Supplemental Materials

This document contains a list of questionnaires collected as part of the research (but not analyzed in this report) and additional results complementing the findings reported in the main text.

1. Questionnaires

Although not analyzed here, each of our studies involved the collection of several questionnaires. Study 1 participants completed the Positive and Negative Affective Schedule (PANAS; Watson, Clark, & Tellegen, 1988), State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1970), Attentional Control Scale (ACS; Judah, Grant, Mills, & Lechner, 2014), Emotion Regulation Questionnaire (ERQ; Gross & John, 2003), Difficulties in emotion regulation scale (DERS; Neumann, van Lier, Gratz, & Koot, 2010), and the Penn State Worry Questionnaire (PSWQ; Kertz, Lee, & Bjorgvinsson, 2014; Meyer, Miller, Metzger, & Borkovec, 1990). The replication study participants completed the PANAS, STAI, ERQ, Mood and Anxiety Symptom Questionnaire (MASQ; Bredemeier et al., 2010), and Behavior Rating Inventory of Executive Function for Adults (BRIEF-A; Roth, Gioia, & Isquith, 2005). Study 2 participants from the student sample completed the PANAS, STAI, ERQ, PSWQ, Beck Depression Inventory (Beck, Erbaugh, Ward, Mock, & Mendelsohn, 1961), and Life Satisfaction Questionnaire (Carlsson & Hamrin, 2002). Study 2 participants from the mother sample (also used for Study 3) completed the PANAS, STAI, ACS, ERQ, DERS, PSWQ, MASQ, BRIEF-A, Behavioral Inhibition/Activation System scale (O’Connor & Forgan, 2007), Mindful Attention Awareness Scale (MacKillop & Anderson, 2007).

2. Additional Eye-Tracking Results

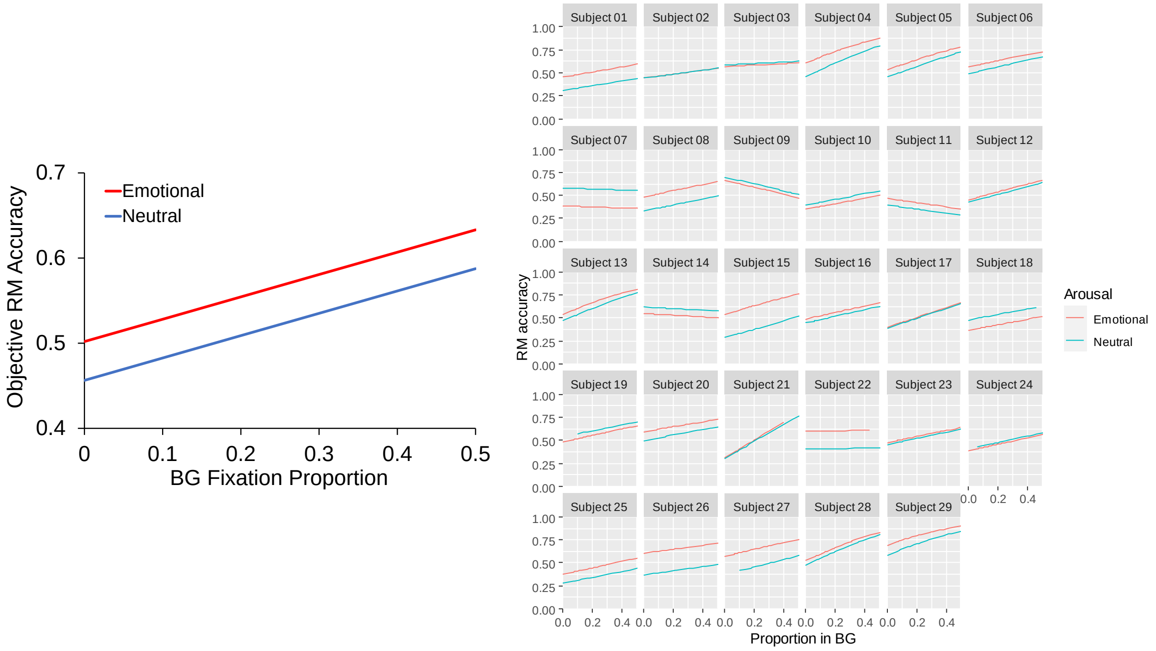
The multilevel regression from the main text can be adapted to predict subjective RM, in place of predicting objective RM. Such a regression, using gaze and emotional rating as predictors, reveals a significant effect of arousal enhancing subjective RM ($\beta = .23$, $p < .001$) and a significant effect of FG gaze also enhancing subjective RM ($\beta = .10$, $p = .047$). Testing for an emotion-gaze-subjective RM mediation revealed a marginal indirect enhancing effect (p = .09). Although a marginally significant effect alone is weak evidence, the results are consistent with the findings from the attentional manipulation designs (see main text), wherein FG focus maximized subjective RM, speaking to the indirect effect's validity. In sum, emotion elicits both a direct enhancing effect on subjective RM and a weaker indirect enhancing effect that is mediated by attentional capture and focus on the FG. This suggests that the emotional enhancement of subjective RM is partially an attentional effect.

