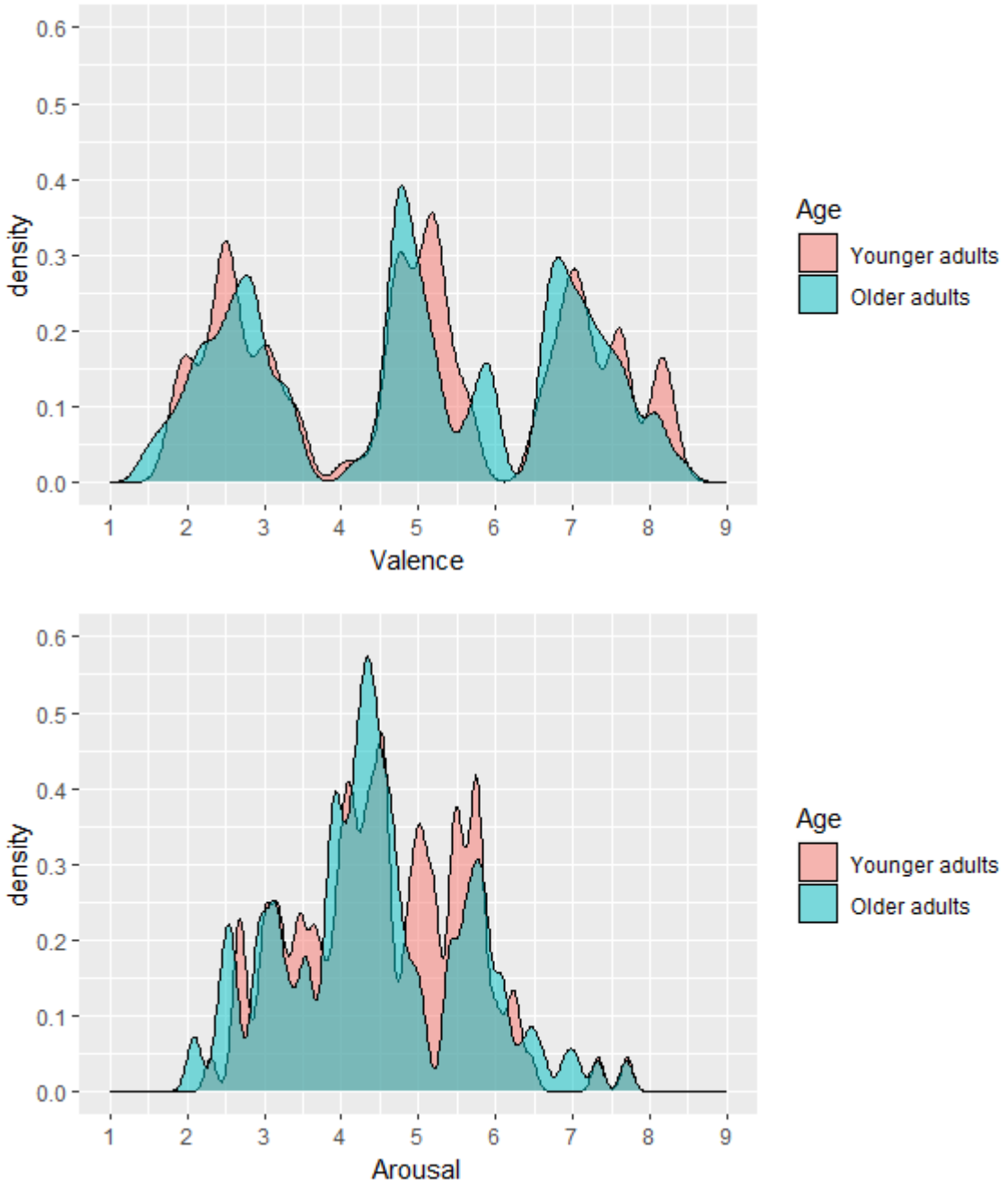
See p. 18-19:

