The locus of recognition memory signals in human cortex depends on the complexity of the memory representations

D. Merika W. Sanders[1,*] and Rosemary A. Cowell[1]

[1]Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA 01003

Abbreviated Title: Recognition memory in visual cortex

*Corresponding author:

D. Merika W. Sanders, merikasanders@gmail.com

Pages: 65

Figures: 11 (9 in main manuscript, 2 in extended data)

Tables: 4

Abstract: 250 words

Introduction: 719 words

Discussion: 1706 words

Acknowledgments: This work was supported by the National Science Foundation Award #1554871. We thank Dave Huber and Jeff Starns for helpful discussions.

Conflict of interest: The authors declare no competing financial interests.

Abstract

Representational theories predict that brain regions contribute to cognition according to the information they represent (e.g., simple versus complex), contradicting the traditional notion that brain regions are specialized for cognitive functions (e.g., perception versus memory). In support of representational accounts, substantial evidence now attests that the Medial Temporal Lobe (MTL) is not specialized solely for long-term declarative memory, but underpins other functions including perception and future-imagining for complex stimuli and events. However, a complementary prediction has been less well explored, namely that the cortical locus of declarative memory may fall outside the MTL if the to-be-remembered content is sufficiently simple. Specifically, the locus should coincide with the optimal neural code for the representations being retrieved. To test this prediction, we manipulated the complexity of the to-be-remembered representations in a recognition memory task. First, participants in the scanner viewed novel 3D objects and scenes, and we used multivariate analyses to identify regions in the ventral visual-MTL pathway that preferentially coded for either simple features of the stimuli, or complex conjunctions of those features. Next, in a separate scan, we tested recognition memory for these stimuli and performed neuroimaging contrasts that revealed two memory signals – feature memory and conjunction memory. Feature memory signals were found in visual cortex, while conjunction memory signals emerged in MTL. Further, the regions optimally representing features via preferential feature-coding coincided with those exhibiting feature memory signals. These findings suggest that representational content, rather than cognitive function, is the primary organizing principle in the ventral visual-MTL pathway.

Keywords: visual cortex, feature, neural representations, recognition memory, fMRI

Significance Statement

Representational theories of cognition propose that the primary principle driving how memory and perception map onto the brain is the type of information that each brain region contains. A key prediction is that a classic "memory region", the medial temporal lobe (MTL), should underpin cognitive tasks beyond memory, if the task stimuli necessitate representations of MTL-appropriate information (e.g., complex objects and scenes). Substantial evidence now supports this prediction. However, the complementary prediction – that declarative memory should engage areas outside of MTL if the stimuli are sufficiently simple – is less thoroughly explored. We provide neuroimaging evidence that extra-MTL regions support long-term recognition memory for simple information, and that the location of a memory signal depends on its information content.

Introduction

The Swiss Army knife analogy of brain function (Cosmides & Tooby, 1994) asserted that the human mind contains many functionally distinct cognitive "tools" – e.g., for language acquisition, reinforcement learning, visual perception – each employing different mechanisms that emerged independently. This analogy greatly influenced cognitive neuroscience, encouraging researchers to focus on the separability of cognitive functions and map them to distinct neuroanatomical substrates (Aggleton & Brown, 1999; Kanwisher, 2006). More recently, some approaches have rejected a one-to-one mapping between cognitive functions and focal brain structures, instead mapping mnemonic functions to distributed networks (Ranganath & Ritchey, 2012; Rugg & Vilberg, 2013; Thakral, Wang, & Rugg, 2017). However, an important legacy of the Swiss army knife view is that most research questions still center on a cognitive function − e.g., episodic memory retrieval, decision-making, or selective attention − and ask which brain region, distributed network, or neural mechanism supports it. That is, cognition is still carved into subdomains that correspond to functional "tools" (Cowell, Barense, & Sadil, 2019).

A partial departure from this approach was made by models that invoke information content to explain the roles of different brain regions in memory (e.g., Davachi, 2006; Diana, Yonelinas, & Ranganath, 2007; Shimamura, 2011). But many such models still explain brain regions' contributions via a *combination* of function and content, e.g., claiming the hippocampus underpins recollection (a mnemonic function) of contextual details (a type of content) (Bastin et al., 2019; Davachi, Mitchell, & Wagner, 2003; Eichenbaum, Yonelinas, & Ranganath, 2007; Staresina, Cooper, & Henson, 2013). We, and others, have suggested that "content" is more than just a label to be used *in combination with* mnemonic functions or processes, such as

recollection; instead, it may be the primary principle driving how memory and perception map onto the brain at the systems level (Bussey & Saksida, 2002; Cowell et al., 2019; Graham, Barense, & Lee, 2010; Nadel & Peterson, 2013).

In this vein, the Representational-Hierarchical view of memory (Bussey & Saksida, 2002; Cowell, Bussey, & Saksida, 2010b) holds that regions in the ventral visual-medial temporal lobe (MTL) pathway engage in a cognitive task to the extent that they contain stimulus representations useful for solving that task. This predicts that the MTL is important not just for memory, but also for perception of complex stimuli such as scenes and 3-dimensional objects. Substantial empirical work now supports this prediction (e.g., Bartko, Winters, Cowell, Saksida, & Bussey, 2007; Buckley, Booth, Rolls, & Gaffan, 2001; Lee et al., 2005; Lee, Yeung, & Barense, 2012), eroding the traditional notion that the MTL specializes in long-term, declarative memory (Squire & Zola-Morgan, 1991). Indeed, the multi-tasking abilities of the MTL are now known to extend to statistical learning (e.g., Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014), decision-making (e.g., Palombo, Keane, & Verfaellie, 2015), and future simulation (e.g., Thakral, Madore, Addis, & Schacter, 2020). But the hypothesis that the MTL is involved in perception has a complementary prediction, which follows just as clearly from representational accounts: extra-MTL regions should underpin long-term, declarative memory whenever they contain representations of the to-be-remembered information. This complementary prediction has been less thoroughly investigated, although recent studies suggest support for it: early visual regions underlie recognition memory for orientation (Cooke, Komorowski, Kaplan, Gavornik, & Bear, 2015) and spatial location (Karanian & Slotnick, 2018; Thakral, Slotnick, & Schacter, 2013), and pattern completion-based recollection occurs outside

of hippocampus for stimuli not containing rapidly-learned associations (Ross, Sadil, Wilson, & Cowell, 2018).

We had two goals: first, to test the hypothesis that extra-MTL regions in "sensory" cortex underpin declarative memory when the to-be-remembered information is sufficiently simple; and second, to ask whether memory signals coincide anatomically with the optimal neural code for the representation being retrieved. In an fMRI study of recognition memory, we manipulated the complexity of the retrieved representations, measuring memory signals for simple features and, separately, for conjunctions of features. In addition, we applied multivariate analyses to an independent dataset collected during stimulus viewing, to characterize whether each brain region preferentially codes for individual features versus whole conjunctions. We predicted that manipulating the complexity of mnemonic information would shift the location of the memory signal within the ventral visual-MTL pathway, from posterior for feature memory to more anterior for conjunction memory. We further expected that regions exhibiting feature-coding would coincide with regions signaling feature memory, with both residing outside of MTL.

Materials and Methods

Design Overview

The study design included two primary objectives: (1) to find the cortical loci of feature-coding (in which neural representations are more informative about individual features of an object or scene than about conjunctions of those features), and conjunction-coding (in which representations are more informative about conjunctions than their features); and (2) to find the cortical loci of both *feature memory* (memory signals that can differentiate familiar from novel *features*) and *conjunction memory* (memory signals that can discriminate familiar from novel *conjunctions*). The R-H account predicts that (1) feature-coding will appear earlier in the ventral

visual-MTL pathway than conjunction-coding, and (2) recognition memory signals will appear in posterior versus anterior areas according to whether feature-based or conjunction-based representations differ between familiar and novel stimuli. We examined these predictions with two sets of stimuli: novel 3D objects, called fribbles (Barry, Griffith, De Rossi, & Hermans, 2014; Williams, 1998), and novel computer-generated scenes.

To achieve our first objective of measuring neural coding, we conducted a pair of scanning sessions (i.e., study scans) – one for each stimulus set, on separate days – in which participants completed an incidental 1-back task (to ensure attention) while viewing a stream of fribble or scene stimuli. Each fribble or scene stimulus was composed of the conjunction of three simple, binary features (see Figure 1, Figure 2 and *Stimuli*). We then calculated the feature conjunction index (FCI) established by Cowell, Leger & Serences (2017), which we describe briefly next, and in *Feature-Conjunction Index Analysis* in more detail.

We used data from the study scans to train a set of classifiers to identify either a specific feature or the specific stimulus that was viewed on a given trial, from the neural activation pattern. Three of the classifiers were trained to identify each of the three separate features (i.e., which of the two possible values of a feature was present). The product of the three two-way feature classifiers, termed the *predicted* item-level classification accuracy, indicated the extent to which a brain region could identify the whole stimulus *based only on a linear combination of its knowledge about individual features*. A final classifier was trained to identify the unique conjunction of features (i.e., the individual fribble or scene). This *empirically observed* item-level classification accuracy indicated the extent to which a brain region contained information about the whole item. The *predicted* and the *empirically observed* item-level classification accuracies were pitted against each other in a log-ratio, giving the FCI:

$$\mathrm{FCI} = \ln\left(\frac{Empirical\ object\ accuracy}{Predicted\ object\ accuracy}\right) \quad (1)$$

A positive FCI value indicated conjunction-coding, and a negative FCI value indicated feature-coding. Thus, although other methods have identified conjunction-coding for simple conjunctions of low-level features in visual cortex (e.g., color, form, motion, and orientation; Anzai, Peng, & Van Essen, 2007; Baumgartner et al., 2013; Engel, 2005; Gegenfurtner, Kiper, & Levitt, 1997; Johnson, Hawken, & Shapley, 2008; Pollmann, Zinke, Baumgartner, Geringswald, & Hanke, 2014; Seymour, Clifford, Logothetis, & Bartels, 2009, 2010; Wilkinson et al., 2000), or conjunction-coding for more complex conjunctions in more anterior brain regions (e.g., Baeck, Wagemans, & Op de Beeck, 2013; Coutanche & Thompson-Schill, 2015; Erez, Cusack, Kendall, & Barense, 2016; MacEvoy & Epstein, 2009; Martin, Douglas, Newsome, Man, & Barense, 2018; van den Honert, McCarthy, & Johnson, 2017; J. Zhang, Liu, & Xu, 2015), the present method is unique in that it measures a *relative* preference for feature- versus conjunction-coding, allowing identification of regions that not only respond to features, but actively prefer features. In this study, we apply this method to complex objects and scenes for the first time.