3. Participant-Level Regressions

To rule out the possibility that the multilevel regression results from Study 1 were linked to outliers, separate linear regressions were fit for each participant (Figure S1). The enhancing effect is seen across most individuals, indicating that the enhancing effect of arousal and the beneficial effect of focusing on the BG were not driven by outlier participants. Note that the individual slopes are not based on best linear unbiased predictions (BLUPs) extracted from the multilevel regression, but instead reflect unique linear regressions fit for each individual. The BLUP plots show similar patterns but less clearly illustrate the between-subject variability, as one would expect.

Figure S1

Effects of gaze and valence on objective RM



Note. The results are based on Study 1, showing both the multilevel regression data (left) and individual participant data (right).

4. Multivariate fMRI Analyses

To supplement the cluster-based univariate results, multivariate analyses were also performed. A primary advantage of multivariate analytic approaches compared to traditional univariate strategies is that multivariate analyses tend to be more sensitive (Bogdan, Iordan, Shobrook, & Dolcos, 2023; Cremers, Wager, & Yarkoni, 2017; Noble, Mejia, Zalesky, & Scheinost, 2022; Zalesky, Fornito, & Bullmore, 2010). One reason for this is that multivariate analyses, by definition, pool multiple variables and do not assume that the effects of an independent variable are homogeneous. For instance, the procedures here examined the medial temporal lobe (MTL) by analyzing every voxel from a given region of interest (ROI) simultaneously, while also not assuming that every voxel's response is similar. By contrast, the univariate analyses from the main text, analyzed each voxel one-by-one and the cluster-based procedures hinge on neighboring voxels showing similar effects. Analyzing many pieces of information can greatly enhance statistical power if effects are distributed.

However, the potentially enhanced sensitivity and statistical power of multivariate approaches come at the expense of precision during interpretation (Bogdan et al., 2023; Cremers et al., 2017; Noble et al., 2022; Zalesky et al., 2010). For instance, the univariate L amygdala (AMY) cluster showing EmoFG RHit-Hit > Emo Miss-Miss suggests that subsequent memory elicits an effect in the superior and posterior portion of the L AMY (see Figure 7 in the main text). However, this level of spatial precision is not possible using a classifier that uses all AMY voxels as features, which can implicate the AMY's involvement but not pinpoint any specific AMY subregions. Hence, the multivariate analysis is complementary to the main text univariate approach. Specifically, multivariate results can speak for the robustness of the findings based on univariate analyses if the multivariate analyses are also combined with stricter significance thresholds (e.g., a family-wise error correction, which the main text results primarily use).

4.1. Methods

4.1.1. Classifiers

Classifiers were fit that attempted to predict subsequent memory based on voxelwise activation within anatomical ROIs. For instance, based on voxelwise activation within the bilateral AMY, a linear support vector machine was fit to distinguish EmoFG RHit-Hit examples from Emo Miss-Miss examples.[1] The classifier treats each voxel's BOLD response as a separate feature. Above-chance classifier accuracy would indicate a difference-due-to-memory effect in the AMY, which would speak to the validity of the activation effects reported in the main text. The analysis was performed for five bilateral MTL ROIs – namely, the AMY, entorhinal cortex, perirhinal cortex, posterior parahippocampal cortex, and hippocampus, identified based on the guidelines described by Moore et al. (2014). Although this includes regions that did not show any univariate effect, this strategy was used for the sake of completeness. Similar analyses were also performed differentiating EmoBG (RM) Hit vs. Miss.