"Based on image ratings collected by Leclerc & Kensinger (2011), we ensured that negative images in our study have a valence between 1 and 3.5, both in younger and older adults (in Sets 1 and 2 combined, younger adults: [1.80, 3.47]; older adults: [1.45, 3.46]); neutral images have a valence between 4 and 6, both in younger and older adults (in Sets 1 and 2 combined, younger adults: [4.04, 5.69]; older adults: [4.18, 5.95]); positive images have a valence between 6.5 and 9, both in younger and older adults (in Sets 1 and 2 combined, younger adults: [6.51, 8.34]; older adults: [6.55, 8.42]). Therefore, valence differed significantly between negative, neutral, and positive images in both younger and older adults (all *ps* < .001). Furthermore, the images in our study were selected so that the arousal of negative and positive images did not differ significantly in either young or older adults (all *ps* > .337). Also, in selecting our images, we sought to keep the arousal of negative and positive images as low as possible (see Waring & Kensinger, 2009), so that the highest arousal for a negative image is 7.71 whether in younger or older adults (in Sets1 and 2 combined, younger adults: [3.67, 6.46]; older adults [3.81, 7.71]) and that the highest arousal for a positive image is 7.70 whether in younger or older adults (in Sets 1 and 2 combined, younger adults: [3.67, 6.46]; older adults [3.80, 7.70]). However, due to the U-shaped relationship between valence and arousal frequently obtained in normative studies (e.g., Lang et al., 1998), arousal for negative and positive images was higher than for neutral images in both younger and older adults (in Sets 1 and 2 combined, younger adults: [2.30, 4.38]; older adults: [2.05, 4.86]) (all *ps* < .001).

Finally, valence and arousal were matched across the two age groups of our study for each category of images (i.e., negative, neutral, and positive images) (all *ps* > .088). The distribution of valence and arousal scores for Set 1 and Set 2 images as a function of age is shown in Figure 1."

Figure 1

Density plot showing the distribution of valence and arousal scores for all images in our study (i.e., both Set 1 and Set 2) as a function of age (i.e., younger vs. older adults)



6. The authors state, "Finally, valence and arousal were matched across the two age groups of our study for each category of images (i.e., negative, neutral, and positive images) (all ps > .05)." Might you want to be more conservative (e.g., .2 or no greater than x effect size)

Response: It was technically impossible for us to be more conservative on the valence and arousal matchings between younger and older adults across the three image categories (i.e., negative, neutral and positive) while respecting the other matchings.
We specified in the manuscript that for the most "lax" matching, the result of the means comparison test provided a p = .088 (comparison of image valence ratings collected from younger vs. older adults for neutral images from Set 2).

See p. 18: "Finally, valence and arousal were matched across the two age groups of our study for each category of images (i.e., negative, neutral, and positive images) (all *ps* > .088)."

7. It seems that the authors strived for low arousal images —why? ("Also, in selecting our images, we sought to keep the arousal of negative and positive images as low as possible (see Waring & Kensinger, 2009"). Don't other studies try to aim for high *arousal, high valence? Sorry if I misunderstood something.*

Response: The choice to select negative and positive images with the lowest possible arousal level is based on results obtained in studies that have simultaneously manipulated valence, arousal and age in item memory tasks (Kensinger, 2008; Laulan et al, 2022) or item-context memory tasks (Waring & Kensinger, 2009). First, in studies by Kensinger (2008) and Laulan et al. (2022), it was shown that the age-related positivity effect in item memory emerged preferentially for low-arousal item vs. high-arousal item. Then, in the study by Waring & Kensinger (2009), it was shown that in young adults, backgrounds associated with negative images were less well remembered than backgrounds associated with positive images; in older adults, no differences in memory were shown between backgrounds associated with negative and positive images. However, if we look more closely at the results, it would appear that in older adults, backgrounds associated with low-arousal positive images were less well remembered than backgrounds associated with low-arousal negative images, while the opposite was true for high-arousal images. Thus, although no significant effect was found in

this sense, the data from Waring and Kensinger (2009) suggest that an age-related positivity effect in item-context memory would be more likely to emerge for low-arousal item rather than high-arousal item. To take this into account, we now state the following in the manuscript:

See p.19: "Also, when selecting images, we sought to keep the arousal of negative and positive images as low as possible to maximize the probability of observing an age-related positivity effect in item memory (Kensinger, 2008; Laulan et al., 2022) and in item-context memory (Waring & Kensinger, 2009)."

We also point out in a footnote that this choice may have certain limitations, particularly in the case of young adults.

See p. 19: "However, this choice may reduce the possibility of observing an effect of emotions on item-context memory in young adults, since it has been shown in individuals of this age group that high-arousal items are remembered more often with spatial and temporal context than low-arousal items (Schmidt et al., 2011)."