To achieve our second objective, we conducted a second pair of scanning sessions (i.e., test scans) – one each for fribbles and scenes, on separate days, with each test scan occurring on the day of the same stimulus set's study scan – in which participants performed a recognition memory task involving the fribble or scene stimuli they had viewed in the first session. Trials belonged to one of three mnemonic classes: Novel, Recombination, and Familiar. On Novel trials, stimuli were entirely novel, meaning all stimulus features and conjunctions of features were novel. On Familiar trials, both the individual features and the conjunction of those features had been seen in the first session, and thus both were familiar. On Recombination trials, the

individual features were familiar from having been seen as parts of items in the first session, but those features were conjoined in a novel way.

We assayed different types of memory by contrasting different pairs of memory trial types. A contrast of Novel and Recombination trials indexed *feature memory* because the familiarity of individual features is the factor that discriminates the two trial types (features are novel during Novel trials and familiar during Recombination trials, whereas the conjunction is novel for both). A contrast of Recombination and Familiar trials indexed *conjunction memory* because the familiarity of the conjunction of features is the factor that discriminates the two trial types (features of both trial types are familiar, but the conjunction on Recombination trials is novel while the conjunction on Familiar trials is familiar). This approach specifically tests feature memory, divorced from conjunction memory, and conjunction memory, divorced from feature memory; this delivers finer control of the content of mnemonic representations than, for example, earlier work demonstrating cortical reinstatement in early visual areas during retrieval of visual information (e.g., Thakral et al., 2013).

Following the R-H theory, we expected feature-coding to reside posterior to conjunction-coding, and feature memory signals to lie posterior to conjunction memory signals. Further, the two neural coding modes (feature versus conjunction) and the two memory signals (feature versus conjunction) were expected to differ in location across stimulus sets, with the more complex scene representations expected to lie anterior to those of fribbles. Finally, the study design allowed us to measure the relative locations of representational content (neural coding type) and signals for mnemonic retrieval. We predicted anatomical correspondence between neural coding type and memory signal, with the greatest signals for feature memory coinciding with feature-coding brain areas, outside of the MTL.

Participants

Twenty-three participants were recruited from the University of Massachusetts-Amherst community. All participants spoke English fluently, had normal or corrected-to-normal vision, no history of neurological illness, and no contraindications for MRI scanning. Participants were compensated $25/hour with an additional performance-based bonus up to $10 per scan session. Table 1 highlights when extreme outliers were identified and subsequently excluded from statistical analyses; for repeated measures analyses all datapoints from a participant identified as an outlier were excluded.

Stimuli

Two different stimulus sets were created. The first set comprised novel 3-D objects (i.e., fribbles) created using Strata Design 3D CX 7.5 (Barry et al., 2014; Williams, 1998). Each fribble was a unique conjunction of three simpler features (3-D colored shapes referred to as "tail", "body", and "head" features) and each fribble belonged to a "family" (see Figure 1). Within a family, there were two possible variants for each of the tail, body and head features and those variants were unique to that family. Therefore, each family contained eight unique fribbles corresponding to all possible conjunctions of a family's binary features. A total of four fribble families were created.

The second stimulus set comprised novel 3-D scenes created using Sweet Home 3D, an indoor planning software. Analogous to the fribbles, each scene was a unique conjunction of three features (room shape, color, and furniture) and each scene belonged to a "family" (see Figure 2). Within a family, there were two possible variants for each of the three room features and those variants were unique to that family. Consequently, each family comprised eight unique

scenes corresponding to all possible conjunctions of a family's binary features. A total of four scene families were created.

Each participant was scanned on two separate days, one day for each of fribbles and scenes; stimulus set presentation order was counterbalanced across participants. Each day's scanning involved two scan sessions – an initial, study phase scan and a second, memory test scan – separated by a short break. For each stimulus set (fribbles/scenes), each participant was assigned two families to view in the study phase, thus creating stimuli designated as Familiar in the later test phase. A second class of test stimuli, Recombination stimuli, was created by combining two features of one Familiar family with a third feature of the other Familiar family (Figure 3). For Recombination stimuli, although all the individual features were seen as part of Familiar stimuli during the study phase, the conjunction of the features into these specific, unique wholes was never seen until the memory test. A final mnemonic stimulus class, Novel stimuli, were created in the same manner as Recombination stimuli (i.e., 1 + 2 feature combination), but using features from the two remaining unseen families not designated as Familiar stimuli. Thus, Novel stimuli were entirely novel at the second memory test scan. The assignment of families to stimulus classes (i.e., which two families appeared in the study scan and which two families were reserved to create Novel stimuli at test) was counterbalanced across participants.

Task

Tasks were identical for fribble and scene stimulus sets. During study scans, participants completed a 1-back detection task (i.e., indicate by button press whether the stimulus on screen is the same or different from that of the previous trial). There were 15 null trials randomly inserted into 15 of the inter-stimulus intervals (ISIs) between study experimental trials, during which

participants saw a white central fixation cross (+) that dimmed briefly. Dimming occurred once or twice per null trial and participants could press either response key whenever it occurred. Participants did not need to make a 1-back detection response during stimulus trials that immediately followed a null trial or occurred at the start of a scan. Before entering the scanner, participants practiced this task on a laptop computer.

During the break between study and test scan sessions, participants received instructions on how to distinguish between the three mnemonic stimulus classes (Familiar, Recombination, and Novel). During test scans, participants indicated the stimulus class of the stimulus on the screen using three button box response keys: (1) “Familiar” (i.e., studied in the previous session); (2) “Recombination” (i.e., made of features that had been studied in the previous session, but combined in a new way); or (3) “Novel” (i.e., not studied in the previous session in any form). In contrast to study scans, null trials occurred during every ISI between test experimental trials.

For both study and test tasks, participants were instructed to respond as accurately as possible while the stimulus was still on the screen, and were informed that a performance-based bonus was available based on accuracy.

Experimental Design

The design was identical for fribble and scene stimulus sets. For each stimulus set, each participant completed ten study and six test functional scans, termed “runs” (Figure 4). During the runs, stimuli were displayed on a 32” LCD monitor positioned at the head end of the magnet bore. Participants viewed the screen via a mirror on the head coil.

In the first, study scan session, stimulus presentation order was blocked across scans by family (e.g., stimuli from one Familiar family were presented in even-numbered study runs and stimuli from the other Familiar family in odd-numbered study runs), but stimuli within a family

block/run were presented in random order. Within each study scan there were 35 stimulus trials: four repetitions each of eight stimuli from a Familiar family, presented singly on a gray background, plus three items repeated a fifth time to create “immediate repeats” (later removed from the analysis) for the 1-back task. Each stimulus was presented for 2000ms, immediately followed by an ISI of between 2000ms and 8000ms, during which a white central fixation cross appeared on a gray background. The response window for stimulus trials began 500ms after stimulus onset and extended 1000ms into the ISI.

At least 500ms after the end of the response window, for 15 of the 35 ISIs in each study run, a null trial occurred. Null trials never occurred during ISIs less than 3000ms; instead, participants simply saw a white central fixation cross on a gray background. For null trials occurring during ISIs between 3000ms and 6000ms, the white cross dimmed to gray once, for 250ms. For null trials during ISIs longer than 6000ms, the white cross dimmed to gray twice, each time for 250ms. The exact onset of the dimming event within the ISI was randomized, but constrained to occur at least 1000ms before the end of the ISI and with at least 1500ms between two dimming events within the same ISI. Null trials measured participants' attention and also provided gaps in the stimulus sequence with minimal “mind-wandering” (Stark & Squire, 2001). After completion of the 10 study runs, participants exited the scanner and took a self-paced break before re-entering for the six test runs.

During test runs, 48 Novel, 48 Recombination and 48 Familiar stimulus trials were presented, with 16 unique stimuli from each mnemonic stimulus class shown three times each. Novel and Recombination stimuli were created by combining one feature of one family (unstudied for Novel or studied for Recombination) with a third feature of the other unstudied/studied family. By implementing all possible 1 + 2 feature combinations, this design

method yielded 48 unique stimuli for both Novel and Recombination stimulus classes. However, because there were only 16 unique Familiar stimuli, only a subset of 16 out of the 48 possible stimuli was selected for presentation at test for each of Novel and Recombination classes. Which 16 stimuli of the 48 were included in this subset was counterbalanced across participants.

Trial order and spacing of the three stimulus types was optimized using the easy-optimize-x MATLAB tool, which finds the most efficient design for later contrasting of conditions. Optimization was constrained such that the three stimulus types were equally distributed across the six test runs (i.e., eight of each type per run) and that repetition was blocked (i.e., all 48 stimuli were shown once before any stimulus was presented a second time). Each test stimulus appeared for 2500ms, with a randomly assigned ISI between 4000ms and 12000ms. The response window for stimulus trials extended 1000ms into the ISI. All ISIs included a null trial, as described above, which did not commence until after the response window had elapsed (3500ms after stimulus trial onset).

Image Acquisition

Scanning was performed on a Siemens 3T Skyra scanner equipped with a 64-channel head coil at the University of Massachusetts-Amherst's Human Magnetic Resonance Center. Functional images were acquired using a T2-weighted EP2D-BOLD sequence (TR: 1250ms; TE: 33ms; flip angle: 70°; FOV: 210mm; 2.5mm^{3} voxels) and forty-two axial slices were acquired during each functional scan. Whole-brain anatomical images were acquired in the middle of the study functional scan session using a T1-weighted (T1w) MP-RAGE sequence (208 sagittal slices; TR: 2000ms; TE: 2.13ms; flip angle: 9°; 1mm^{3} voxels).

