

Behavioural separation of face memory and face perception

Corresponding Author: Dr Michal Ramot

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this paper, Kadlec and colleagues report results using their Face Memory and Perception (FMP) task in order to evaluate if and how face perception and face memory (specifically, working memory) processes may interact. The FMP task is a match-to-sample paradigm and was designed to allow for independent manipulation of both face memory and face perception task demands/difficulty. For face perception, there were three perceptual difficulty levels. For face memory, there were also three difficulty levels, consisting of three different memory tasks (matching task, unfilled delay task, and filled delay tasks). A long-term face recognition memory test was also completed for all conditions, and inversion effects were evaluated for the unfilled delay task. Over four independent datasets and hundreds of participants, main effects of perception and memory were observed. An interaction between perception and memory difficulty was also observed in 3 of the 4 datasets. The authors conclude that face perception and memory tasks are fundamentally independent but do interact when there is working memory interference that impacts holistic processing.

Overall, this study addressed an interesting question, and the manuscript was well written, but I have one major concern about the factors driving the key interaction and several smaller concerns.

Major points

Restricted range: Performance in the Matching and the Unfilled Delay conditions was reduced as perceptual difficulty increased. In contrast, perceptual difficulty had smaller (or no) effects on the Filled Delay Emotions condition. The differences in the effect of perceptual difficulty on the three conditions produced the interaction, which most of the manuscript's inferences depend on. However, performance in the easiest perceptual condition in the Filled Delay Emotions condition was substantially lower than performance in the easiest perceptual conditions in the other two conditions, and the proportion correct in the easiest Filled Delay Emotions condition was close enough to chance performance that restrictions of range may have contributed to or even been responsible for the smaller effects of perceptual difficulty. To convincingly demonstrate an interaction driven by the task differences rather than restrictions of range in the FDE condition, I'd want to see performance in the easiest FD condition that was comparable to the easiest Unfilled Delay condition (roughly 0.80). Similarly, restrictions of range may be contributing to the reduced effect of perceptual difficulty in the inverted condition.

Use of identical test and target images: Participants were required to identify a test image that was identical to the target image. A better design would require identification of the test image that displayed a different view of the target face (e.g., viewpoint, lighting)(Hay & Young, 1986). That way, participants could not simply identify the target image from lower-level information, so performance would require greater reliance on the higher-level processes that the study is attempting to address.

Figures: It is tough to see the individual participant data for Figures 2-4 on the blue bars in the graphs, which masks the variability of this data. I suggest using a different color here for the bars or the participant data so that this information is available to the reader. Additionally, to make the figures more helpful to the reader, I would try to use colors that are more different from one another for individuals with color blindness.

Minor points

I found the absence of an effect of the math problems surprising and interesting.

I was curious if there are any learning effects during these tasks. For example, are the differences reported here more or less prevalent in earlier blocks of a task when compared to later blocks of a task? If so, what could this tell us about the

interaction and distinction between face perception and memory?

Do participants complete any practice trials before starting the task? Please clarify these details in the manuscript.

Can the authors expand on why they think the morphed and unmorphed faces used as the cue stimulus impacted overall memory accuracy?

Similar to the above point, can the authors provide possible explanations on why they think Dataset 2 didn't have significant long-term memory results for the different conditions?

Would the authors expect any face presentation during the filled delay (FD) task to have a similar interference effect/impact on accuracy? That is, is an active face task (such as determining the expression of a face) required to obtain the same results, or is it just the presence of the face? This may speak to the portion of the discussion section that discusses cognitive load theory and the Working Memory Model.

I'm less confident than the authors that face identity recognition is the defining goal of the face system. I'm not sure that there is a single defining goal of the face system, and it wouldn't surprise me if face processing was well developed in our remote ancestors for tasks such as face detection, gaze perception, sex discrimination, etc. before there was selection for face identity recognition.

Reviewer #2

(Remarks to the Author)

This study was designed to characterize the relationship between perceptual and mnemonic processes related face stimuli. The experiments are cleverly designed to manipulate difficulty along two orthogonal dimensions, i.e., perceptual difficulty and task difficulty. Across datasets, the authors report significant main effects of perceptual difficulty and task difficulty. Their central claim hinges on the logical assumption that if face perception and face memory are at least partially dependent processes then there should be a statistical interaction between the factors perceptual difficulty and task difficulty. This result was indeed obtained. Further probing of this interaction suggested that it was capturing dependence that was unique to the filled distractor task. This effect was not present when math problems were used in place of face distractors. Moreover, the authors used a face inversion manipulation to demonstrate that increased perceptual difficulty drove a shift toward feature-based processing. Overall, the paper is well written and the authors have done a wonderful job of attempting to falsify their hypotheses. I am of the opinion that these data are of high value for the typically orthogonal fields focused on perception and memory. However, I do have a few concerns that, if addressed properly, would strengthen the claims being made and potential impact.

1. The difficult memory task includes either a face distractor or an arithmetic problem distractor. The authors argue that domain-specific interference reveals a unique interaction between perception and memory. However, the strongest version of this claim requires a non-face visual distractor that is matched to faces at the level of visual complexity. Given extensive neuropsychological, neuroimaging, and intracranial EEG evidence pointing to specialization within the ventral temporal cortex for faces, it seems that a strong test of the hypothesis would include discrete objects matched to faces for real-world size as distractors.

2. Related to my previous point, I am curious to know how critical it is that faces be presented during the delay stage. The authors frame their results in terms of domain-specific cognitive load, but I wonder whether (again) the results might alternatively reflect an accumulation of interference from face stimuli that can be observed irrespective of when the stimuli are presented. For example, would the results be the same if participants were to complete the emotion recognition task using all distracting faces prior to completing the unfilled FMP task?

3. It's possible that I've misunderstood the face morphing procedure, but it seems that all faces of a particular gender were blended with just one face. If this is the case, there should be considerable overlap in individual features across trials. How do the authors account for the potential trial-to-trial interference that might emerge from the accumulation of feature-based representations?

4. The authors refer to the matter of entropy multiple times throughout the manuscript. However, this concept or its impact on perception/memory is not well described. Some elaboration is necessary.

5. A minor concern pertains to potential differences in screen resolution/size across participants. Given the nature of the task, a participant using their phone may perform rather differently than a participant using a large display. Were relevant data collected?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

My main concern with the initial submission was that restrictions of range were contributing to the interaction that the key inferences depended on. To address this concern, the authors restricted their analysis to the participants who scored in the top half. This approach brought the scores up to a level where I am no longer concerned about floor effects, so I'm now convinced by the interaction.

Everything else looks good to me.

Reviewer #2

(Remarks to the Author)

The authors have done a lovely job of addressing my concerns. My sense is that the data now better support the claims and that this paper will be of broad interest.

Reviewer #3

(Remarks to the Author)

I very much liked this paper; it is well motivated, well written, and the study is rigorously conducted. It asks an important question about the interaction of perception and memory whose answer has important implications for the mechanisms of cognition. However, I have a few questions about some of the claims made and some of the modelling methods, and in one case I have a suggestion for a reanalysis of some of the data. I will provide my comments in the order that they arise from the manuscript.

line 64 – prosopagnosia constitutes “strong evidence for separation between face perception and face memory” – I was surprised by (and skeptical of) this claim. This claim seems to imply that prosopagnosia is a deficit solely in face memory, since it is subsequently described as a deficit in “face recognition.” Surely, in this disorder, face perception (e.g., ability to discriminate two faces that are simultaneously present) is also impaired; I have seen evidence of this in multiple accounts of prosopagnosia. The authors mention another disorder – prosopamnesia – that may more clearly differentiate face memory from face perception, but to the best of my knowledge this is an exceedingly rare disorder with scant documentation and empirical investigation. Even if prosopamnesia provides good evidence for a separation between memory and perception for faces, it is still unclear to me why we should consider *prosopagnosia itself* evidence for such separation. I think this claim should be modified.

Related to the above point is a more general issue: It is not clear what the authors are using as their definition of memory versus perception. Does “face recognition” count as perception or as memory? It would be reasonable to operationalize the two functions (perception vs memory) as specific tasks, for the purposes of this paper. To some extent this is what the authors do: face memory is taxed by using a delay between study and test and then increasing the memory load by filling the delay with potentially interfering activities; face perception is taxed by having a simultaneous discrimination of which the difficulty is manipulated by making target and lure stimuli more visually similar. But I think it should be explicitly stated that these task manipulations are being used as a functional definition of the terms “memory” and “perception” and, importantly, it should be acknowledged that it's unclear how real-world face-processing functions such as “face recognition” map onto this dichotomy – face recognition seems to involve both.