4.1.2. Dataset

For each trial of each participant, four examples were created, based on the volumes near the peak of the hemodynamic response following image presentation. Specifically, examples were created based on the 4th, 5th, 6th, and 7th volumes (8-14 s) following stimulus onset. Relative to using an entire trial's worth of data for each example, using individual volumes sacrifices data quality in favor of achieving a large quantity of examples. This strategy was motivated by earlier fMRI classification research showing the benefits of having a large number of examples even if it requires lowering data quality (e.g., Dvornek, Yang, Ventola, & Duncan, 2018; Wang et al., 2020). Using this strategy for creating examples, two binary classification datasets were organized for the investigation of the two memory effects: EmoFG RHit-Hit vs. Emo Miss-Miss and EmoBG Hit vs EmoBG Miss. The EmoFG RHit-Hit condition yields 620 examples, Emo Miss-Miss yields 652 examples, EmoBG Hit yields

[1] A linear support vector machine was used, given earlier fMRI research showing its efficacy relative to similar classifiers (Plitt, Barnes, & Martin, 2015).

1312 examples, and EmoBG Miss yields 1208 examples. The datasets were stratified on a subject-by-subject basis. For example, if one participant yielded 28 EmoFG RHit-Hit examples and 32 Emo Miss-Miss examples, then 4 of the latter would be randomly dropped.

4.1.3. Cross-validation and permutation-testing

Because each participant contributes multiple examples, the dataset has a hierarchical structure. Hence, cross-validation defined each participant as a group to account for this structure, meaning that no participant's examples were split between folds – e.g., the first participant's examples may be entirely in the first fold. Earlier research has demonstrated that preserving the hierarchical structure of data during cross-validation is critical to avoid overfitting and, essentially, training-testing contamination (Winkler, Ridgway, Webster, Smith, & Nichols, 2014; Winkler, Webster, Vidaurre, Nichols, & Smith, 2015). Cross-validation was performed using two folds. This was repeated 100 times, with each instance having a random organization of participants into folds, and accuracy was averaged across the 100 repetitions. This cross-validation procedure, which averages across a large number of repetitions, has been shown to enhance the statistical sensitivity of multivariate analyses for detecting the effects of a dependent variable, relative to more traditional techniques such as 5-fold cross-validation or leave-one-out cross-validation (Bogdan et al., 2023; Valente, Castellanos, Hausfeld, De Martino, & Formisano, 2021). In other words, 100-repeated 2-fold cross-validation increases the likelihood that the classifier will yield statistically significant (above-chance) accuracy if the underlying neural data shows a genuine effect.

Permutation-testing was used to identify the level of accuracy that would be expected due to chance, and hence determine the accuracy threshold needed for statistical significance. For permutation-testing, labels were shuffled within-subject to preserve the dataset's hierarchical structure (Winkler et al., 2014; Winkler et al., 2015). The accuracy threshold required for statistically significant EmoFG RHit-Hit vs. Emo Miss-Miss classification is 52.4%, and the threshold required for significant EmoBG Hit vs. Miss classification is 51.9%; these accuracy thresholds reflect uncorrected $p < .01$ (i.e., $p < .05$ following Bonferroni correction for testing on five ROIs). The thresholds differ between the classification problems due to the different number of examples used for each analysis (see above).[2]

4.2. Results

First, for classification of EmoFG RHit-Hit vs. Emo Miss-Miss, the AMY classifier achieved 52.8% accuracy, which was significant. This result speaks to the validity of the two AMY clusters reported in the main text as showing increased activation linked to subjective confirmed by objective RM. In addition, the entorhinal cortex classifier achieved 53.0% accuracy. Although the univariate contrast did not identify a significant cluster in this region, the area notably showed a significant cluster in a previous report examining solely subjective RM using this dataset (Dolcos et al., 2020). The other three classifiers did not yield significant accuracy for this dependent variable. Second, for classification of EmoBG Hit vs. Miss, the parahippocampal cortex classifier achieved 52.0% accuracy, which was significant. The parahippocampal cortex did not show a significant activation effect of EmoBG hit, but it notably contained a suggestive sub-threshold cluster (14 voxels), which is consistent with the present findings.