Statistical Analyses

fMRI Data Preprocessing

Preprocessing was performed using fMRIPrep 1.5.2 (Esteban et al., 2019; Esteban, Markiewicz, DuPre, et al., 2018), which is based on Nipype 1.3.1 (Esteban, Markiewicz, Johnson, et al., 2018; Gorgolewski et al., 2011). T1w images were corrected for intensity non-uniformity using N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants, Epstein, Grossman, & Gee, 2008). T1w images were then skull-stripped with antsBrainExtraction.sh, using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid, white-matter and gray-matter was performed on the brain-extracted T1w images using FAST (FSL 5.0.9; Y. Zhang, Brady, & Smith, 2001). An unbiased registration of the two T1w images using mri_robust_template (FreeSurfer 6.0.1; Reuter, Rosas, & Fischl, 2010) produced a single T1w-reference map used throughout the workflow. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1; Dale, Fischl, & Sereno, 1999) and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs- and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (MNI152NLin2009cAsym; Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both the T1w reference and the standard space template.

For each of the 34 BOLD scans per participant (i.e., 20 study runs, 12 test runs, and 2 functional localizer), a BOLD reference image and its skull-stripped version were generated using a custom methodology of fMRIPrep. Using this BOLD reference, head-motion parameters were estimated using mcflirt (FSL 5.0.9; Jenkinson, Bannister, Brady, & Smith, 2002) and susceptibility distortions were corrected using a custom workflow of fMRIPrep that estimated a deformation field based on a field map that was co-registered to the BOLD reference. This

estimated susceptibility distortion was used to calculate an unwarped BOLD reference for more accurate co-registration with the T1w reference using bbregister (FreeSurfer). In a single interpolation step that concatenated the above transformations, the BOLD time-series were resampled into subject native space and standard space – the latter only used for the searchlight analysis to define group-level FCI-regions of interest (ROIs). Volumetric (gridded) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Surface (nongridded) resamplings to subject native space (fsnative) and the fsaverage template space were performed using mri_vol2surf (FreeSurfer) and the resulting transformation matrices were later used to apply probabilistic ROI atlases. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

After fMRIPrep's preprocessing pipeline, functional data was high-pass filtered with a 128s cutoff. Data used in univariate analyses (i.e., functional data from test scan sessions) also underwent smoothing with a 5mm FWHM Gaussian kernel. The data were analyzed with a combination of custom MATLAB scripts utilizing SPM 12 software and the CoSMoMVPA toolbox (Oosterhof, Connolly, & Haxby, 2016). Immediate-repeat trials from study scan sessions and incorrect trials from test scan sessions were omitted from all neuroimaging data analyses.

Definition of Regions of Interest

Each participant completed one localizer scan, but, ultimately, we used probabilistic atlases to define ventral visual stream areas (i.e., V1, V2, V3, V3AB, hV4, ventral occipital cortex [VOC], LOC, temporal cortex [TOC]; Wang, Mruczek, Arcaro, & Kastner, 2015) and MTL areas (i.e., PHC, PRC, HC; Ritchey, Montchal, Yonelinas, & Ranganath, 2015). In all

analyses, data from left and right hemispheres were combined into a single ROI. A subset of representative ROIs were selected for analysis, namely, V1, V2, V3, LOC, PHC, PRC, and HC.

In addition to the probabilistically defined ROIs, we defined a conjunction-coding and a feature-coding ROI for each stimulus set, as detailed below. These ROIs' definition was non-circular because they were functionally defined from study scan data, then used in the analysis of entirely separate test scan data (Nikolaus Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

Feature-Conjunction Index Analysis

The FCI analysis followed the methods described in Cowell et al. (2017). For each stimulus set, a separate GLM was estimated for each stimulus trial of the study phase, excluding immediate repeats (Mumford, Turner, Ashby, & Poldrack, 2012). Each of the 32 single-trial GLMs had one regressor for the trial, one regressor for all remaining trials, and six motion nuisance regressors. Each regressor combined a boxcar model of the study stimulus time-series with a canonical hemodynamic response function (HRF). The model provided activation estimates for every voxel within an ROI for each trial separately. Estimates of single-trial activation patterns across all voxels within an ROI were then used in classification analyses.

Study scan data from the two Familiar families had been collected separately (i.e., blocked by family) and were kept separate for classifier training. For each of the two Familiar families of a stimulus set (i.e., five study runs), four non-probabilistic linear discriminant analysis (LDA) classifiers were trained using leave-one-run-out cross-validation: three 2-way feature-level classifiers and one 8-way conjunction-level classifier. This cross-validation approach involved training the classifier with activation patterns from four runs of 32 trials (i.e., 128 trials) and returning predicted conjunction or feature labels for the 32 trials of the held out

fifth run. Predicted labels were the label that had the highest discriminant score for each trial. For each family, the process was repeated five times, with each run serving as the test sample once.

For each trial and for all classifiers, predicted labels scored 0 or 1 for incorrect versus correct classification. The three feature-level classification accuracies for a trial were then multiplied together to provide a *predicted* conjunction classification accuracy (i.e., 0 or 1) for that trial. The *predicted* conjunction could only be correct if all three features were classified accurately. Next, both the *empirically observed* conjunction classification accuracy (i.e., 0 or 1 depending on whether the 8-way conjunction-level classifier predicted the correct label) and the *predicted* conjunction classification accuracy were averaged across trials.

The averaged *empirically observed* and the averaged *predicted* conjunction classification accuracy were compared by taking the natural log of their ratio, producing the FCI. Positive FCIs indicate that activation within a region is modulated by conjunctions more than features and negative FCIs indicate that activation within a region is modulated by features more than conjunctions (Cowell et al., 2017). FCIs were averaged across the two Familiar families for each participant.

Memory Scores Analysis

For each stimulus set, a GLM was constructed with ten regressors: correct responses to each of Familiar, Recombination and Novel stimulus types; incorrect responses; and six motion nuisance regressors. Each non-motion regressor combined a boxcar model of the test stimulus time-series with a canonical HRF. For each participant, the model provided activation estimates as beta weights (β) for every voxel within an ROI, for each of the four modelled conditions.

Memory signal indices were derived from two contrasts, defined in (2) and (3). To index feature memory, we contrasted the beta weights from Novel and Recombination correct trials

(because familiarity differences between these trial types are dominated by features: all features are familiar in Recombination stimuli but all features are novel in Novel stimuli, whereas the three-feature conjunction is novel in both). To index conjunction memory, we contrasted the beta weights from Recombination and Familiar correct trials (because familiarity differences between these trial types stem from conjunctions: the conjunctions are novel in Recombination stimuli but familiar in Familiar stimuli, whereas all individual features are familiar in both). For feature memory, the beta weight estimates from the GLM for correct Recombination trials ($\beta_{Recombination}$) were subtracted from the beta weights for correct Novel trials (β_{Novel}). For conjunction memory, the beta weights for correct Familiar trials ($\beta_{Familiar}$) were subtracted from the beta weights for correct Recombination trials ($\beta_{Recombination}$).

For both memory contrasts, the mean difference across voxels within an ROI was divided by the standard deviation of that difference to obtain a directional Cohen's *d* statistic. Calculating effect sizes allows comparison of the magnitude of effects across ROIs with different numbers of voxels and different signal-to-noise levels (Erez et al., 2016; Lakens, 2013; Rosenthal, 1994). These effect sizes are taken as feature memory scores (Novel versus Recombination contrast) and conjunction memory scores (Recombination versus Familiar contrast):

$$Feature\ Memory\ Score = Cohen's\ d_{\ Feature} = \frac{\bar{x}_{\beta_{Novel} - \beta_{Recombination}}}{sd_{\beta_{Novel} - \beta_{Recombination}}} \quad (2)$$

$$Conjunction\ Memory\ Score = Cohen's\ d_{\ Conjunction} = \frac{\bar{x}_{\beta_{Recombination} - \beta_{Familiar}}}{sd_{\beta_{Recombination} - \beta_{Familiar}}} \quad (3)$$

Where $\beta_{Novel} - \beta_{Recombination}$ is a subscript denoting the difference of two vectors: β_{Novel} refers to a vector of beta weights from all voxels in an ROI on "Novel" trials in

which the participant responded correctly; and $\beta_{Recombination}$ refers to a vector of beta weights from all voxels in the same ROI on correct "Recombination" trials. Thus, $\bar{x}_{\beta_{Novel} - \beta_{Recombination}}$ is the mean $\beta_{Novel} - \beta_{Recombination}$ difference vector for an ROI, and $sd_{\beta_{Novel} - \beta_{Recombination}}$ is the standard deviation of the $\beta_{Novel} - \beta_{Recombination}$ difference vector calculated across all voxels in the ROI. Analogously, the $\beta_{Recombination} - \beta_{Familiar}$ subscript denotes the difference of two vectors: $\beta_{Recombination}$ refers to a vector of beta weights from all voxels in an ROI on correct "Recombination" trials; and $\beta_{Familiar}$ refers to a vector of beta weights for all voxels in the same ROI on correct "Familiar" trials. Thus, $\bar{x}_{\beta_{Recombination} - \beta_{Familiar}}$ is the mean $\beta_{Recombination} - \beta_{Familiar}$ difference vector for an ROI, and $sd_{\beta_{Recombination} - \beta_{Familiar}}$ is the standard deviation of the $\beta_{Recombination} - \beta_{Familiar}$ difference vector.

Searchlight Analysis of the Intersection of Representational Content and Memory Signals

The above analyses produced two types of dependent variable: FCI scores from the study phase, and memory scores (both feature and conjunction) from the test phase. The R-H account predicts that because feature memory should rely primarily on feature representations to produce a familiarity difference, feature-coding regions should show greater feature memory scores than conjunction-coding regions. Similarly, because conjunction memory should draw upon conjunction representations to produce a familiarity difference, conjunction-coding regions should show greater conjunction-memory scores than feature-coding regions.