I think many in the field will find the above operationalization of memory and perception to be reasonable. However, there's at least one theoretical position that disagrees with this. In a series of papers – Cowell et al. (2006), Bartko et al. (2007), Bartko et al. (2010) – a connectionist model of object recognition memory suggests that the *same mechanism* (namely, feature-level interference) is responsible for deficits in perirhinal cortex-lesioned individuals in the face of increased memory load *and* increased perceptual difficulty. In other words, whether you make a memory discrimination task harder by making the stimuli more perceptually similar, by increasing the retention delay, or by filling the delay with interfering material, the same mechanism can account for deficits following perirhinal lesions. So, there is a theoretical interpretation of the present study's findings that says the two manipulations (perceptual vs mnemonic) might be taxing the same resources. However, that theory would still have to explain why, in this study, the manipulations have an additive effect across the lower levels of load/difficulty and a sub-additive effect at the highest levels.

line 81: “independence as a likely universal cognitive principle across face and object recognition.” This seems to imply that object recognition (as in, identifying what an object is) is separate from object memory (e.g., including object recognition memory). As noted above, for object recognition specifically (rather than face recognition), there is a line of work including connectionist modeling and a number of empirical papers in rats and humans suggesting strongly that the mechanisms of object perception and memory are not separable (Cowell et al., 2006, J. Neurosci; Bartko et al., 2007, J. Neurosci; Bartko et al., 2010, Neuropsychologia; Barense et al., 2012 Neuron; Yeung et al., 2013; JEP: General).

line 200 – I understand the logic of using Bayesian mixed-modeling to explore the nature of the interactions, but I have some questions about how the models were set up. Specifically, the two models under comparison were:

1. Interaction model: $y \sim \text{condition} * \text{difficulty} + (1 | \text{participant})$
2. No interaction model: $y \sim \text{condition} + \text{difficulty} + (1 | \text{participant})$

And the goal was to test whether the data justify an interaction term (as indicated in the frequentist ANOVAs) as well as to probe which specific pairings of the levels of each factor were driving the interaction. My question concerns the interaction model. I was expecting to see a nested set (or pair) of models, in which the interaction model contained both terms for the

main effects AND the interaction term; then a comparison would evaluate if there is evidence in the form of better model fit to justify including the interaction term. By using a model that did not include the lower-order main effects, isn't it possible that the interpretation of the interaction term could change dramatically, because it becomes partly a main effect in disguise. Put another way, could that single interaction term inadvertently absorb the main effects' variance, giving misleading inferences? I am not overly concerned about this possibility because the authors also performed frequentist ANOVAs which revealed significant interactions, but they should justify and explain their choice of interaction model, to make the rationale clear.

line 230 – “The above analysis gave us all interactions” – “gave us” implies possibly that they were all significant (which was not the case); better to say “the above analysis assessed all interactions” just to be clear.

Figure 3d and 3e – the figure or its caption should clearly state that the morphed image is used as baseline. Otherwise, it's a little confusing.

line 416: It is stated that “there is no difference between models with and without an interaction term” but then later in the same paragraph various significant interactions are reported, e.g., “The previously observed significant interaction between the FD-e condition, and perceptual difficulty in the UD and M conditions was observed in this analysis as well”. This all refers to Supplementary Figure 10. Panel a of that figure confirms that there was no benefit to including the interaction term in the model. Then panel b appears to show the posterior distributions for individual interaction terms (presumably from the model “with interaction”), which implies that something is being made of the interaction term even though, per panel a, we should not be paying any attention to the interaction term because it does not improve model fit? Please explain and justify this figure and clarify whether we should make anything of these interactions.

line 456: “As perceptual difficulty increases, the probe faces become more similar. This might prompt participants to shift from a more holistic face processing approach to a more feature-based one.” This hypothesis is the opposite of what some theories would predict e.g., simulations from the model of Bussey and Saksida (2003; Eur J Neurosci) suggest that as a discrimination problem becomes more perceptually difficult through greater morph overlap, there should be a greater reliance on holistic, “whole-is-greater-than-sum-of-parts” representations. That said, I can also see an argument for needing to try to select a single distinguishing feature as the overlap reaches ceiling levels (e.g., the common experience of telling apart identical twins by a birthmark or haircut), so which strategy dominates may be a question of exactly how much overlap exists between the morphed discriminanda. The authors marshal evidence for their hypothesis in the form of the face-inversion effect (indexing holistic processing), which appears to reduce as morph overlap increases. However, I'm concerned that this analysis does not take into account the non-interval scale of the accuracy measure (proportion scores). That is, % correct is a non-linear measure, in this case bounded at 50% and 100% (chance is 50%). The % correct measure likely follows a cumulative normal-like curve as a function of the underlying (latent) discriminability signal that we care about. This means that a given change in the measured variable (% correct) at low values near 50% corresponds to a larger change in the underlying discriminability signal than the same observed change in the % correct variable near 75%. Thus, subtracting inverted % correct from upright % correct is not a theoretically sound thing to do: the authors should instead transform the scores using an inverse normal function to put them on an interval scale before subtracting one from another. I am concerned that this would show no difference in the size of the inversion effect between the different levels of perceptual difficulty. One further question: the inversion effect is often measured by looking at reaction time differences between upright and inverted – did the authors try this?

Supplementary (line 960): “Standard deviation of RT was less than 2 SD below the standard deviation of the group”. Do you mean MORE than 2 SD below?

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

The authors have done an excellent job responding to all of my concerns, and I was glad to see that most of them stemmed from minor ambiguities in the writing or equations (e.g., the Bayesian model with the interaction *did* in fact include the lower-order terms). I think this round of clarifications has improved the manuscript, which now makes a fine contribution to the literature.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We would like to thank the reviewers for their thoughtful, constructive, and encouraging feedback on our manuscript. We are particularly grateful for the overall enthusiasm expressed for our study and for the recognition of the novelty and rigour of our approach to investigating the interaction between face perception and memory.

Revising the manuscript and addressing the reviewers' comments led us to substantially deepen and extend our work. In particular, we have now conducted and included new analysis and experiments specifically designed to address key points raised in the reviews. We introduced a new visual interference experiment using complex, non-face stimuli (food plates) to directly test whether visual complexity alone, without face-specific processing demands, could explain the critical interaction observed in our original data. We collected 83 participants for this experiment and the results of this new dataset support our original conclusion: the interaction between face perception and memory is not simply due to visual interference or task difficulty, but reflects domain-specific processing related to faces.

In addition to this major addition, we also performed further control analyses (e.g., addressing floor effects, learning and fatigue, or task order) and clarified multiple aspects of our methods, analysis, and interpretation throughout the manuscript. This resulted in an additional results section with one new main figure (Figure 5), five new supplementary figures (Supplementary Figures 1, 5, 6, 10, 13), two Supplementary tables (Supplementary Tables 5, 7) and changes to the discussion and methods. These efforts significantly increase the robustness of our findings and the clarity of our claims.

We are confident that the revised manuscript now offers a more complete and compelling account of our findings, and we hope the reviewers and editors will find that their suggestions have been fully and thoroughly addressed.

Reviewer #1:

In this paper, Kadlec and colleagues report results using their Face Memory and Perception (FMP) task in order to evaluate if and how face perception and face memory (specifically, working memory) processes may interact. The FMP task is a match-to-sample paradigm and was designed to allow for independent manipulation of both face memory and face perception task demands/difficulty. For face perception, there were three perceptual difficulty levels. For face memory, there were also three difficulty levels, consisting of three different memory tasks (matching task, unfilled delay task, and filled delay tasks). A long-term face recognition memory test was also completed for all conditions, and inversion effects were evaluated for the unfilled delay task. Over four independent datasets and hundreds of participants, main effects of perception and memory were observed. An interaction between perception and memory difficulty was also observed in 3 of the 4 datasets. The authors conclude that face perception and memory tasks are fundamentally independent but do interact when there is working memory interference that impacts holistic processing.

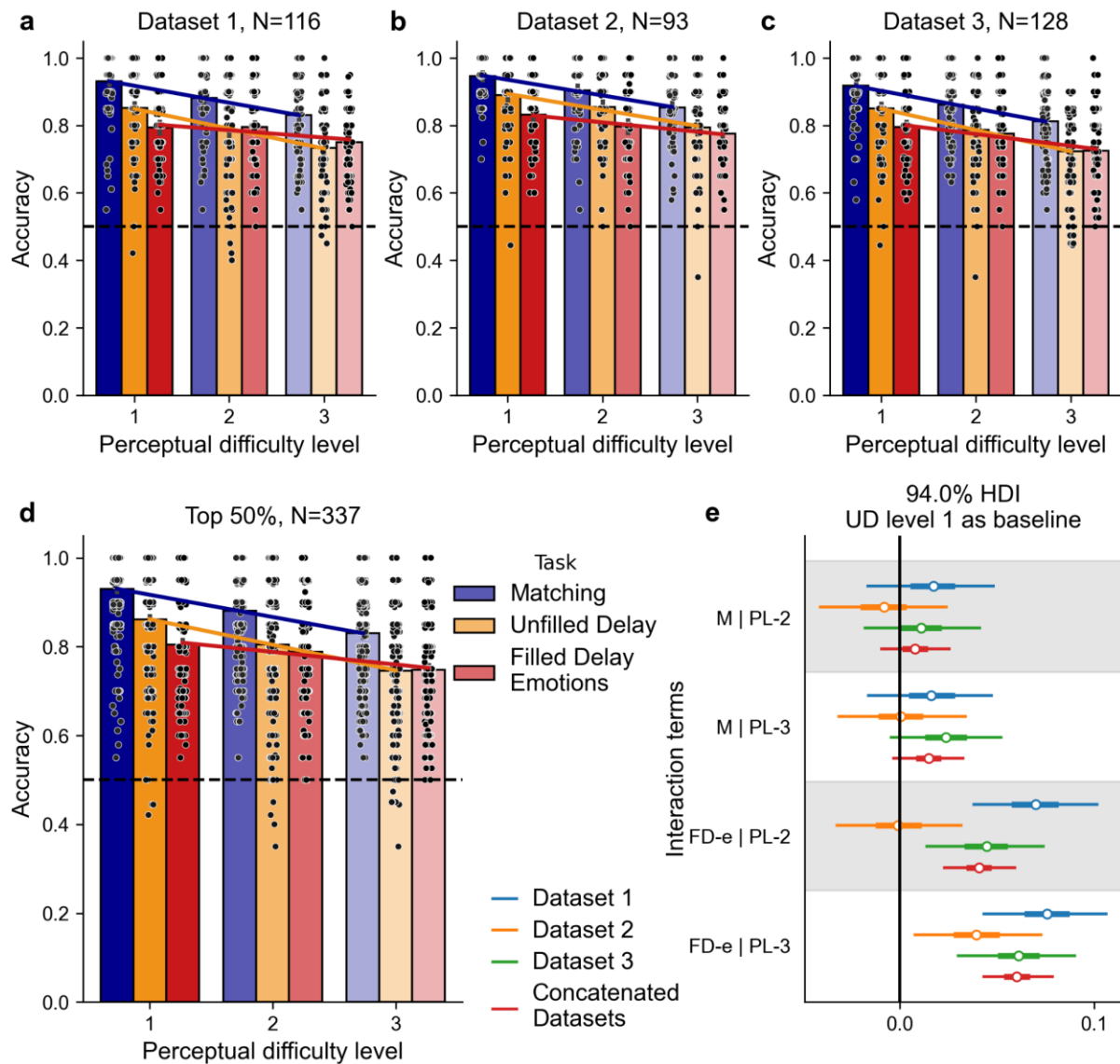
Overall, this study addressed an interesting question, and the manuscript was well written, but I have one major concern about the factors driving the key interaction and several smaller concerns.