8. Did the authors match the categories of images (landscapes, animals etc across images) as is common in emotional memory studies. These categories are provided in NAPS but can be done for IAPS too. This will keep more of a distinctiveness of items within condition and avoid confusability in the memory tasks, (I am sure the authors are aware that this can be an issue in emotional memory studies as emotional stimuli tend to come from narrower semantic or perceptual themes (see Deborah Talmi's work on this).

Response: Although not indicated in version 1 of the registered report, efforts were made to match the different image categories (i.e., negative, neutral and positive) according to their content, in sets 1 and 2 separately. This information has been added to the manuscript.

See p. 19: "In addition, three research assistants coded the content of images from the study by Leclerc and Kensinger (2011) following the instructions described in the article by Marchewka et al. (2014). Thus, each image was categorized as animal, object, landscape, people or face, and only images for which all three coders provided identical responses were selected as potential candidates for inclusion in our study. In the final materials of our study,

we ensured that image content did not differ between negative, neutral and negative images in both Set 1 and Set 2[3] (all *ps* > .731)."

The exact distribution of image content has also been added.

See p. 19 (footnote):
"The content of the images in our study for each set and for each category was as follows:
- *Set* 1: negative images = 2 with animals, 6 with objects, 2 with landscapes, 7 with people, 3 with faces; neutral images = 1 with animals, 7 with objects, 2 with landscapes, 7 with people, 3 with faces; positive images = 3 with animals, 5 with objects, 3 with landscapes, 6 with people, 3 with faces.
- Set 2: negative images = 2 with animals, 5 with objects, 3 with landscapes, 8 with people, 2 with faces; neutral images = 1 with animals, 7 with objects, 3 with landscapes, 5 with people, 4 with faces; positive images = 3 with animals, 5 with objects, 3 with landscapes, 6 with people, 3 with faces."

9. The authors state that "Each image will be displayed for 3 s (e.g., see Palumbo et al., 2018)"—> did the authors mean Palombo et al., 2021 instead? And do timings need to be adjusted in light of the inclusion of older folks?

Response: In agreement with the reviewer, we believe in hindsight that a time of 3 s is too short to allow all participants to respond to the trials of the item recognition task, particularly for the older participants. Consequently, we decided to extend the response time to 8 s, i.e. a time equivalent to that allowed for the temporal judgment task and the associative recognition task.

See p. 21: "Each image will be displayed for 8 s *vs.* 4 s in the study by Palombo et al. (2021), given the inclusion of older adults in our study, and participants will have to determine if they have seen the image before by answering "Old" or "New" using two keys on the keyboard of the computer ("Old": Q; "New": P)."

10. Are there any concerns about administering "emotional" type questionnaires in the delay period? Could that affect early consolidation?

[3]

Response: To avoid any bias in our data resulting from the administration of emotion-related questionnaires to participants, we decided to have participants complete these questionnaires at the very end of the experiment. We have therefore modified the experimental procedure in the manuscript so that the last phase (i.e., phase 5) is devoted to evaluating the participants' characteristics and debriefing (see p. 24).

See p. 23: "After a 45-minute break during which participants can walk around and do some coloring (see Palombo et al., 2021), the images presented during the encoding phase will be shown again to the participants, as well as new images serving as distractors."

See p.24: "***Phase 5: Assessment of participants' characteristics and debriefing*** **"**

11. How will the sequential analyses pertain to the different outcomes--are sequential analyses valid in this way when one needs to observe significant effects across the board?

*Response: It seems unlikely that we would be able to stop the inclusion of participants at interim analyses 1 or 2 (i.e., after having included 50 or 75% of the final sample determined with a power analysis) having validated the 6 hypotheses we wish to test in our study. Indeed, a sensitivity analysis performed on G*Power 3.1.9 (Faul et al., 2007) revealed that: at time 1, with 106 participants, a target power of 0.964 and an alpha of .003, and using an RM-ANOVA, we will be able to highlight within-between interaction effects of size greater than or equal to d = 0. 75; at time 2, with 160 participants, a target power of 0.964 and an alpha of .018, and using an RM-ANOVA, we will be able to demonstrate within-between interaction effects of size greater than or equal to d = 0.58. Consequently, in order to increase the probability of stopping the experimental runs earlier and thus reduce the costs associated with our study (time, money), we decided to add stop conditions for futility, i.e. thresholds for which we consider that the chances of obtaining a significant effect of size d = 0.4 by adding participants will be null or too low to be relevant (see Lakens et al., 2021). These futility thresholds were determined via an O'Brien and Fleming-type beta spending function, on the model of the procedure used to determine efficacy thresholds. Thus, we will stop the analyses at Time 1 if, for each hypothesis test, p < .003; we will stop the analyses at Time 2 if, for each hypothesis test, p < .018 or p > .298; finally, at Time 3, we will consider that an effect is significant if p < .044.*