To test this prediction, ROIs exhibiting feature-coding and conjunction-coding were defined empirically, via a searchlight analysis of the study scan data (Kriegeskorte, Goebel, & Bandettini, 2006). For each stimulus set separately, a searchlight was swept through the brain by defining overlapping, spherical sets of voxels (radius 5 functional voxels), computing the FCI within each spherical ROI, and recording the FCI value at the center voxel. The group average of

the spherical ROIs was calculated by averaging over participants' FCI values at each center voxel, in normalized space. A group-level *t*-test at each voxel was then used to compare the group mean FCI to zero (uncorrected for multiple comparisons because these tests were used only to define ROIs, not to make inferences from the dependent variable of interest). This comparison was used to define brain regions – to be used as group-level ROIs – that demonstrated extremes of feature-coding (i.e., negative FCI values less than zero) or conjunction-coding (i.e., positive FCI values greater than zero). We then calculated conjunction and feature memory scores within these group-level ROIs. ROIs were stimulus-specific (fribbles versus scenes), and memory scores computed within each negative and positive FCI-ROI drew only upon the corresponding fribble or scene test scan data.

Bayes Factor Analysis

We supplemented the frequentist analyses with Bayes factors, which are useful for interpreting frequentist analyses in which the *p*-value does not reach the traditional significance threshold of 0.05. A *p*-value greater than 0.05 does not discern between evidence in *favor* of the null hypothesis versus a lack of evidence for either the null or alternative hypothesis. Instead, a Bayes factor captures the relative evidence of one hypothesis over another, given the observed data. Bayes factors near one indicate that the data are equally likely to occur under both hypotheses, while values much larger than one and fractional values much smaller than one indicate evidence for one hypothesis over the other. Although Bayes factors are continuous, there are some categorical guidelines to gauge how strongly the evidence favors a hypothesis. Values greater than three indicate "substantial" evidence, greater than 10 "strong evidence", greater than 30 "very strong" evidence, and greater than 100 "decisive" evidence (the same guidleines can be

applied to the hypothesis in the denominator by making these values fractions; Jeffreys, 1961; Raftery, 1995).

We used the *BayesFactor* R package (Morey, Rouder, & Jamil, 2015) with the default input values for how many Monte Carlo simulations to generate (10,000) and the scale value for the Cauchy priors on effect size ($^1/_2$). The null hypothesis is placed in the denominator of the Bayes factor model comparison; therefore, any values larger than one provide evidence for the alternative hypothesis. For ANOVA designs with nested models (i.e., a full model containing all main effects and their interactions, as well as restricted models with fewer effects) all models are compared against the null (i.e., a model with only random subject effects). Because Bayes factors from these nested model comparisons share a common denominator, we can directly compare nested models by taking the ratio of their Bayes factors.

Results

Behavioral Analysis

Participants performed well during the study scan on the incidental one-back detection task, in which the correct response was "same" if the stimulus on the previous trial was the same as the current trial, or "different" otherwise. Discriminability between same and different trials, *d'*, was the difference between the normalized proportions of hits (i.e., participants responded "same" and the trial was same) and false alarms (i.e., participants responded "same" and the trial was different). The mean *d'* score was 3.21 and 3.26 for fribbles and scenes, respectively; these values did not significantly differ, t (22) = -0.26, p = 0.8.

Performance on the recognition memory task during the test scan session was also measured with *d'*, which indexed the discriminability of the three classes of mnemonic status that were present among the test stimuli: Familiar, Recombination, Novel. Because there were

more than two possible responses on a given trial, *d'* scores were calculated for each possible pairwise comparison of trial types (see Table 2). In each case, trials containing stimuli belonging to the third outstanding trial type were removed from the analysis. Any (erroneous) responses to the two trial types under comparison that invoked the outstanding trial type were binned as a hit, miss, false alarm or correct rejection, depending on the comparison. For example, in one comparison between Familiar and Novel trials, trials whose true status was Recombination were removed, and on the remaining trials, any "Recombination" response was considered a hit on a Familiar trial, but a false alarm on a Novel trial. In a second comparison of the same two trial types (Familiar and Novel), a "Recombination" response was considered a miss on Familiar trials, but a correct rejection on Novel trials.

Table 2 shows the average *d'* scores for all such comparisons; overall, memory discrimination performance was good. Performance did not differ significantly between stimulus sets ($p = 0.53$) but differed significantly between different pairwise comparisons (F [1.27, 26.57] = 89.07, $p < 0.001$). Unsurprisingly, *d'* scores were lowest, but still above chance, in the comparison of Familiar and Recombination trials; these trials both contained stimuli in which all features were familiar from the study scan, making for a difficult mnemonic discrimination.

fMRI Analysis

Feature-Conjunction Index Analysis

Although not of primary interest, mean classifier accuracies within each ROI for fribbles and scenes, separately, are presented in Figure 5. Accuracy was lowest in MTL (PHC, PRC, HC) where the BOLD signal has a lower signal-to-noise ratio (SNR; Olman, Davachi, & Inati, 2009). A heuristic for classifier chance performance was determined with a binomial test for two-way feature and eight-way conjunction classifiers, uncorrected for multiple comparisons. This

relatively liberal criterion was chosen because the outcome of this test was not of primary importance to the experimental results; that is, classifier accuracy was not a dependent variable for which we made predictions, it was simply an indicator of noise in each brain region. Classifier performance did not exceed this binomial chance threshold in MTL regions. Because FCIs derived from classifiers performing at or below chance tend toward zero (Cowell et al., 2017), we should be cautious when interpreting the FCI results in these MTL areas. It is possible that zero FCIs reflect "hybrid" coding (i.e., areas that demonstrate a mixture of feature- and conjunction-coding), but, in the absence of reliably above-chance classifier accuracy, zero FCIs could also reflect the existence of neither feature nor conjunction knowledge.

However, for regions outside of MTL with reliably above chance classifier performance, regional variation in SNR is much less likely to account for the FCI results described below. Importantly, because the FCI places evidence for feature- and conjunction-knowledge within the same ROI into a ratio, both measures should be similarly affected by noise. Simulations have shown that significantly negative FCI values can be produced only by a feature-based code and significantly positive values only by a conjunction-based code (see Cowell et al., 2017).

First, a two-way (2x7) ANOVA of FCI with factors of stimulus set (fribbles, scenes) and ROI (V1, V2, V3, LOC, PRC, PHC, HC) revealed a main effect of ROI, F (6, 126) = 4.41, p < 0.001, η^2_G = 0.08. There was no significant interaction of stimulus set and ROI (p = 0.06, η^2_G = 0.04), nor a main effect of stimulus set (p = 0.51, η^2_G = 0.002). Correspondingly, the Bayes factor analysis indicated very strong evidence for the model with only the effect of ROI compared to the null (BF = 43.45). We can directly compare two nested models by taking the ratio of their Bayes factors. When comparing the ROI only model to the ROI and stimulus set

interactive model ($BF_{ROI\ x\ Stimulus}$ = 6.63) there was substantial evidence in favor of the ROI only model (43.45/6.63 = 6.55).

Because there was no evidence of an effect of stimulus set, we performed a one-way ANOVA of ROI on FCIs averaged across stimulus set (Figure 6; though see Figure 6-1 for FCIs separated by stimulus set). In this follow-up analysis, the main effect of ROI was still significant, F (6, 132) = 3.94, $p < 0.001$, $\eta^2_G = 0.13$, indicating that FCI values differed between individual ROIs. V1 and V2 located in early ventral visual stream areas had significantly negative FCIs (two-tailed *t*-tests Bonferroni corrected for seven comparisons, FCI = -0.08, $p = 0.004$; and FCI = -0.1, $p = 0.0002$, respectively) and FCI values became numerically positive, though not significantly so, in MTL areas PHC (FCI = 0.02) and PRC (FCI = 0.01). Similarly, the Bayes factor analyses revealed very strong evidence in favor of the model with an effect of ROI compared to the null (BF = 45.14) and very strong evidence that FCI values were less than zero in early VVS (BF_{V1} = 121.81; BF_{V2} > 150). Interestingly, for PHC and PRC, which demonstrated numerically positive FCIs that did not reach significance in the frequentist analyses, there was substantial evidence for neither the alternative hypothesis, in which FCI values were greater than zero, nor the null (BF_{PHC} = .42; BF_{PRC} = .35). This suggests that failing to reject the null in the follow-up *t*-test analyses reflects a lack of power in these MTL areas, rather than support for the null.

Memory Scores Analysis

Memory score, measured by Cohen's D, served as the dependent variable in a three-way (2x7x2) ANOVA with factors of stimulus set (fribbles, scenes), ROI (V1, V2, V3, LOC, PRC, PHC, HC) and memory score type (feature memory, conjunction memory; see Figure 7-1 for memory scores by stimulus set). There was a main effect of ROI, F (2.32, 46.32) = 17.87, $p <$

0.001, $\eta^2_G = 0.14$, and memory score type, F (1, 20) = 84.75, $p < 0.001$, $\eta^2_G = 0.36$, and an interaction between these two factors, F (2.37, 47.32) = 28.4, $p < 0.001$, $\eta^2_G = 0.19$. However, contrary to our prediction, there was no main effect of stimulus set, nor any interactions between stimulus set and other factors ($p > 0.05$, $\eta^2_G < 0.006$). In line with this, in the Bayes factor analysis, the most favored of the nested alternative models omitted stimulus set and any of its interactions, but included ROI, memory score type and their interaction (BF > 150). Consequently, memory scores were again averaged across stimulus sets in a follow-up two-way (7x2) ANOVA.