Major points

Restricted range: Performance in the Matching and the Unfilled Delay conditions was reduced as perceptual difficulty increased. In contrast, perceptual difficulty had smaller (or no) effects on the Filled Delay Emotions condition. The differences in the effect of perceptual difficulty on the three conditions produced the interaction, which most of the manuscript's inferences depend on. However, performance in the easiest perceptual condition in the Filled Delay Emotions condition was substantially lower than performance in the easiest perceptual conditions in the other two conditions, and the proportion correct in the easiest Filled Delay Emotions condition was close enough to chance performance that restrictions of range may have contributed to or even been responsible for the smaller effects of perceptual difficulty. To convincingly demonstrate an interaction driven by the task differences rather than restrictions of range in the FDE condition, I'd want to see performance in the easiest FD condition that was comparable to the easiest Unfilled Delay condition (roughly 0.80). Similarly, restrictions of range may be contributing to the reduced effect of perceptual difficulty in the inverted condition.

We thank the reviewer for a thorough review of our paper and for suggesting this important control. To address this, we selected the top 50% of participants on the FD-e task and reran the analysis with this cohort. We decided to select participants based on performance on the hardest condition, FD-e, as that was a condition that could suffer from floor effects (restricted range) and chose the top 50% to match the overall accuracy, as the reviewer suggested. The results corroborate our previous findings and show no interaction between UD and M and a significant interaction between UD and FD-e for all datasets and levels, except for PL-2 in Dataset 2 as before. See the figure below that we added to the manuscript as Supplementary Figure 4 (shown below). The following paragraph was added to the Results on p. 8:

To validate that the observed interaction in the FD-e condition does not stem from a floor effect or restricted range since this is the most difficult condition, we selected the top 50% of participants on the FD-e condition and reran the analysis. Performance on FD-e level 1 in this subsample of participants (0.80 ± 0.10 , proportion correct $\pm$ standard deviation) was comparable to that of the full sample on UD level 1 (0.81 ± 0.14). The results, shown in Supplementary Figure 5, corroborate results from the full sample – independence between M and UD on all levels and interaction between UD and FD-e on all levels and in all datasets except for Dataset 2 and perceptual level 2 (as before).



Use of identical test and target images: Participants were required to identify a test image that was identical to the target image. A better design would require identification of the test image that displayed a different view of the target face (e.g., viewpoint, lighting)(Hay & Young, 1986). That way, participants could not simply identify the target image from lower-level information, so performance would require greater reliance on the higher-level processes that the study is attempting to address.

We appreciate the reviewer's thoughtful suggestion. However, our goal in this study was not to target higher-level or view-invariant face processing mechanisms, but rather to isolate and examine face *perception* in its most basic form, as distinct from face *memory*. Using the same image was a deliberate methodological decision designed to reduce reliance on memory processes and minimise confounding factors.

While using a different view of the same identity could indeed recruit higher-level representations (e.g., view-invariant face identity encoding), it would also introduce a

substantial memory component. Participants would have to recall and match identity across transformations, thereby engaging face memory and potentially conflating perception and memory processes, making it more difficult to disentangle the two. We now clarify this motivation in the Discussion of the revised manuscript (p. 23)

We should also note that we explicitly chose to focus on morphing manipulations for studying perception, rather than different views or images of the same face (as done, for instance, in the Glasgow Face Matching Task). While other types of manipulation would allow us to study higher-order identity representations, it is difficult to disentangle the perceptual processes from other inherent memory processes involved in these tasks, such as those that might be required to match identity across different representations of the same face.

Figures: It is tough to see the individual participant data for Figures 2-4 on the blue bars in the graphs, which masks the variability of this data. I suggest using a different color here for the bars or the participant data so that this information is available to the reader. Additionally, to make the figures more helpful to the reader, I would try to use colors that are more different from one another for individuals with color blindness.

We thank the reviewer for this very important comment. In response, we have adjusted the colour of the individual participant data points to enhance contrast and improve visibility. From the initial submission, we aimed to select colours and contrasts that would remain clear and distinguishable in both black-and-white print and for readers with colour vision deficiencies.

To ensure accessibility, we verified the appearance of the figures under various types of colour blindness using the following website: <https://www.color-blindness.com/coblis-color-blindness-simulator/>. In general, the order of the labels in legends is the same as in the plots, allowing for straightforward comparison. If the changes are still insufficient, we would greatly appreciate any specific suggestions regarding which colours are problematic and how they might be improved further.

Minor points

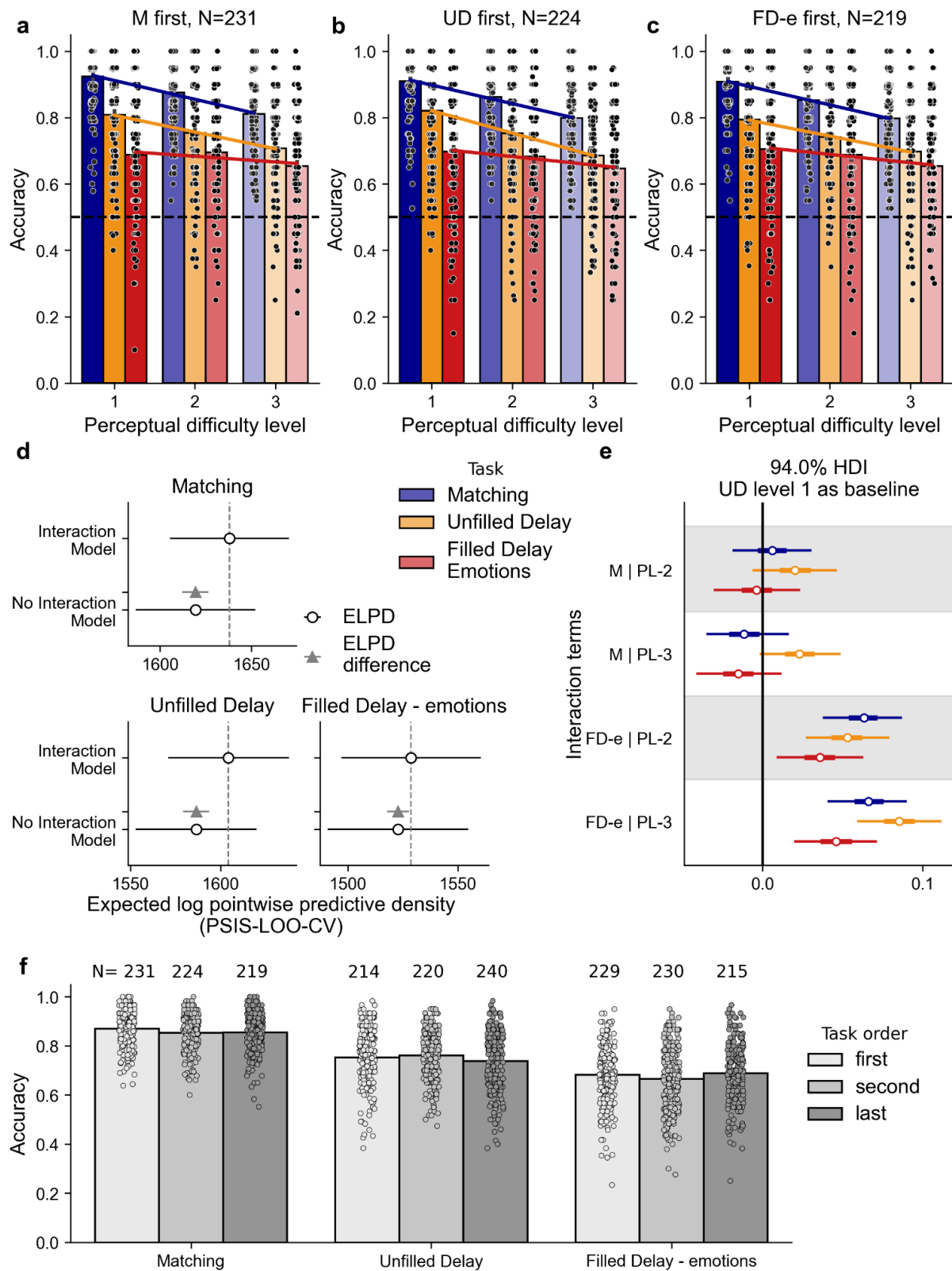
I found the absence of an effect of the math problems surprising and interesting.