[2] Although ~52% accuracy may initially seem to hardly surpass chance accuracy (50%), it is important to recognize that the analytic plan was structured such that individual examples are of low quality (individual volumes) but in high quantity. Earlier papers reporting extremely high classification accuracy (e.g., over 80%) have tended to use much more data (e.g., classification based on a whole session's worth of data). Additionally, the present use of repetitions during cross-validation increases the dataset's "effective size" even further because classification is tested on each example multiple times. For further discussion of this methodological strategy, which focuses on whether classification accuracy is above chance level rather than on classification accuracy itself *per se*, see Valente et al. (2021) and Bogdan et al. (2023).

References

Beck, A. T., Erbaugh, J., Ward, C. H., Mock, J., & Mendelsohn, M. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry, 4*(6), 561-&.

Bogdan, P. C., Iordan, A., Shobrook, J., & Dolcos, F. (2023). ConnSearch: A framework for functional connectivity analyses designed for interpretability and effectiveness at limited sample sizes *NeuroImage*, *278*, 120274.

Bredemeier, K., Spielberg, J. M., Silton, R. L., Berenbaum, H., Heller, W., & Miller, G. A. (2010). Screening for depressive disorders using the Mood and Anxiety Symptoms Questionnaire Anhedonic Depression Scale: a receiver-operating characteristic analysis. *Psychological assessment, 22*(3), 702-710.

Carlsson, M., & Hamrin, E. (2002). Evaluation of the life satisfaction questionnaire (LSQ) using structural equation modelling (SEM). *Quality of Life Research, 11*, 415-426.

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PloS ONE, 12*(11), e0184923.

Dolcos, F., Katsumi, Y., Bogdan, P. C., Shen, C., Jun, S., Buetti, S., . . . Dolcos, S. (2020). The impact of focused attention on subsequent emotional recollection: a functional MRI investigation. *Neuropsychologia, 138*, 107338.

Dvornek, N. C., Yang, D., Ventola, P., & Duncan, J. S. (2018). *Learning generalizable recurrent neural networks from small task-fmri datasets.* Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology, 85*(2), 348.

Judah, M. R., Grant, D. M., Mills, A. C., & Lechner, W. V. (2014). Factor structure and validation of the attentional control scale. *Cognition & emotion, 28*(3), 433-451.

Kertz, S. J., Lee, J., & Bjorgvinsson, T. (2014). Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychol Assess, 26*(4), 1146-1154. doi: 10.1037/a0037251

MacKillop, J., & Anderson, E. J. (2007). Further psychometric validation of the mindful attention awareness scale (MAAS). *Journal of psychopathology and behavioral assessment, 29*, 289-293.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behav Res Ther, 28*(6), 487-495. doi: 10.1016/0005-7967(90)90135-6

Moore, M., Hu, Y., Woo, S., O'Hearn, D., Iordan, A. D., Dolcos, S., & Dolcos, F. (2014). A comprehensive protocol for manual segmentation of the medial temporal lobe structures. *JoVE (Journal of Visualized Experiments)*(89), e50991.

Neumann, A., van Lier, P. A., Gratz, K. L., & Koot, H. M. (2010). Multidimensional assessment of emotion regulation difficulties in adolescents using the difficulties in emotion regulation scale. *Assessment, 17*(1), 138-149.

Noble, S., Mejia, A. F., Zalesky, A., & Scheinost, D. (2022). Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proceedings of the National Academy of Sciences, 119*(32), e2203020119.

O’Connor, R. C., & Forgan, G. (2007). Suicidal thinking and perfectionism: The role of goal adjustment and behavioral inhibition/activation systems (BIS/BAS). *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 25*(4), 321-341.

Plitt, M., Barnes, K. A., & Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical, 7*, 359-366.

Roth, R. M., Gioia, G. A., & Isquith, P. K. (2005). BRIEF-A: behavior rating inventory of executive function-adult version: Psychological Assessment Resources.

Spielberger, C., Gorsuch, R., Lushene, R., Vagg, P., & Jacobs, G. (1970). Manual for the state-trait anxiety inventory: Palo Alto: Consulting Psychologists Press.

Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: validity of permutation strategies and power of cross-validation schemes. *NeuroImage*, 118145.

Wang, X., Liang, X., Jiang, Z., Nguchu, B. A., Zhou, Y., Wang, Y., . . . Wu, F. (2020). Decoding and mapping task states of the human brain via deep learning. *Human Brain Mapping, 41*(6), 1505-1519.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage, 92*, 381-397.

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., & Smith, S. M. (2015). Multi-level block permutation. *NeuroImage, 123*, 253-268.

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *NeuroImage, 53*(4), 1197-1207.