The follow-up analysis revealed main effects of memory type, F (1, 18) = 73.6, $p < 0.001$, $\eta^2_G = 0.47$, and ROI, F (2.26, 40.72) = 17.56, $p < 0.001$, $\eta^2_G = 0.22$, and an interaction between memory type and ROI, F (2.27, 40.9) = 35.6, $p < 0.001$, $\eta^2_G = 0.30$ (Figure 7). We explored this interaction further by separating the data for feature and conjunction memory and performing two one-way ANOVAs. This revealed a significant main effect of ROI in both cases (feature memory: F [2.4, 43.2] = 32.59, $p < 0.001$, $\eta^2_G = 0.48$; conjunction memory: F [2.44, 43.84] = 3.41, $p = 0.03$, $\eta^2_G = 0.08$). The greatest signal for feature memory occurred in ROIs located in early ventral visual stream (V1, V2, V3) and this signal decreased moving anteriorly toward MTL. Additionally, the greatest signal for conjunction memory occurred in MTL (PHC, PRC, HC) and late-stage ventral visual stream (LOC), and decreased moving posteriorly toward early ventral visual stream. Conjunction memory scores clustered near zero, and thus we tested whether there was any reliable conjunction memory signal, that is, conjunction memory scores that significantly exceeded zero. Out of the seven individual ROIs, only four (LOC, PHC, PRC, HC) had numerically positive conjunction memory scores, therefore comparisons were made for these four ROIs only. Conjunction memory was significantly positive in PRC (one-tailed *t*-tests

Bonferroni corrected for four comparisons, Cohen's D = 0.16, p = 0.01), but was not different from zero in any other ROI with a numerically positive score.

Bayes factors corresponding to the two-way (7x2) ANOVA of memory scores reflected decisive evidence in favor of the interactive model (memory type x ROI) over the null (BF > 150). There was also decisive evidence for a main effect of ROI for feature memory (BF > 150) and substantial evidence in favor of a main effect of ROI for conjunction memory (BF = 8.59). Additionally, in line with the follow-up *t*-tests, there was strong evidence for conjunction memory scores greater than zero in PRC (BF = 13.68), and for the other three ROIs with numerically positive conjunction memory scores there was a lack of substantial evidence for either the alternative or the null (BF_{LOC} = .61; BF_{PHC} = .82; BF_{HC} = 1.18). Again, this suggests that failing to reject the null may reflect a lack of power rather than evidence against conjunction memory.

Searchlight Analysis of the Intersection of Representational Content and Memory Signals

In the above analyses, we averaged FCIs and memory scores across fribbles and scenes but, here, we defined FCI-ROIs *separately* for each stimulus set, for two reasons. First, in terms of our theory, we did not expect to find stimulus-invariant negative and positive FCI brain regions (i.e., "universal conjunction-coder" regions that are the site of conjunction-coding for both fribbles and scenes), so constraining the searchlight analysis to "shared" ROIs was not warranted *a priori*. Second, the absence of stimulus set related differences in the FCIs and memory scores, and the absence of stimulus set by ROI interactions, in the *a priori* ROIs above, does not necessarily entail that the same regions of the brain will contain the highest positive and lowest negative FCIs for the two stimulus sets.

The FCI-ROIs for fribbles and scenes are shown in Figure 8. Table 3 and Table 4 list the coordinates for the ten most negative and ten most positive FCI-voxels included in the memory score analysis, and the general areas where these coordinates fell within the Harvard-Oxford probability atlas (distributed with FSL, Smith et al., 2004). Fribbles and scenes contained a similar number of voxels in their respective negative and positive FCI-ROIs. For fribbles, the negative FCI-ROI consisted of 5194 voxels and the positive FCI-ROI consisted of 1544 voxels. For scenes, the negative FCI-ROI contained 6088 voxels and the positive FCI-ROI contained 1173 voxels. Between the two stimulus sets, 1948 voxels were shared between negative FCI-ROIs, and 55 voxels were shared between positive FCI-ROIs.

For the fribble dataset, a two-way (2x2; [feature memory, conjunction memory] x [negative FCI-ROI, positive FCI-ROI]) ANOVA of memory score revealed an interaction between memory score type and ROI, F (1, 22) = 23.94, $p < 0.001$, $\eta^2_G = 0.16$, as well as main effects of memory score type, F (1, 22) = 17.89, $p < 0.001$, $\eta^2_G = 0.18$, and ROI, F (1, 22) = 17.55, $p < 0.001$, $\eta^2_G = 0.17$ (Figure 9, left panel). For scenes, the same effects were found in the analogous two-way (2x2) ANOVA of memory score (Type x ROI interaction: F [1, 20] = 11.49, $p = 0.003$, $\eta^2_G = 0.16$; Type: F [1, 20] = 26.17, $p < 0.001$, $\eta^2_G = 0.28$; ROI: F [1, 20] = 9.01, $p = 0.007$, $\eta^2_G = 0.06$; Figure 9, right panel). For both stimulus sets, feature-coding regions (i.e., negative FCI-ROIs) had greater feature memory signals than conjunction-coding regions (i.e., positive FCI-ROIs; fribbles: t [22] = 5.59, $p < 0.001$; scenes: t [20] = 3.96, $p < 0.001$). For scenes but not for fribbles, conjunction memory scores were numerically (but not significantly) greater in the positive FCI-ROI (Cohen's D = 0.15) than in the negative FCI-ROI (Cohen's D = -0.07). However, for both stimulus datasets, conjunction memory scores did not significantly differ from zero ($p > 0.05$) in either positive or negative FCI-ROIs.

The Bayes factor analysis again corroborated the frequentist ANOVAs. For both stimulus sets, there was decisive evidence in favor of a model with main effects of ROI and Memory Type, and their interaction, when compared to the null model ($BF_{Fribbles} > 150$; $BF_{Scenes} > 150$) or the alternative nested model with only additive effects ($BF_{Fribbles} > 150$; $BF_{Scenes} = 150$). Additionally, for both stimulus sets, there was very strong evidence in favor of greater feature memory scores in negative FCI-ROIs (i.e., feature-coding regions) than in positive FCI-ROIs (i.e., conjunction-coding regions; $BF_{Fribbles} > 150$; $BF_{Scenes} = 91.03$). Finally, there was a lack of substantial evidence in favor of greater-than-zero conjunction memory scores in positive FCI-ROIs for either fribbles or scenes, but, for scenes only, there was substantial evidence for greater-than-zero feature memory scores in the *positive* FCI-ROI compared to the null (BF = 8.70). This may indicate that the feature memory measure can reflect a familiarity signal for relatively "complex" features in the scene stimuli (i.e., room furniture counts as a "feature", but is conjunctive in nature).

Discussion

In this study, we tested two hypotheses: first, that the cortical locus of recognition memory signals depends on the complexity of the to-be-remembered content; and second, that the location of memory signals tracks the location of the optimal neural code for the representations being retrieved. These hypotheses entail that recognition memory does not always depend on the MTL (contrasting with dominant memory theories, e.g., Brown & Aggleton, 2001; Davachi, 2006; Diana et al., 2007; Eichenbaum et al., 2007; Norman & O'Reilly, 2003; Squire & Zola-Morgan, 1991; Yonelinas, Ranganath, Ekstrom, & Wiltgen, 2019), but can rely on extra-MTL sites when the representations being retrieved are sufficiently

simple or low-dimensional (Cowell et al., 2019; Cowell, Bussey, & Saksida, 2010a; Graham et al., 2010).

To test the first hypothesis – that the cortical locus of recognition memory depends on complexity – we manipulated the to-be-remembered representations in two ways: by contrasting feature memory (simpler) with conjunction memory (more complex), and by comparing fribbles (simpler) with scenes (more complex). We found few differences between fribbles and scenes in terms of neural signals (discussed below), but we found robust differences between feature memory and conjunction memory. Feature memory signals were strongest in posterior visual sites (V1 through LOC) and waned in MTL regions, while conjunction memory signals increased from posterior to anterior ROIs, emerging in PRC. These findings directly support the R-H account of memory, which postulates that each region in the ventral visual-MTL pathway contributes to memory and perception according to whether it contains the representations required for the task (Bussey & Saksida, 2007).

We tested our second hypothesis – that cortical memory signals should coincide with the optimal neural code for the representations being retrieved – by applying a novel MVPA technique (Cowell et al., 2017) to data from the study phase. This technique measures feature- and conjunction-coding, via the FCI, in which negative values indicate preferential coding of features, and positive values preferential coding of conjunctions. In our ventral visual-MTL ROIs, FCIs did not differ for fribbles versus scenes (discussed below), so we collapsed across stimulus sets. An ROI-based analysis of collapsed FCIs revealed robust feature-coding, i.e., negative FCIs, in early visual regions (V1 and V2), transitioning to numerically positive FCIs in MTL regions PHC and PRC. The finding of feature-coding in posterior regions replicates Cowell et al. (2017) and supports the widely held claim that early visual cortex houses simple visual

primitives (Hubel & Wiesel, 1962, 1965; Kamitani & Tong, 2005; Mazer, Vinje, McDermott, Schiller, & Gallant, 2002; Serences, Saproo, Scolari, Ho, & Muftuler, 2009). The near-zero and numerically positive FCIs in MTL regions may reflect hybrid-coding (a mixture of feature- and conjunction-codes) or conjunction-coding. However, we hesitate to conclude this, because no region contained an FCI significantly above zero, and classifier accuracies for these regions were, on average, below the binomial threshold that we defined heuristically as chance performance. FCIs derived from classifiers performing at chance tend toward zero (Cowell et al., 2017), thus the FCIs we observed in MTL areas could simply reflect the absence of either feature or conjunction knowledge. Nonetheless, there was clear correspondence between the sites of feature-coding, which emerged in early visual cortex (V1 and V2) and the strongest feature-memory signals, which were found in early visual ROIs (including V1 and V2). Furthermore, the sole structure demonstrating significantly positive conjunction memory, PRC, was located further anterior, in MTL, and was also found to contain numerically positive FCI values. These findings together support the prediction of anatomical correspondence between memory signals and the optimal code for the to-be-remembered representations.