This was a surprising finding for us as well. As we write in the text, our initial assumption was that the Math interference would be more difficult and show lower accuracy than the UD task, but higher than the FD-e. This was the first result which prompted us to reconsider our hypothesis. Similarly interesting is the result of FD-v that we have now added to the paper (see section Visual interference does not drive the interaction). Both these results suggest that it is the active and specific interference with face processing, and not just any interference, that is detrimental to performance, causing the observed interaction, probably through a strategy shift, as we discuss.

I was curious if there are any learning effects during these tasks. For example, are the differences reported here more or less prevalent in earlier blocks of a task when compared to later blocks of a task? If so, what could this tell us about the interaction and distinction between face perception and memory?

There is no direct way of testing the effect or presence of learning in our present dataset, as task blocks, each corresponding to a specific memory condition, appear only once per participant. We can, however, indirectly attempt to quantify the possible effect of learning on our results. To do so, we compared participants who completed the M/UD/FD-e tasks as their first task vs participants who completed them as their second or last task, as task order was counterbalanced across participants. This analysis showed no evidence of learning effects, with similar interaction effects and change in performance between memory tasks regardless of task order (see new Supplementary Figure 5 below). We report this now in the Results in the manuscript on p. 8:

Even though the order of the conditions was counterbalanced across participants, we wanted to test if there were any learning or exposure effects, and if these might explain any part of our results. We therefore repeated the main analyses on three subsets of participants: those who began with the M condition, the UD condition, or the FD-e condition, respectively. As shown in Supplementary Figure 6a–c, the results remain consistent across these subsets. Model comparisons and interaction patterns are preserved (Supplementary Fig. 6d, e), with no qualitative changes. We further examined whether condition order influenced performance by comparing accuracy across participants who completed the M/UD/FD-e conditions first, second, or last. While a significant order effect was found for the M condition ($F(2, 671) = 3.20$, $p = 0.041$; $\eta_p^2 = 0.0094$), post-hoc analysis revealed this was driven by slightly higher accuracy when the condition appeared first compared to second (Post-hoc Tukey test: first vs second, $T=2.35$, $p=0.049$, Cohen's $d=0.23$), but not compared to last (Post-hoc Tukey test: first vs last, $T=1.96$, $p=0.12$, Cohen's $d=0.19$; last vs second, $T=0.42$, $p=0.90$, Cohen's $d=0.04$, see SI Table 7). Note that this comparison is across different participants, and most likely reflects either minor differences between the groups, or a possible small effect of fatigue. In either case, the result is inconsistent with a learning effect (which would predict increased performance with later condition order), accumulation of interference (which would predict decreased performance with later condition order), see Supplementary Figure 6f.



Do participants complete any practice trials before starting the task? Please clarify these details in the manuscript.

We thank the reviewer for pointing out this confusion in the Methods. Yes, participants are first shown illustrative screenshots and a walk-through of a single trial of each task. Following this, they complete a short training phase consisting of four practice trials for the first task they encounter, and two practice trials for each subsequent task. This design balances the need for familiarisation with the tasks while minimising participant fatigue, as the core structure of the tasks remains consistent, with only minor differences between them. All practice trials include feedback. We added the following information to the manuscript to the Experimental design section in Methods (p. 28) to reflect these details:

Before beginning each task in the main experiment, participants were first shown example screenshots and guided through a single trial for that specific task. They then completed practice trials with feedback – four trials for the first task they encountered and two trials for each subsequent task. This structure (walkthrough, practice, task) was repeated for each task. As the overall task format remained consistent, this approach helped ensure understanding while avoiding unnecessary repetition.

Can the authors expand on why they think the morphed and unmorphed faces used as the cue stimulus impacted overall memory accuracy?

The reason for the lack of symmetry in the morphing is a very interesting question for which we currently have no definitive answer. It might potentially be related to the smoothing induced by the morphing, and how this affects participants' strategies. We are currently planning new experiments to explore this.

Similar to the above point, can the authors provide possible explanations on why they think Dataset 2 didn't have significant long-term memory results for the different conditions?

We thank the reviewer for this question. The answer is likely a comparatively smaller N within this dataset in general and the UD task in particular. Figure 7 features a between-subject analysis, and as can be seen from the figure itself, both the significance and the means shift as N increases, showing the sensitivity of the analysis to the number of participants. Even though the results don't reach significance, they follow the same trend as the other datasets, with UD having higher d' and forming thus better and more stable memories of faces than in the M and FD-e tasks. We therefore believe that the lack of statistical significance may be a result of lack of statistical power, rather than a true negative finding, and doesn't change our conclusions. We added the following sentence to clarify it in the main text on p. 18:

Dataset 2 shows the same trend, though the differences do not reach significance, perhaps because Dataset 2 is the smallest of all datasets. This is a between-participant analysis, making it particularly sensitive to the relatively small sample size, with only 43 participants in the UD condition.

Would the authors expect any face presentation during the filled delay (FD) task to have a similar interference effect/impact on accuracy? That is, is an active face task (such as determining the expression of a face) required to obtain the same results, or is it just the

presence of the face? This may speak to the portion of the discussion section that discusses cognitive load theory and the Working Memory Model.

We thank the reviewer for this question. We considered running such an experiment, but without an active task, it would be impossible to ascertain that participants attend to the face, or even look at it (nothing stopping them from looking away). This is why in all our Filled Delay conditions, we include a task that ensures active engagement with the distractor, whether that is recognising an emotion, solving a mathematical equation, or classifying an image (see new added experiment in Fig. 5). Having said that, in light of the new results of the FD-v task, which involves comparable visual interference and a similar active task but has none of the effects of the FD-e task, our guess would be that probably the mere presence of the face, as long as it is attended, is likely to cause similar reductions in performance/interactions. We have added a section in the Discussion on this point (p. 23):

For instance, it is an open question as to whether an active judgement on the emotional faces is necessary to create the observed interaction in the FD-e condition, or whether merely viewing emotional faces in absence of a judgement would be enough to impair performance. This question is difficult to test in online experimental paradigms without eye tracking to confirm that participants attend to the face distractor during the delay period without an overt response. However, given the lack of similar effects in the FD-m and FD-v conditions, which both involve active tasks, it is more likely that these effects result from the nature of the distractor stimuli (faces), rather than the nature of the task.

I'm less confident than the authors that face identity recognition is the defining goal of the face system. I'm not sure that there is a single defining goal of the face system, and it wouldn't surprise me if face processing was well developed in our remote ancestors for tasks such as face detection, gaze perception, sex discrimination, etc. before there was selection for face identity recognition.

This was indeed an unnecessarily strong statement. We have amended it to read "*One of the defining goals*" (p. 2).

Reviewer #2:

This study was designed to characterize the relationship between perceptual and mnemonic processes related face stimuli. The experiments are cleverly designed to manipulate difficulty along two orthogonal dimensions, i.e., perceptual difficulty and task difficulty. Across datasets, the authors report significant main effects of perceptual difficulty and task difficulty. Their central claim hinges on the logical assumption that if face perception and face memory are at least partially dependent processes then there should be a statistical interaction between the factors perceptual difficulty and task difficulty. This result was indeed obtained. Further probing of this interaction suggested that it was capturing dependence that was unique to the filled distractor task. This effect was not present when math problems were used in place of face distractors. Moreover, the authors used a face inversion manipulation to demonstrate that increased perceptual difficulty drove a shift toward feature-

based processing. Overall, the paper is well written and the authors have done a wonderful job of attempting to falsify their hypotheses. I am of the opinion that these data are of high value for the typically orthogonal fields focused on perception and memory. However, I do have a few concerns that, if addressed properly, would strengthen the claims being made and potential impact.

We thank the reviewer for their insightful review. We appreciate positive remarks on the significance of our contributions.

1. The difficult memory task includes either a face distractor or an arithmetic problem distractor. The authors argue that domain-specific interference reveals a unique interaction between perception and memory. However, the strongest version of this claim requires a non-face visual distractor that is matched to faces at the level of visual complexity. Given extensive neuropsychological, neuroimaging, and intracranial EEG evidence pointing to specialization within the ventral temporal cortex for faces, it seems that a strong test of the hypothesis would include discrete objects matched to faces for real-world size as distractors.

We thank the reviewer for this insightful and constructive suggestion. In response, we conducted an additional experiment to directly address this question, for which we created a new Filled Delay condition that uses food plates as complex visual stimuli that are matched in real-world size, complexity and salience to human (emotional) faces.

We first piloted a set of 90 food plates, asking participants (N=50) to classify whether the dish is vegetarian or not (in a similar design to the FD-e and FD-m distractor tasks). We chose 60 stimuli based on the results that showed, on average, high inter-rater agreement (>73%) as distractor stimuli. We then introduced a new task, FD-v, in which participants are first presented with a face (3s), followed by a food plate (4s) after which they have to match the initial face (<3s) and answer whether the dish was vegetarian or not (Yes/No, <3s). The design and results of this study are presented in Figure 5 (also displayed below).

We added the following sections to the manuscript (Methods). On p. 26:

Validation and selection of food stimuli for Visual interference condition

This was a pilot experiment to select and assess the difficulty of food distractors for the vegetarian classification task in the FD-v condition described above. In this pilot, 90 images of food plates, 45 vegetarian and 45 non-vegetarian, were presented to participants for 4s (like in the Filled Delay condition of the FMP task) after which participants were asked whether the dish was vegetarian or not (Yes/No). In total, N=51 participants (mean age 35.3 ± 11.0 , 18 females, 32 males, 1 not stated) were recruited and one was rejected due to technical problems during the task. Participants took on average 15 min to complete the task.