See p.15: "By performing sequential analyses, it is possible to reduce the experimental sample size and thus minimize the cost of the experiment (i.e., time, money) thanks to the opportunity of rejecting the null hypothesis or stopping the study for futility at an interim look while controlling both Type 1 and Type 2 error rates. Efficacy and futility boundaries for interim analyses were calculated with an online Shiny application using functions from the R *rpact* package (Wassmer & Pahlke, 2020). Analyses indicated that with a two-tailed test with an alpha of .05, a power of 0.964 and three analyses using O'Brien and Fleming-type alpha and beta spending functions, the efficacy boundary for the first analysis (time = 0.50) is .003; the efficacy and futility boundaries for the second analysis (time = 0.75) are respectively .018 and .298; the efficacy boundary for the third analysis (time = 1.00) is .044[4]. Thus, if the hypothesis tests all return a $p < .003$ after the first interim analyses (i.e., after 106 participants, including 53 young adults and 53 older adults), data collection will be interrupted; if the hypothesis tests all return a $p < .018$ or a $p > .298$ after the second interim analyses (i.e., after 160 participants, including 80 young adults and 80 older adults), data collection will be interrupted."

We would also like to inform the reviewer that we have modified the parameters concerning the power analysis. We initially set the target power for each interaction effect at 0.90, but then realized that since our study includes six independent hypotheses, with this approach we would in fact only have a power of 0.53 to test simultaneously all 6 hypotheses (i.e., 0.90^6). Consequently, we ran our power analysis again, this time with a target power of 0.964 (i.e., the sixth root of 0.80), and the results showed that a sample size of 212 participants is required in our study (vs. 160 in the old analysis).

See p. 14-15: "A power analysis performed with MorePower (version 6.0.4; Campbell & Thompson, 2012) indicated that for a repeated-measures ANOVA (RM-ANOVA), a total of 212 participants is required to show a within-between interaction effect (i.e., 3 [Valence: positive vs. neutral vs. negative] x 2 [Age: young vs. older adults]; see hypotheses 1b, 2b and 3b) of size $d = 0.40$ with an alpha of .05 and a power of .964."

[4] A simulation study showed that the cumulative error probabilities for ANOVAs are similar to the probabilities of t-tests, especially when the number of looks of the data is small (i.e., $\leqslant 3$; Lang, 2017), so it is valid to use the computational methods of the R package *rpact* (Wassmer & Pahlke, 2020) in our situation.

12. In the study design template table, the reference to the amygdala seems ill fitting and uses reverse interference.

Response: We agree with the reviewer. Reference to the amygdala has been removed from the study design table.

13. Perhaps I missed this but will participants have a chance to practice the task first? I think practice is important given the inclusion of older adults.

Response: Thank you for your comment. In order to make sure that all participants have understood the instructions of the encoding phase, and in particular concerning the evaluation of congruence between an image and the video in which it was presented, participants will perform a training session consisting of 3 trials.

See p. 23: To ensure that the instructions are fully understood, all participants will perform three practice trials before starting the experimental task.

14. I do not have strong expertise in GEE; I am familiar with the approach and the analyses seem reasonable but I flag this as a place of weakness in my expertise

Response: We had our experimental approach looked through and informed by a statistician.

15. Minor: Consider rephrasing "power deficits" as "insufficient power"

Response: We have followed the reviewer's recommendation in the section where we discuss statistical power in studies on the effect of emotions on item-context memory.

See p.21: "In order to determine whether the studies by Waring and Kensinger (2009) and Nashiro and Mather (2010) also had insufficient power, we conducted sensitivity analyses on these two studies with the same parameters used by Ceccato et al. (2022), i.e., an alpha of .05 and a desired power of .80."