We also tested the anatomical correspondence prediction more directly in a second analysis. Via a searchlight analysis of the study phase data, we created empirically defined ROIs that demonstrated extremes of conjunction-coding and feature-coding, for fribbles and scenes separately. We then analyzed the corresponding fribble or scene memory test data within these stimulus-specific conjunction- and feature-coding ROIs, expecting feature-coding regions to be engaged for feature memory and conjunction-coding regions for conjunction memory. For both fribbles and scenes, feature-coding ROIs were located primarily in posterior areas, including bilateral LOC and fusiform and lingual gyri. In contrast, conjunction-coding ROIs were localized

to more anterior sites, including many previously implicated in memory, such as anterior cingulate, lateral orbitofrontal cortex, precuneus (Ranganath & Ritchey, 2012) and medial prefrontal cortex (Ranganath & Ritchey, 2012; Restivo, Vetere, Bontempi, & Ammassari-teule, 2009; Rugg & Vilberg, 2013). Within these empirically defined ROIs, we found no evidence of above-zero conjunction memory signals, regardless of whether the FCI used to define the ROI was positive (conjunction-coding) or negative (feature-coding), and regardless of stimulus set. It is possible that the areas most engaged by conjunction memory (i.e., MTL) were excluded from conjunction-coding ROIs in the searchlight analysis because they exhibit lower BOLD signal SNR (Olman et al., 2009), resulting in poor classification performance and near-zero FCIs (as in the ROI-based FCIs). However, feature-coding ROIs did demonstrate greater feature memory signals than conjunction-coding regions, for both fribbles and scenes. Therefore, brain areas containing the optimal code for representing features were indeed preferentially engaged during mnemonic retrieval of features.

Our findings accord with prior studies examining memory for low-level stimulus properties (Cooke et al., 2015; Karanian & Slotnick, 2018; Thakral et al., 2013), which have reported recruitment of extra-MTL brain regions, sometimes as early as V1. However, to our knowledge, this is the first study to manipulate the complexity of the mnemonic material in a recognition memory task to test the prediction that the cortical location of the memory signal shifts with complexity (but see Ross et al., 2018, for a similar manipulation with a recall task). The manipulation of complexity via the class of stimulus was unsuccessful, perhaps because fribbles and scenes differed insufficiently, but the manipulation of complexity via representational requirements – features versus conjunctions – produced robust shifts in memory signal location. Importantly, the measures of feature and conjunction memory captured signals

that were constrained to arise from representations with specific properties. Feature memory was constrained to arise from features because the contrast of New vs. Recombination trials excluded contributions from memory for conjunctions (conjunctions were novel in both), and conjunction memory had to arise from higher-level conjunctions because the contrast of Recombination vs. Old excluded contributions from feature memory (features were familiar in both).

Just as the representational constraints on the feature memory signal ensure that it precisely targets feature-based information, the neural coding measure, FCI, precisely targets the coding of feature-based information. That is, for a brain region to exhibit a negative FCI, it cannot merely respond to features or contain information that discriminates them; it must contain *more* information about features than the conjunctions of those features. Thus, whereas prior methods for examining neural coding have revealed conjunction-coding (e.g., when the representations contain superadditive information about the conjunction of features relative to each feature alone; Baumgartner et al., 2013; Pollmann et al., 2014), they have not probed feature-coding defined as a *preference* for features. Moreover, a search for regions that simply respond to features usually reveals wide swaths of cortex (e.g., Erez et al., 2016; Martin et al., 2018). Because our methods targeted feature memory and feature-coding precisely, we were able to localize these phenomena to posterior cortical sites and confirm their anatomical correspondence.

Unexpectedly, we did not find any differences between the neural representations of fribbles and scenes. We detected no differences in the locations of feature- or conjunction-coding, and no differences in the strength or location of the memory signals. Because fribbles are "simpler" than scenes, comprising a single object rather than an array of objects, we expected

fribbles to engage more posterior sites than scenes. However, two factors may have worked against this outcome.

First, for features, there may be little difference between fribbles and scenes. Although the two stimulus sets were designed to possess features at a mid- versus high-level of complexity (i.e., fribbles differed from one another by mid-level “head”, “body” and “tail” features, whereas scenes differed by complex, high-level features that were sets of whole objects such as beds and tables), these mid- and high-level features were conflated with low-level features. For example, the two values of each mid-level “body” feature of a fribble took on an identical appearance in all fribbles, differing in a consistent, unchanging way in terms of low-level attributes like color and shape. Thus, any fribble features could be discriminated using either mid-level or low-level visual properties. The same was true for scene features – their low-level attributes varied in a consistent way. Thus, it is possible that low-level properties of features (e.g., line orientation or color) were exploited for both fribbles and scenes, recruiting representations in early visual cortex for *both* stimulus sets. Low-level features may have dominated both in feature-classification, determining the sites of feature-coding measured with the FCI, and in the generation of memory signals, determining the sites of feature memory.

Second, for conjunction memory, the predicted anatomical distance between fribble and scene signals was small. Both stimulus sets were expected to elicit conjunction-level signals in the MTL, in adjacent structures (PRC versus PHC/HC). The measure used to assess conjunction memory – the contrast of Recombination with Familiar trials – was a weak signal, in line with the behavioral discrimination being subtle and difficult to make. It is possible that, after applying the spatial smoothing appropriate for detecting subtle changes in BOLD, in the memory score

univariate analysis (Molloy, Meyerand, & Birn, 2014; Triantafyllou, Hoge, & Wald, 2006), the statistical power was insufficient to differentiate adjacent conjunction memory signals.

In sum, the present findings demonstrate that visual regions outside of MTL underpin memory retrieval when the representational demands of a task are sufficiently simple and low-dimensional, as was the case with the features of our fribbles and scenes. Moreover, using the same stimuli, we showed that memory signals for more complex representations – conjunctions of features – reside further anterior, in PRC. Finally, our results demonstrate anatomical correspondence between the regions that optimally represent features via preferential feature-coding and the regions that exhibit a feature memory signal when participants perform retrieval. Taken together, these findings suggest that it is representational content, rather than cognitive function, that constitutes the primary organizing principle in the ventral visual-MTL pathway (Bussey & Saksida, 2002; Cowell et al., 2010b).

References

Abraham, A., Pedregosa, F., Eickenberg, M., & Gervais, P. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*(14), 1–10. https://doi.org/10.3389/fninf.2014.00014

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, *22*, 425–489.

Anzai, A., Peng, X., & Van Essen, D. C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, *10*(10), 1313–1321. https://doi.org/10.1038/nn1975

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2013). The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule. *NeuroImage*, *70*, 37–47. https://doi.org/10.1016/j.neuroimage.2012.12.023

Barry, T. J., Griffith, J. W., De Rossi, S., & Hermans, D. (2014). Meet the fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*, 1–8. https://doi.org/10.3389/fpsyg.2014.00103

Bartko, S. J., Winters, B. D., Cowell, R. A., Saksida, L. M., & Bussey, T. J. (2007). Perceptual functions of perirhinal cortex in rats: zero-delay object recognition and simultaneous oddity discriminations. *The Journal of Neuroscience*, *27*(10), 2548–2559.

https://doi.org/10.1523/JNEUROSCI.5171-06.2007

Bastin, C., Besson, G., Simon, J., Delhaye, E., Geurten, M., Willems, S., & Salmon, E. (2019). An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*, *42*, e281. https://doi.org/10.1017/S0140525X19000621

Baumgartner, F., Hanke, M., Geringswald, F., Zinke, W., Speck, O., & Pollmann, S. (2013). Evidence for feature binding in the superior parietal lobule. *NeuroImage*, *68*, 173–180. https://doi.org/10.1016/j.neuroimage.2012.12.002

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, *2*(1), 51–61. https://doi.org/10.1038/35049064

Buckley, M. J., Booth, M. C., Rolls, E. T., & Gaffan, D. (2001). Selective perceptual impairments after perirhinal cortex ablation. *J Neurosci*, *21*(24), 9824–9836. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11739590&retmode=ref&cmd=prlinks

Bussey, T. J., & Saksida, L. M. (2002). The organization of visual object representations: A connectionist model of effects of lesions in perirhinal cortex. *The European Journal of Neuroscience*, *15*, 355–364. https://doi.org/10.1046/j.0953-816x.2001.01850.x

Bussey, T. J., & Saksida, L. M. (2007). Memory, Perception, and the Ventral Visual-Perirhinal-Hippocampal Stream: Thinking Outside of the Boxes. *Hippocampus*, *17*, 898–908. https://doi.org/10.1002//hipo.20320

Cooke, S. F., Komorowski, R. W., Kaplan, E. S., Gavornik, J. P., & Bear, M. F. (2015). Visual recognition memory, manifest as long-term habituation, requires synaptic plasticity in V1. *Nature Ne*, *18*(2), 262–271. https://doi.org/10.1038/nn.3920

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. *Cognition*, *50*, 41–77. https://doi.org/10.1016/0010-0277(94)90020-5

Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating Concepts from Converging Features in Human Cortex. *Cerebral Cortex*, *25*(9), 2584–2593. https://doi.org/10.1093/cercor/bhu057

Cowell, R. A., Barense, M. D., & Sadil, P. S. (2019). A Roadmap for Understanding Memory: Decomposing Cognitive Processes into Operations and Representations. *ENeuro*, *6*(4).

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2010a). Components of recognition memory: Dissociable cognitive processes or just differences in representational complexity? *Hippocampus*, *20*, 1245–1262. https://doi.org/10.1002/hipo.20865

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2010b). Functional dissociations within the ventral object processing pathway: cognitive modules or a hierarchical continuum? *Journal of Cognitive Neuroscience*, *22*(11), 2460–2479. https://doi.org/10.1162/jocn.2009.21373

Cowell, R. A., Leger, K., & Serences, J. T. (2017). Feature-coding transitions to conjunction-coding with progression through human visual cortex. *Journal of Neurophysiology*, jn.00503.2017. https://doi.org/10.1152/jn.00503.2017

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, *16*(6), 693–700. https://doi.org/10.1016/j.conb.2006.10.012

Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National*

Academy of Sciences, *100*(4), 2157–2162. https://doi.org/10.1073/pnas.0337195100

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386. https://doi.org/10.1016/j.tics.2007.08.001

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Engel, S. A. (2005). Adaptation of Oriented and Unoriented Color-Selective Neurons in Human Visual Areas. *Neuron*, *45*(4), 613–623. https://doi.org/10.1016/j.neuron.2005.01.014

Erez, J., Cusack, R., Kendall, W., & Barense, M. D. (2016). Conjunctive Coding of Complex Object Features. *Cerebral Cortex*, *26*(5), 2271–2282. https://doi.org/10.1093/cercor/bhv081

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Esteban, O., Markiewicz, C. J., DuPre, E., Goncalves, M., Kent, J. D., Ciric, R., … Gorgolewski, K. J. (2018). FMRIPrep. *Software*, Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.852659.