The images were subsequently ordered based on participants' accuracy (agreement), and we selected the top 30 non-vegetarian and the top 30 vegetarian dishes as stimuli for the full experiment. These stimuli were correctly identified at least by 73% of participants (average classification accuracy (vegetarian/non-vegetarian) $87.5\% \pm 7\%$ ranging from 73% to 100%).

On p. 27:

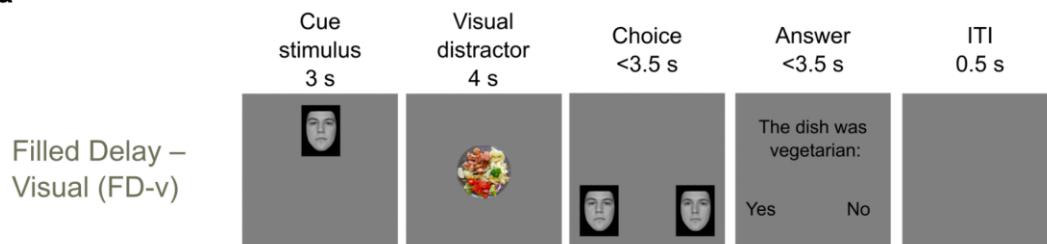
Lastly, to examine whether visual interference alone could account for the interaction observed between UD and FD-e conditions, we introduced a Filled Delay with visual non-face distractor (FD-v) condition using complex, non-face visual stimuli (Dataset 5). The FD-v condition was structurally identical to the FD-e and FD-m conditions, but instead of emotional faces or equations, participants were shown images of food plates during the 4s delay period. Each image was accompanied by a binary categorisation task in which participants had to judge and remember whether the dish was vegetarian or not (Yes/No). As in the FD-e condition, participants responded to this question immediately after selecting which of the two probe stimuli matched the initially presented face. As in FD-m, the location (left or right) of the True/False labels on the screen was randomised across trials.

On p. 29:

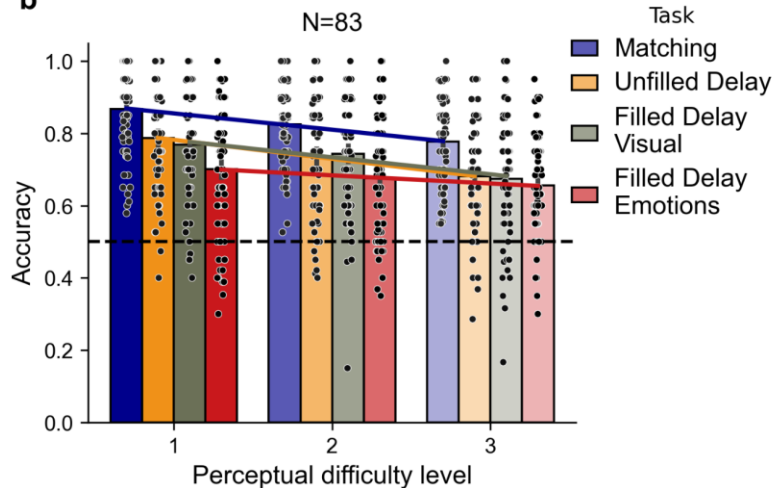
Dataset 5 – Visual interference

This design contained three perceptual difficulty levels and four memory levels -- the regular three described in Figure 1 (M, UD, FD-e) and an additional interference level orthogonal to face processing with visual stimuli matched in size and complexity to faces (Filled Delay with Visual interference, Fig. 5a). The long-term memory task was not included in this experiment. A total of 99 participants started the study, out of which sixteen were excluded during preprocessing due to the same exclusion criteria as Datasets 1-4, leaving N=83 (mean age 32.5 ± 12.5 , 38 females, 45 males).

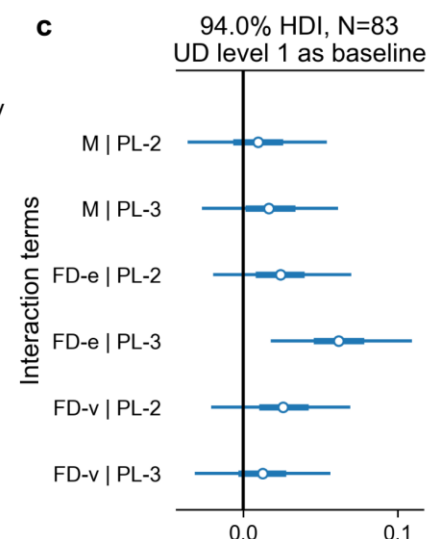
a



b



c



Our results showed that performance in the FD-v condition was indistinguishable from the Unfilled Delay (UD) condition, both at the behavioural level and through Bayesian modelling analyses. This finding suggests that visual interference alone, even with stimuli matched in complexity and salience to faces, is not sufficient to generate the observed interaction between perception and memory. These new results further strengthen our conclusion that the interaction between face perception and face memory is not a general consequence of increased visual complexity or cognitive load, but instead reflects a domain-specific effect that requires face-related processing. The text on pages 13-15 now reads:

Non-specific visual interference does not drive the interaction

Another possible explanation could be that it is the visual interference – processing highly complex visual stimuli (emotional faces), rather than the cognitive nature of the task, that produces the interaction. The FD-m condition lacks visual complexity, and therefore cannot address this possibility. To test this hypothesis more directly, we introduced a new condition, FD-v (Filled Delay with Visual non-face distractor), in which participants viewed images of food plates during the delay period and had to classify whether the food was vegetarian or not, a question they were asked following the probes, as in the FD-e and FD-m conditions (Fig. 5a). Food plates were selected as visually complex and socially salient stimuli, approximately matched to faces in real-world size and perceptual richness. The stimuli were piloted to ensure that categorisation difficulty was comparable to the emotional judgement task used in the FD-e condition (see Methods).

Behavioural results showed that performance in the FD-v condition was indistinguishable from that in the UD condition across all levels of perceptual difficulty (Fig. 5b), a finding further supported by Bayesian modelling, which revealed no significant difference between FD-v and UD (Supplementary Fig. 10). This pattern closely mirrors the findings from the previous section (Fig. 4) and applies to all aspects of the results described there. There is no difference between models with and without an interaction term in Bayesian modelling (Supplementary Fig. 10). Looking at the posterior distributions, we found that the only significant interaction was between UD and FD-e at perceptual level 3 (Fig. 5c, Supplementary Fig. 10b). We did not observe any interactions for the FD-v condition (Fig. 5c). The previously observed significant interaction between the FD-e condition, and perceptual difficulty in the UD and M conditions was observed in this analysis as well (Fig. 5c, Supplementary Fig. 10b). Similarly to before, no significant interaction was observed when we set the M condition as a baseline except for an interaction of FD-e and M on perceptual level 3, mirroring the interaction of FD-e and UD on perceptual level 3.

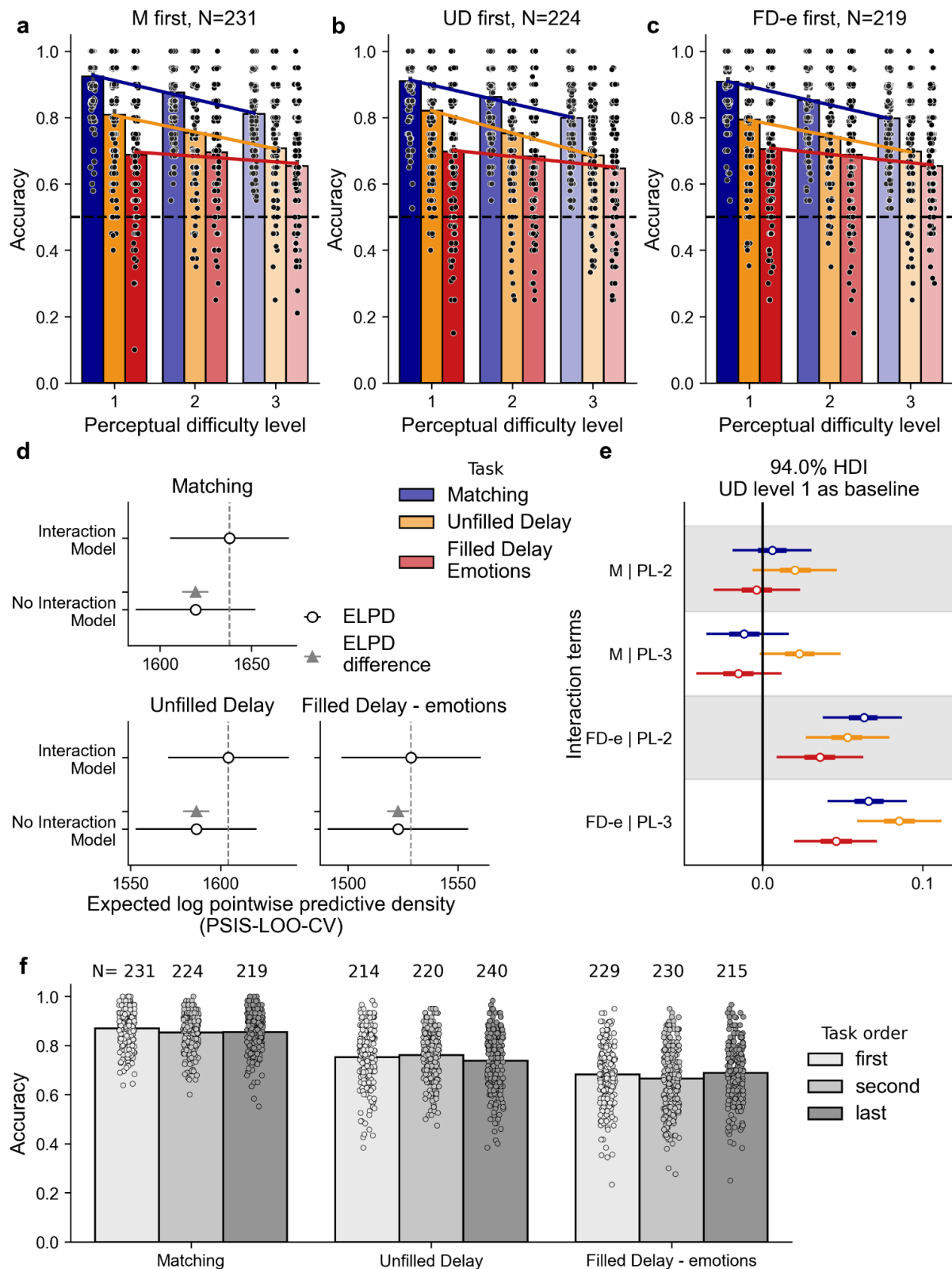
Together, these findings replicate the pattern observed with the FD-m condition and reinforce the conclusion that the interaction between face perception and memory depends on domain-relevant interference, rather than non-specific visual content or general cognitive load.

2. Related to my previous point, I am curious to know how critical it is that faces be presented during the delay stage. The authors frame their results in terms of domain-specific cognitive load, but I wonder whether (again) the results might alternatively reflect an accumulation of interference from face stimuli that can be observed irrespective of when the stimuli are presented. For example, would the results be the same if participants were to

complete the emotion recognition task using all distracting faces prior to completing the unfilled FMP task?

We thank the reviewer for their comment. To address the potential impact of the accumulation of interference from the distractor faces, and from all other faces in other task blocks (as per the reviewer's next comment), as well as to test for the opposite confound of learning, we compared participants who completed the M/UD/FD-e tasks as their first task vs participants who completed them as their second or last task, as task order was counterbalanced across participants. This analysis showed no evidence of an accumulation of interference, certainly not as a driving factor for the FD-e results. We report this now in the Results in the manuscript on p. 8, and in the new Supplementary Figure 6:

Even though the order of the conditions was counterbalanced across participants, we wanted to test if there were any learning or exposure effects, and if these might explain any part of our results. We therefore repeated the main analyses on three subsets of participants: those who began with the M condition, the UD condition, or the FD-e condition, respectively. As shown in Supplementary Figure 6a–c, the results remain consistent across these subsets. Model comparisons and interaction patterns are preserved (Supplementary Fig. 6d, e), with no qualitative changes. We further examined whether condition order influenced performance by comparing accuracy across participants who completed the M/UD/FD-e conditions first, second, or last. While a significant order effect was found for the M condition ($F(2, 671) = 3.20$, $p = 0.041$; $\eta_p^2 = 0.0094$), post-hoc analysis revealed this was driven by slightly higher accuracy when the condition appeared first compared to second (Post-hoc Tukey test: first vs second, $T=2.35$, $p=0.049$, Cohen's $d=0.23$), but not compared to last (Post-hoc Tukey test: first vs last, $T=1.96$, $p=0.12$, Cohen's $d=0.19$; last vs second, $T=0.42$, $p=0.90$, Cohen's $d=0.04$, see SI Table 7). Note that this comparison is across different participants, and most likely reflects either minor differences between the groups, or a possible small effect of fatigue. In either case, the result is inconsistent with a learning effect (which would predict increased performance with later condition order), accumulation of interference (which would predict decreased performance with later condition order), see Supplementary Figure 6f.



3. It's possible that I've misunderstood the face morphing procedure, but it seems that all faces of a particular gender were blended with just one face. If this is the case, there should be considerable overlap in individual features across trials. How do the authors account for

the potential trial-to-trial interference that might emerge from the accumulation of feature-based representations?

We thank the reviewer for pointing out that we did not explicitly address potential confounds that arise from the morphing procedure, which is indeed as understood by the reviewer, all faces were morphed to one of two template faces, one for each gender. We did not believe this to be a problem, as even the highest morphing level (which is only a third of the trials) is 36%, so that different exemplars are still quite distinct. We have added a new Supplementary Figure 13 with examples of several different faces at all morph levels to better illustrate the appearance of the morphed stimuli (all images originate from the RADIATE database). More importantly though, as in the above analysis in response to the previous question, we do not see evidence of either learning (if participants were getting better at spotting the shared features of the morph) or accumulation of interference over time.

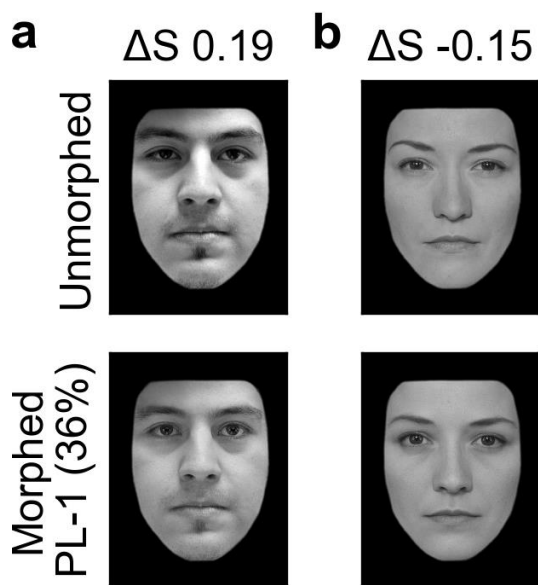


4. The authors refer to the matter of entropy multiple times throughout the manuscript. However, this concept or its impact on perception/memory is not well described. Some elaboration is necessary.

We thank the reviewer for this question. In Dataset 1, we saw a visual difference between the morphed images that we could quantify using image entropy (a measure of the local

variation among neighbouring pixels), prompting us to control for this in Dataset 2 by randomising all images across conditions, and removing the most affected outlier images. We have added a new Supplementary Figure (Supplementary Figure 1) displayed below that shows how change in entropy might manifest differently across different images at the same morph level, as well as updated it in the text (p. 6).

To account for the slightly divergent effect of the morphing procedure on different images, which also resulted in differential changes in image entropy (a measure of the local variation among neighbouring pixels, see Methods for a detailed description and examples in Supplementary Fig. 1), we created a fully randomised stimuli set with entropy-matched stimuli, involving N=153 participants (Dataset 2, Fig. 2b).



5. A minor concern pertains to potential differences in screen resolution/size across participants. Given the nature of the task, a participant using their phone may perform rather differently than a participant using a large display. Were relevant data collected?

We thank the reviewer for raising this important point. We took a number of specific steps to minimise variation in screen resolution and device type. During recruitment on Prolific, we explicitly restricted participation to desktop and laptop users only, excluding phones and tablets via platform settings. Additionally, prior to entering the experiment, participants encountered a browser-based device check implemented in PHP. This script automatically screened out mobile and tablet users, displaying the following message:

“This experiment works on **desktops or laptops only**. If you have a phone or tablet, it will not let you further and your screen will freeze. Therefore, please consent that you are accessing from a desktop. Thanks.”

The detection was based on a commonly used method described in this Stack Overflow question: <https://stackoverflow.com/questions/4117555/simplest-way-to-detect-a-mobile-device-in-php>

We also recorded relevant technical information, including browser type, screen resolution, operating system and more. We acknowledge that even among desktop users, variability in screen size, viewing distance, and visual angle remains a limitation of online data collection.

We clarified this point in the Methods on p. 28:

To minimise variability in device type and display conditions, only participants using desktop or laptop computers were eligible to take part in the studies. Prior to the task, participants were informed that the experiment would not run on phones or tablets. This restriction was enforced through Prolific's prescreening options and a PHP-based device check that blocked access for mobile or tablet users. We also recorded technical data including screen resolution, browser, and operating system. While some variation in screen size and viewing conditions may still be present, this is an inherent limitation of online experiments.

Response to reviewers, November 2025

We would like to thank the reviewers for their thoughtful, constructive, and encouraging feedback on our manuscript.

Reviewer #1:

My main concern with the initial submission was that restrictions of range were contributing to the interaction that the key inferences depended on. To address this concern, the authors restricted their analysis to the participants who scored in the top half. This approach brought the scores up to a level where I am no longer concerned about floor effects, so I'm now convinced by the interaction.

Everything else looks good to me.

Thank you for the positive assessment. We're grateful for your careful review.

Reviewer #2:

The authors have done a lovely job of addressing my concerns. My sense is that the data now better support the claims and that this paper will be of broad interest.

Thank you for your feedback. We're glad the revisions address your concerns and that the data now better support the claims. We appreciate your view that the paper will be of broad interest.

Reviewer #3

I very much liked this paper; it is well motivated, well written, and the study is rigorously conducted. It asks an important question about the interaction of perception and memory whose answer has important implications for the mechanisms of cognition. However, I have a few questions about some of the claims made and some of the modelling methods, and in one case I have a suggestion for a reanalysis of some of the data. I will provide my comments in the order that they arise from the manuscript.

We thank the reviewer for their insightful review. We appreciate their positive comments on the significance of our contributions.

line 64 – prosopagnosia constitutes “strong evidence for separation between face perception and face memory” – I was surprised by (and skeptical of) this claim. This claim seems to imply that prosopagnosia is a deficit solely in face memory, since it is subsequently described as a deficit in “face recognition.” Surely, in this disorder, face perception (e.g., ability to discriminate two faces that are simultaneously present) is also impaired; I have seen evidence of this in multiple accounts of prosopagnosia. The authors mention another disorder – prosopamnesia – that may more clearly differentiate face memory from face

perception, but to the best of my knowledge this is an exceedingly rare disorder with scant documentation and empirical investigation. Even if prosopamnesia provides good evidence for a separation between memory and perception for faces, it is still unclear to me why we should consider *prosopagnosia itself* evidence for such separation. I think this claim should be modified.