Esteban, O., Markiewicz, C. J., Johnson, H., Ziegler, E., Manhães-Savio, A., Jarecka, D., … Gon, S. (2018). Nipype. *Software*, Zenodo.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, Supplement 1: S102. https://doi.org/https://doi.org/10.1016/S1053-8119(09)70884-5

Gegenfurtner, K. R., Kiper, D. C., & Levitt, J. B. (1997). Functional Properties of Neurons in

Macaque Area V3. *Journal of Neurophysiology*, *77*(4), 1906–1923. https://doi.org/10.1152/jn.1997.77.4.1906

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*(13), 1–10. https://doi.org/10.3389/fninf.2011.00013

Graham, K. S., Barense, M. D., & Lee, A. C. H. (2010). Going beyond LTM in the MTL: A synthesis of neuropsychological and neuroimaging findings on the role of the medial temporal lobe in memory and perception. *Neuropsychologia*, *48*(4), 831–853. https://doi.org/10.1016/j.neuropsychologia.2010.01.001

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106–154. Retrieved from https://doi.org/10.1113/jphysiol.1962.sp006837

Hubel, D. H., & Wiesel, T. N. (1965). Receptive Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat. *Journal of Neurophysiology*, *28*, 229–289.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Johnson, E. N., Hawken, M. J., & Shapley, R. (2008). The orientation selectivity of color-responsive neurons in macaque V1. *Journal of Neuroscience*, *28*(32), 8096–8106. https://doi.org/10.1523/JNEUROSCI.1404-08.2008

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human

brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanwisher, N. (2006). What's in a Face ? *Science*, *311*(February), 617–618.

Karanian, J. M., & Slotnick, S. D. (2018). Confident false memories for spatial location are mediated by V1. *Cognitive Neuroscience*, *9*(3–4), 139–150. https://doi.org/10.1080/17588928.2018.1488244

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., … Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, Nikolaus, Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1167/8.6.88

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV). https://doi.org/10.3389/fpsyg.2013.00863

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76–85.

Lee, A. C. H., Buckley, M. J., Pegman, S. J., Spiers, H., Scahill, V. L., Gaffan, D., … Graham, K. S. (2005). Specialization in the medial temporal lobe for processing of objects and scenes. *Hippocampus*, *15*, 782–797. https://doi.org/10.1002/hipo.20101

Lee, A. C. H., Yeung, L.-K., & Barense, M. D. (2012). The hippocampus and visual perception.

Frontiers in Human Neuroscience, *6*(April), 91. https://doi.org/10.3389/fnhum.2012.00091

MacEvoy, S. P., & Epstein, R. A. (2009). Decoding the Representation of Multiple Simultaneous Objects in Human Occipitotemporal Cortex. *Current Biology*, *19*(11), 943–947. https://doi.org/10.1016/j.cub.2009.04.020

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L. Y., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *ELife*, *7*, 1–29. https://doi.org/10.7554/elife.31873

Mazer, J. A., Vinje, W. E., McDermott, J., Schiller, P. H., & Gallant, J. L. (2002). Spatial frequency and orientation tuning dynamics in area V1. *Proceedings of the National Academy of Sciences*, *99*(3), 1645–1650. https://doi.org/10.1073/pnas.022638499

Molloy, E. K., Meyerand, M. E., & Birn, R. M. (2014). The influence of spatial resolution and smoothing on the detectability of resting-state and task fMRI. *NeuroImage*, *86*(5), 221–230. https://doi.org/10.1016/j.neuroimage.2013.09.001

Morey, R. D., Rouder, J. N., & Jamil, X. (2015). BayesFactor: Computation of Bayes factors for common designs.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Nadel, L., & Peterson, M. A. (2013). The hippocampus: Part of an interactive posterior representational system spanning perceptual and memorial systems. *Journal of Experimental Psychology: General*, *142*(4), 1242–1254. https://doi.org/10.1037/a0033690

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological*

Review, *110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Olman, C. A., Davachi, L., & Inati, S. (2009). Distortion and signal loss in medial temporal lobe. *PLoS ONE*, *4*(12). https://doi.org/10.1371/journal.pone.0008160

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*(July), 1–27. https://doi.org/10.3389/fninf.2016.00027

Palombo, D. J., Keane, M. M., & Verfaellie, M. (2015). The medial temporal lobes are critical for reward-based decision making under conditions that promote episodic future thinking. *Hippocampus*, *25*(3), 345–353. https://doi.org/10.1002/hipo.22376

Pollmann, S., Zinke, W., Baumgartner, F., Geringswald, F., & Hanke, M. (2014). The right temporo-parietal junction contributes to visual feature binding. *NeuroImage*, *101*, 289–297. https://doi.org/10.1016/j.neuroimage.2014.07.021

Raftery, A. E. (1995). *Bayesian Model Selection in Social Research*. *Source: Sociological Methodology* (Vol. 25).

Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, *13*(10), 713–726. https://doi.org/10.1038/nrn3338

Restivo, L., Vetere, G., Bontempi, B., & Ammassari-teule, M. (2009). The Formation of Recent and Remote Memory Is Associated with Time-Dependent Formation of Dendritic Spines in the Hippocampus and Anterior Cingulate Cortex. *Journal of Neuroscience*, *29*(25), 8206–8214. https://doi.org/10.1523/JNEUROSCI.0966-09.2009

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Ritchey, M., Montchal, M. E., Yonelinas, A. P., & Ranganath, C. (2015). Delay-dependent contributions of medial temporal lobe regions to episodic memory retrieval. *ELife*, *4*, 1–19. https://doi.org/10.7554/eLife.05025

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Ross, D. A., Sadil, P., Wilson, D. M., & Cowell, R. A. (2018). Hippocampal engagement during recall depends on memory content. *Cerebral Cortex*, *28*(8), 2685–2698. https://doi.org/10.1093/cercor/bhx147

Rugg, M. D., & Vilberg, K. L. (2013). Brain networks underlying episodic memory retrieval. *Current Opinion in Neurobiology*, *23*(2), 255–260. https://doi.org/10.1016/j.conb.2012.11.005

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The Necessity of the Medial Temporal Lobe for Statistical Learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Serences, J. T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L. T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, *44*, 223–231. https://doi.org/10.1016/j.neuroimage.2008.07.043

Seymour, K., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2009). The Coding of Color, Motion, and Their Conjunction in the Human Visual Cortex. *Current Biology*, *19*(3), 177–183. https://doi.org/10.1016/j.cub.2008.12.050

Seymour, K., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*(8), 1946–1954.

https://doi.org/10.1093/cercor/bhp265

Shimamura, A. P. (2011). Episodic retrieval and the cortical binding of relational activity. *Cognitive, Affective, & Behavioral Neuroscience*, *11*, 277–291. https://doi.org/10.3758/s13415-011-0031-4

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-berg, H., … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*(Suppl. 1), S208-219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*, 1380–1386. https://doi.org/10.1126/science.1896849

Staresina, B. P., Cooper, E., & Henson, R. N. (2013). Reversible Information Flow across the Medial Temporal Lobe: The Hippocampus Links Cortical Modules during Memory Retrieval. *Journal of Neuroscience*, *33*(35), 14184–14192. https://doi.org/10.1523/JNEUROSCI.1987-13.2013

Stark, C. E., & Squire, L. R. (2001). When zero is not zero: the problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(22), 12760–12766. https://doi.org/10.1073/pnas.221462998

Thakral, P. P., Madore, K. P., Addis, D. R., & Schacter, D. L. (2020). Reinstatement of Event Details during Episodic Simulation in the Hippocampus. *Cerebral Cortex*, *30*(4), 2321–2337. https://doi.org/10.1093/cercor/bhz242

Thakral, P. P., Slotnick, S. D., & Schacter, D. L. (2013). Conscious processing during retrieval can occur in early andlate visual regions. *Neuropsychologia*, *51*(3), 482–487. https://doi.org/10.1016/j.neuropsychologia.2012.11.020.Conscious

Thakral, P. P., Wang, T. H., & Rugg, M. D. (2017). Decoding the content of recollection within the core recollection network and beyond. *Cortex*, *91*, 101–113. https://doi.org/10.1016/j.cortex.2016.12.011.Decoding

Triantafyllou, C., Hoge, R. D., & Wald, L. L. (2006). Effect of spatial smoothing on physiological noise in high-resolution fMRI. *NeuroImage*, *32*(2), 551–557. https://doi.org/10.1016/j.neuroimage.2006.04.182

Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie Zheng, Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

van den Honert, R. N., McCarthy, G., & Johnson, M. K. (2017). Holistic versus feature-based binding in the medial temporal lobe. *Cortex*, *91*, 56–66. https://doi.org/10.1016/j.cortex.2017.01.011

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wilkinson, F., James, T. W., Wilson, H. R., Gati, J. S., Menon, R. S., & Goodale, M. A. (2000). An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings. *Current Biology*, *10*, 1455–1458. https://doi.org/10.1016/s0960-9822(00)00800-9

Williams, P. (1998). Representational organization of multiple examplars of object categories, 1–21.

Yonelinas, A. P., Ranganath, C., Ekstrom, A., & Wiltgen, B. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*.

https://doi.org/10.1038/s41583-019-0150-4

Zhang, J., Liu, J., & Xu, Y. (2015). Neural Decoding Reveals Impaired Face Configural Processing in the Right Fusiform Face Area of Individuals with Developmental Prosopagnosia. *Journal of Neuroscience*, *35*(4), 1539–1548. https://doi.org/10.1523/JNEUROSCI.2646-14.2015

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. https://doi.org/10.1109/42.906424

Figure Legends

Figure 1. Fribble Stimulus Set Examples. *Left*: a conjunctive fribble (top) and "tail", "body" and "head" features (bottom, left to right); *Right*: A family containing eight unique conjunctive fribbles, constructed from three binary features.

Figure 2. Scene Stimulus Set Examples. *A*: a conjunctive scene (top) and room furniture, color and shape features (bottom, left to right); *B*: A family containing eight unique conjunctive scenes composed using three binary room features.