We thank the reviewer for this comment and view. We have updated the introduction to more clearly reflect this debate on p. 2:

Additional studies show a differential effect of age on face memory and face perception³⁹.

Others have argued that evidence for separation between face perception and memory comes from patients suffering from developmental or acquired prosopagnosia^{40–45}, a condition involving a selective impairment in face recognition (see reviews^{46,47}). However, it is unclear how specific the deficit in prosopagnosia is to memory per se, rather than to more general mechanisms of face recognition, which involve both perception and memory.

Related to the above point is a more general issue: It is not clear what the authors are using as their definition of memory versus perception. Does “face recognition” count as perception or as memory? It would be reasonable to operationalize the two functions (perception vs memory) as specific tasks, for the purposes of this paper. To some extent this is what the authors do: face memory is taxed by using a delay between study and test and then increasing the memory load by filling the delay with potentially interfering activities; face perception is taxed by having a simultaneous discrimination of which the difficulty is manipulated by making target and lure stimuli more visually similar. But I think it should be explicitly stated that these task manipulations are being used as a functional definition of the terms “memory” and “perception” and, importantly, it should be acknowledged that it’s unclear how real-world face-processing functions such as “face recognition” map onto this dichotomy – face recognition seems to involve both.

We thank the reviewer for this remark. We write in the introduction on p. 1 that “*In the real world, these processes are inherently linked and most natural behaviours, such as face recognition, will involve elements of both.*” and we completely agree with the reviewer that it is unclear how those two functions (face perception and face memory) map onto real-world processing functions such as face recognition.

We also agree that a clear, explicit definition of how we use those terms is needed. Therefore, we added the following text into the introduction on p. 3:

It should be noted here that for the purpose of this paper, we operationalise these task manipulations as our functional definitions of “perception” and “memory”, though these cover only particular aspects of what these terms might encompass.

To make both points more explicit, we now also reiterate this in the discussion on p. 23:

In general, our design is limited to studying specific aspects of perception and memory. In more naturalistic settings, it remains uncertain how the ability to recognise a person in real-

world conditions, with their inherent variability in viewpoint, lighting, expression, and context, would relate to this dichotomy.

I think many in the field will find the above operationalization of memory and perception to be reasonable. However, there's at least one theoretical position that disagrees with this. In a series of papers – Cowell et al. (2006), Bartko et al. (2007), Bartko et al. (2010) – a connectionist model of object recognition memory suggests that the *same mechanism* (namely, feature-level interference) is responsible for deficits in perirhinal cortex-lesioned individuals in the face of increased memory load *and* increased perceptual difficulty. In other words, whether you make a memory discrimination task harder by making the stimuli more perceptually similar, by increasing the retention delay, or by filling the delay with interfering material, the same mechanism can account for deficits following perirhinal lesions. So, there is a theoretical interpretation of the present study's findings that says the two manipulations (perceptual vs mnemonic) might be taxing the same resources. However, that theory would still have to explain why, in this study, the manipulations have an additive effect across the lower levels of load/difficulty and a sub-additive effect at the highest levels.

We thank the reviewer for bringing this literature to our attention. We have included it now in the introduction on p. 2:

In the same vein, the connectionist model of object recognition memory^{32–34}, together with a number of empirical papers in rats and humans^{35,36}, suggests that the mechanisms of object perception and memory are overlapping. These studies show that the same mechanism, namely feature-level interference, can account for deficits following perirhinal lesions, although it is unclear whether this applies to the special case of faces as well as to general objects.

and explain in the discussion on p. 21 why it is an unlikely explanation for our results:

Similarly, the connectionist theory^{32–34} would also predict a stronger hyperadditive effect as memory difficulty increases, as it posits that increasing either memory or perceptual load would be taxing the same resources.

line 81: “independence as a likely universal cognitive principle across face and object recognition.” This seems to imply that object recognition (as in, identifying what an object is) is separate from object memory (e.g., including object recognition memory). As noted above, for object recognition specifically (rather than face recognition), there is a line of work including connectionist modeling and a number of empirical papers in rats and humans suggesting strongly that the mechanisms of object perception and memory are not separable (Cowell et al., 2006, J. Neurosci; Bartko et al., 2007, J. Neurosci; Bartko et al., 2010, Neuropsychologia; Barense et al., 2012 Neuron; Yeung et al., 2013; JEP: General).

We thank the reviewer for this point and the references. We added the connectionist account (see the previous point) and clarified the point also on this line. The new wording on p.3 now says:

highlighting independence as a possible universal cognitive principle across face and object recognition (though note the connectionist theory mentioned above).

line 200 – I understand the logic of using Bayesian mixed-modeling to explore the nature of the interactions, but I have some questions about how the models were set up. Specifically, the two models under comparison were:

1. Interaction model: $y \sim \text{condition} * \text{difficulty} + (1 \mid \text{participant})$
2. No interaction model: $y \sim \text{condition} + \text{difficulty} + (1 \mid \text{participant})$

And the goal was to test whether the data justify an interaction term (as indicated in the frequentist ANOVAs) as well as to probe which specific pairings of the levels of each factor were driving the interaction. My question concerns the interaction model. I was expecting to see a nested set (or pair) of models, in which the interaction model contained both terms for the main effects AND the interaction term; then a comparison would evaluate if there is evidence in the form of better model fit to justify including the interaction term. By using a model that did not include the lower-order main effects, isn't it possible that the interpretation of the interaction term could change dramatically, because it becomes partly a main effect in disguise. Put another way, could that single interaction term inadvertently absorb the main effects' variance, giving misleading inferences? It am not overly concerned about this possibility because the authors also performed frequentist ANOVAs which revealed significant interactions, but they should justify and explain their choice of interaction model, to make the rationale clear.

We thank the reviewer for raising this important point. Our interaction model does include the lower-order terms. In the Bambi specification, the * operator expands to main effects plus the interaction, so the fitted model was

$y \sim \text{condition} + \text{difficulty} + \text{condition}:\text{difficulty} + (1 \mid \text{participant})$

and in the comparison, the second model omitted only the interaction term ($\text{condition}:\text{difficulty}$). We have now stated this explicitly in the Methods on p. 30 to avoid ambiguity and to make clear that the interaction term cannot “absorb” main-effect variance because the model includes the lower-order effects.

line 230 – “The above analysis gave us all interactions” – “gave us” implies possibly that they were all significant (which was not the case); better to say “the above analysis assessed all interactions” just to be clear.

We thank the reviewer for pointing out this confusion. We have amended the text as proposed by the reviewer.

Figure 3d and 3e – the figure or its caption should clearly state that the morphed image is used as baseline. Otherwise, it's a little confusing.

We thank the reviewer for bringing this important point to our attention. It is now specified in both places in the figure caption (previously only in d) and also directly in the figure in panel

titles that the morphed image is used as a baseline. We also updated Supplementary Figure 7 to reflect this.

line 416: It is stated that “there is no difference between models with and without an interaction term” but then later in the same paragraph various significant interactions are reported, e.g., “The previously observed significant interaction between the FD-e condition, and perceptual difficulty in the UD and M conditions was observed in this analysis as well”. This all refers to Supplementary Figure 10. Panel a of that figure confirms that there was no benefit to including the interaction term in the model. Then panel b appears to show the posterior distributions for individual interaction terms (presumably from the model “with interaction”), which implies that something is being made of the interaction term even though, per panel a, we should not be paying any attention to the interaction term because it does not improve model fit? Please explain and justify this figure and clarify whether we should make anything of these interactions.

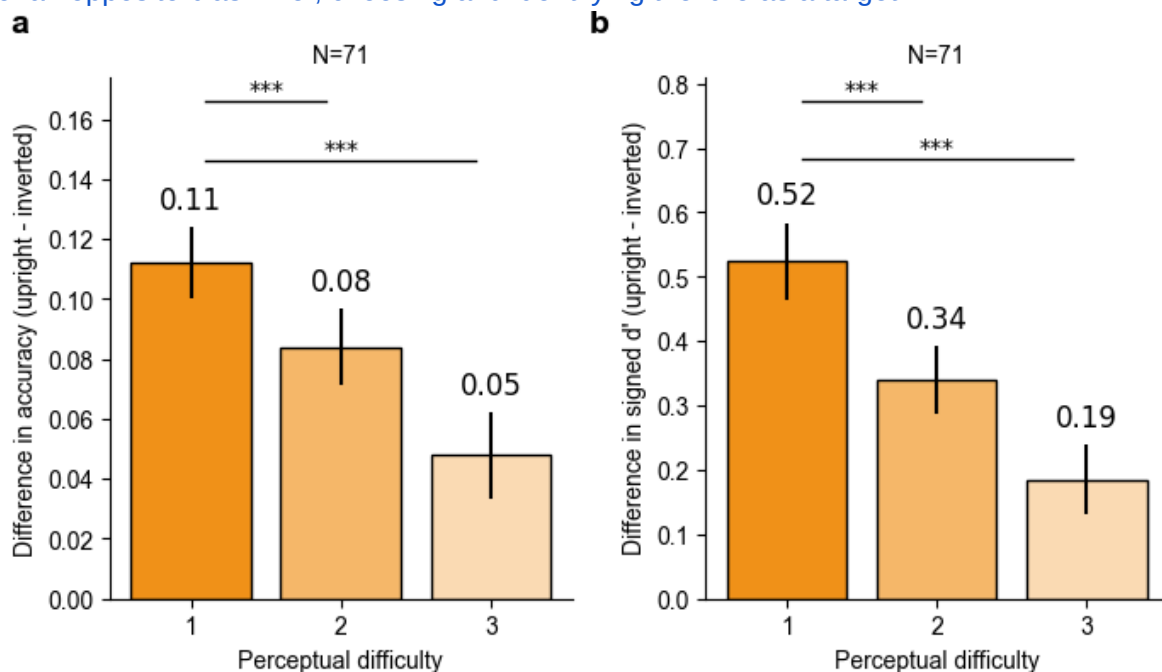
We thank the reviewer for pointing out this confusion. Even though the model with the interaction was not significantly better than the model without, we wanted to test individual interaction terms to see if we can find traces of the same pattern we saw across all our datasets, namely specific interactions between FD-e/PL-3 and M-PL1/UD-PL1. As a standalone result, this would not be meaningful, but as it recapitulates the consistent pattern we saw across all datasets, we felt that it was worth including. We have now rewritten this section on p. 14 to better explain the findings and our interpretation:

There is no significant difference between models with and without an interaction term in Bayesian modelling (Supplementary Fig. 10), and indeed looking at the posterior distributions of all the possible interactions, we find that the only significant interaction was between UD and FD-e at perceptual level 3 (Fig. 5c, also between FD-e perceptual level 3 and M set as baseline in Supplementary Fig. 10b). Though this interaction was not strong enough to make the model with an interaction term significantly better than the one without, these results replicate the pattern observed with the FD-m condition and reinforce the conclusion that the interaction between face perception and memory depends on domain-relevant interference, rather than non-specific visual content or general cognitive load.

line 456: “As perceptual difficulty increases, the probe faces become more similar. This might prompt participants to shift from a more holistic face processing approach to a more feature-based one.” This hypothesis is the opposite of what some theories would predict e.g., simulations from the model of Bussey and Saksida (2003; Eur J Neurosci) suggest that as a discrimination problem becomes more perceptually difficult through greater morph overlap, there should be a greater reliance on holistic, “whole-is-greater-than-sum-of-parts” representations. That said, I can also see an argument for needing to try to select a single distinguishing feature as the overlap reaches ceiling levels (e.g., the common experience of telling apart identical twins by a birthmark or haircut), so which strategy dominates may be a question of exactly how much overlap exists between the morphed discriminanda. The authors marshal evidence for their hypothesis in the form of the face-inversion effect (indexing holistic processing), which appears to reduce as morph overlap increases. However, I’m concerned that this analysis does not take into account the non-interval scale of the accuracy measure (proportion scores). That is, % correct is a non-linear measure, in

this case bounded at 50% and 100% (chance is 50%). The % correct measure likely follows a cumulative normal-like curve as a function of the underlying (latent) discriminability signal that we care about. This means that a given change in the measured variable (% correct) at low values near 50% corresponds to a larger change in the underlying discriminability signal than the same observed change in the % correct variable near 75%. Thus, subtracting inverted % correct from upright % correct is not a theoretically sound thing to do: the authors should instead transform the scores using an inverse normal function to put them on an interval scale before subtracting one from another. I am concerned that this would show no difference in the size of the inversion effect between the different levels of perceptual difficulty. One further question: the inversion effect is often measured by looking at reaction time differences between upright and inverted – did the authors try this?

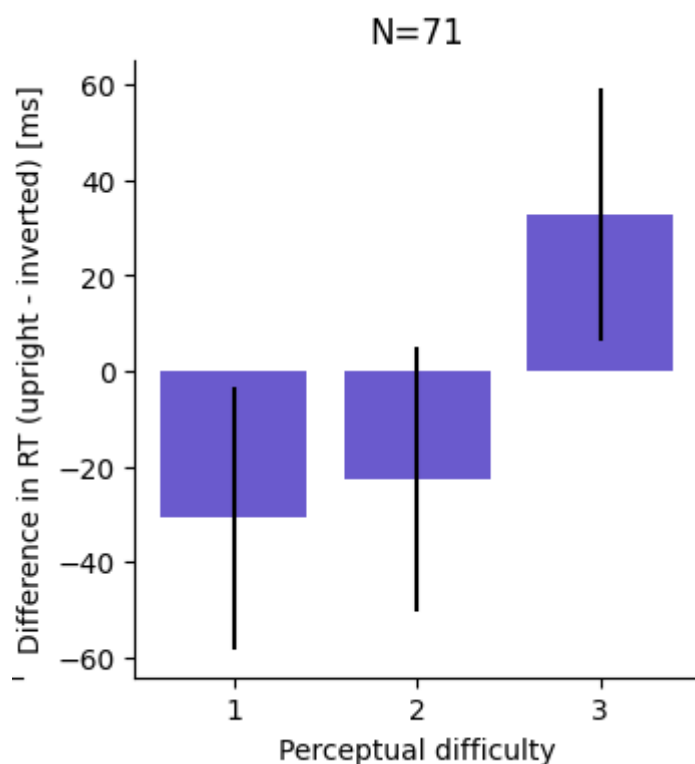
We thank the reviewer for their comment. To address the concern, we first converted accuracy to d' using a formula for 2AFC (Green and Swets, 1966; DeCarlo, 2012), then subtracted d' on upright and inverted trials and performed statistical testing. We tested two cases: 1) where the accuracy is bound to [0.5, 1] (and d' then in [0,1]), and 2) where the accuracy is in [0,1], allowing negative d' values (from -1 to 1), which captures the possibility of an opposite bias – i.e., choosing and identifying the lure as a target.



Using d' instead of accuracy replicated the results from the main text shown in Figure 6 (for easier comparison displayed also here in panel a). The results displayed above in panel b revealed effects that were even stronger and more pronounced than those reported in the original paper: significant main effects of both inversion and perceptual difficulty on d' , as well as a significant interaction between these factors (2-way ANOVA, inversion: $F(1, 420)=65.71$; $p=5.7 \cdot 10^{-15}$; $\eta_p^2=0.14$, perceptual difficulty: $F(2, 420)=23.17$; $p=2.8 \cdot 10^{-10}$; $\eta_p^2=0.10$, interaction: $F(2, 420)=5.14$; $p=6.2 \cdot 10^{-3}$; $\eta_p^2=0.02$). Following post-hoc tests showed a significant difference between the easiest level and all other levels (Post-hoc Tukey test: level 1 vs 2, $T=4.03$, $p=1.9 \cdot 10^{-4}$, Cohen's $d=0.44$; level 1 vs 3, $T=6.19$, $p=4.2 \cdot 10^{-9}$, Cohen's $d=0.72$; level 2 vs 3, $T=2.16$, $p=0.079$, Cohen's $d=0.29$).

These results are not surprising given that the accuracy on inverted trials is very similar (66%, 63%, 62% for PL-1, PL-2, PL-3, respectively), indicating that the main effect of inversion is eliminating the benefit on the easier perceptual levels.

The use of response times (RTs) to demonstrate an inversion effect is one of the pioneering concepts in face research, and we agree that it is a good way to examine the inversion effect. However, the design of the tasks in the original papers (e.g., Haxby, 1999) differs significantly from ours. The classical experiment would have very easily recognisable stimuli, leading to accuracies on any kind of trials (upright, inverted, different categories) to be over 90 % and leaving RT as the alternative way to show a different process is in play (faster RT for stimuli with “an advantage”). The FMP task, on the other hand, is demanding, and the stimuli are much closer perceptually, resulting in a much wider range in the accuracy domain but also a much narrower range in the RT domain. That was the reason we didn’t focus on RT in our manuscript. Nevertheless, we do observe an effect on RT with increasing perceptual difficulty, although not statistically significant, see the figure below.



We added these results to the manuscript. This resulted in a new Supplementary Figure (11) and an updated section in the main text on p. 16.

To account for the non-linear nature of accuracy, we converted accuracy scores to d' using a formula for 2AFC tasks^{75,76} and computed the difference in d' between upright and inverted trials. The resulting differences were even stronger for the easy compared with the harder perceptual levels than what we observed with accuracy (Supplementary Fig. 11a). We next examined response times (RT), a measure that was previously successfully used to demonstrate an inversion effect⁷⁷. Although we saw numerically larger differences in RT for easier perceptual levels, none of the RT differences were significantly different from 0, and

there was no significant difference in RT change between perceptual levels (Supplementary Fig. 11b), which is not unusual given that RT effects often diminish when participants prioritise accuracy or when differences in perceptual discriminability are already well captured by accuracy.

Supplementary (line 960): "Standard deviation of RT was less than 2 SD below the standard deviation of the group". Do you mean MORE than 2 SD below?

We thank the reviewer for this question. The phrasing is indeed ambiguous. We rephrased the sentence to clarify that those participants were an extreme low outlier in RT variability:

Standard deviation of RT was two or more SD below the standard deviation of the group

Response to reviewer(s), February 2026

We would like to thank all three reviewers for their thoughtful, constructive, and encouraging feedback on our manuscript throughout the process. We are grateful for the time, care, and expertise you devoted to the reviews.

Reviewer #3

The authors have done an excellent job responding to all of my concerns, and I was glad to see that most of them stemmed from minor ambiguities in the writing or equations (e.g., the Bayesian model with the interaction *did* in fact include the lower-order terms). I think this round of clarifications has improved the manuscript, which now makes a fine contribution to the literature.

Thank you for your generous and careful evaluation. We appreciate your close reading, and we are pleased that these clarifications improved the manuscript. We sincerely appreciate your view that the revised paper makes a solid contribution to the literature.