Figure 3. Scene Stimulus Set Recombination Examples. *Top*: counterbalanced across participants, two families from the scene stimulus set were designated to be studied (i.e., presented in the first scan session). In this example, a Recombination stimulus is created by combining the color feature from the first family with the room shape and furniture features from the second family. All potential 1 + 2 feature combinations within a given family pairing yielded 48 Recombination stimuli; *Bottom*: the three mnemonic stimulus classes differed on the basis of whether features and the conjunction of features were studied prior to the memory test.

Figure 4. Experimental Design. During a study scan, participants indicated if the current image was identical to the image immediately before (i.e., 1-back repetition task). During a test scan, participants distinguished between Familiar, Novel, and Recombination stimuli. Both study and test scans featured null trials in which participants indicated when a fixation cross dimmed. Participants completed ten study runs and then exited the scanner for a self-paced break. After re-entering the scanner, participants completed six test runs. Scene stimuli are shown here, but the protocol was identical for the fribble scan sessions, completed on a separate day.

Figure 5. Study Phase Classifier Accuracies by Stimulus Set. Classifier accuracies were averaged over the two studied families from a stimulus set, for each participant. The three separate two-way feature-classifiers were also averaged (striped bars) for ease of comparison with the single eight-way conjunction-classifier (solid white bars). Accuracy levels indicating classifier performance above chance were determined heuristically with a binomial test, uncorrected for multiple comparisons (see text). This accuracy level was 0.5625 for feature-classifiers and 0.1628 for the conjunction-classifier (dashed lines). Solid lines indicate actual chance accuracy for feature-classifiers (0.125) and the conjunction classifier (0.5). Error bars are within-subject SEM. Both feature- and conjunction-classifier accuracies were not significantly greater than binomial chance performance for MTL structures, regardless of stimulus set (fribbles [left] versus scenes [right]).

Figure 6. Study Phase Feature-Conjunction Indexes (FCIs). FCIs were derived by taking the natural log of the ratio of conjunction classifier accuracy (the *empirically observed* conjunction accuracy) to the product of three feature classifier accuracies (the *predicted* conjunction accuracy; see Methods). A negative FCI reflects feature-coding, while a positive FCI reflects conjunction-coding. FCIs were first averaged across the two studied families from a given stimulus set, and then across the two stimulus sets for each participant. White bars show group means; plotted points show individual participants, where each unique color-marker combination corresponds to the same participant across ROIs. Error bars are within-subject SEM. Feature-coding (negative FCIs) appeared to transition to a different form of coding — either a "hybrid" combination of feature- and conjunction-coding (zero FCIs) or conjunction-coding (positive FCIs) — when moving anteriorly along the VVS-MTL pathway. MTL FCI values should be interpreted with caution; classifier accuracies were not above chance, meaning zero FCIs could indicate either "hybrid" coding or an absence of feature- and conjunction-knowledge altogether.

Figure 7. Test Phase Memory Scores. Memory scores were derived by calculating effect size, as measured by Cohen's D, for the contrast of Novel and Recombination trials (i.e., Feature Memory) and for the contrast of Recombination and Familiar trials (i.e., Conjunction Memory). Memory scores were averaged across fribbles and scenes for each participant. Striped and solid white bars show group means; plotted points show individual participants, where each unique color-marker combination corresponds to the same individual across ROIs. Error bars are within-subject SEM. Signals for feature memory were greatest in posterior areas along the VVS-MTL pathway, whereas signals for conjunction memory were greatest in MTL areas.

Figure 8. FCI-ROIs Defined via Searchlight Analysis of Study Phase Data. For each stimulus set and participant separately, FCI was recorded at each center voxel of a spherical ROI swept throughout the brain. Voxels that demonstrated FCIs significantly less than or greater than zero via a group-level *t*-test were grouped into negative and positive FCI-ROIs, respectively. Negative FCI-ROIs (blue) indicate regions that demonstrated extremes of feature-coding and positive FCI-ROIs (red) indicate regions that demonstrated extremes of conjunction-coding. For both fribbles (*A*) and scenes (*B*), top and bottom rows show the same FCI-ROIs, but from different views to highlight positive and negative areas.

Figure 9. Test Phase Memory Scores for FCI-defined ROIs. For both fribbles (left) and scenes (right), negative FCI-ROIs included brain areas that contain more information about individual features than conjunctions of features (i.e., feature-coding), based on study phase data collected while participants viewed the stimuli. In contrast, positive FCI-ROIs included brain areas that contain more information about conjunctions of features than about individual features (i.e., conjunction-coding). Memory scores for stimulus-specific, FCI-defined ROIs were derived from corresponding fribble/scene memory test phase data in an identical manner to the memory scores for *a priori* ROIs. Feature memory was measured by the effect size (Cohen's D) of the contrast between Novel and Recombination trials; conjunction memory was measured by the effect size of the contrast between Recombination and Familiar trials. Error bars are within-subject SEM. For both fribbles and scenes, signals for feature memory were greater in feature-coding ROIs than in conjunction-coding ROIs.

Figure 6-1. Study Phase Feature-Conjunction Indexes (FCIs) by Stimulus Set Type. FCIs were derived by taking the natural log of the ratio of conjunction classifier accuracy (*empirically observed* conjunction accuracy) to the product of three feature classifier accuracies (*predicted* conjunction accuracy; see Methods). A negative FCI reflects feature-coding, while a positive FCI reflects conjunction-coding. FCIs were averaged across the two studied families from a given stimulus set for each subject. Error bars are within-subject SEM. We found no evidence that FCI values differed between fribbles (*left*) and scenes (*right*), so FCI values were averaged across stimulus sets for all further analyses.

Figure 7-1. Test Phase Memory Scores by Stimulus Set Type. Memory scores (*top row*) were derived by calculating the effect size, measured by Cohen's D, for the contrast of Novel versus Recombination trials (Feature Memory) and for the contrast of Recombination versus Familiar trials (Conjunction Memory). Specifically, for Feature Memory, beta weight estimates from the GLM for correct Recombination trials were subtracted from beta weight estimates for correct Novel trials. Similarly, for Conjunction Memory, beta weight estimates for correct Familiar trials were subtracted from beta weight estimates for correct Recombination trials. Here, we display the mean of the beta weights used in the calculation of memory scores for fribbles (*bottom left*) and scenes (*bottom right*), separately. Error bars are within-subject SEM. There was insufficient evidence of stimulus set-related differences in memory scores to justify treating the fribbles (*top left*) and scenes (*top right*) separately in this analysis of memory. Consequently, memory scores were averaged across stimulus sets for all further analyses.

Tables

Table 1: Extreme Outliers Omitted from Statistical Analyses

Analysis	Stimulus	Subject	Level
Behavioral Memory performance	Scenes	3	Familiar vs. Novel (I)
			Recombination vs. Novel
FCI: 2x7 ANOVA	Fribbles	11	V1
Memory score: 2x7x2 ANOVA	Scenes	8	PRC Conj. Memory
		23	PRC Feat. Memory
Memory score: 7x2 ANOVA	Collapsed	8	PRC Conj. Memory
		13	PRC Feat. Memory
		18	LOC Feat. Memory
		23	PRC Feat., Memory
			PRC Conj. Memory
FCI-ROI Memory score: 2x2 ANOVA	Scenes	8	Neg. Conj. Memory
		23	Pos. Conj. Memory

Note: Values either three times the interquartile range above the 75th quantile or below the 25th quantile were identified as extreme outliers. For repeated-measures analyses, all observations for a subject were omitted, rather than just the extreme outlier point.

Table 2: Discriminability of Mnemonic Stimulus Classes for Fribbles and Scenes

Comparison Condition	Fribble	Scene	Trial Types	Hit/Correct Rejection	Miss/False Alarm
Familiar vs. Novel (I)	4.47	4.44	Familiar	"Familiar" & "Recombination"	"Novel"
			Novel	"Novel"	"Familiar" & "Recombination"
Familiar vs. Novel (II)	3.28	3.59	Familiar	"Familiar"	"Novel" & "Recombination"
			Novel	"Novel" & "Recombination"	"Familiar"
Familiar vs. Recombination	2.32	2.47	Familiar	"Familiar"	"Recombination" & "Novel"
			Recombination	"Recombination" & "Novel"	"Familiar"
Recombination vs. Novel	4.37	4.22	Recombination	"Recombination" & "Familiar"	"Novel"
			Novel	"Novel"	"Recombination" & "Familiar"

Note: The *d'* statistics shown here are for each possible pairwise comparison. In each comparison, only two of the three trial types were compared, while any trials corresponding to the third trial type were removed from the analysis. (Trial types included are indicated in the "Comparison Condition" and "Trial Types" columns). However, because there were three possible *responses* on a given trial, responses that referred to the third outstanding trial type (i.e., the trial type *not* shown in the "Comparison Condition" column) were binned as a hit, miss, false alarm or correct rejection in a manner that depended on the comparison being made. The exact treatment of responses referring to the outstanding trial type is indicated in the "Hit/Correct Rejection" or "Miss/False Alarm" columns. For example, in the first comparison, Familiar vs. Novel (I), only Familiar and Novel trials were included, but a participant could still respond "Recombination". If a participant responded "Recombination" to a Familiar trial, it was counted as if they had said "Familiar", and contributed to the hit rate; if a participant responded "Recombination" to a Novel trial, it was again counted as if they had said "Familiar" and thus contributed to the false alarm rate. In the second comparison, Familiar vs. Novel (II), these same two trial types were considered again, but with an alternative treatment of "Recombination" responses. A "Recombination" response to a Familiar trial was now counted as a miss, and a "Recombination" response to a Novel trial as a correct rejection. We computed only one version of the "Familiar vs. Recombination" and "Recombination vs. Novel" comparisons, because there was only one sensible way to treat the outstanding response type, for each of these.

Table 3: Ten Most Positive and Negative Voxels within Fribble FCI-ROIs

FCI-ROI	*t*-Statistic	Coordinates	Harvard-Oxford brain area
Positive	4.46	43, 36, 49	rh-precuneus
	4.45	39, 76, 29	rh-medialorbitofrontal
	4.12	36, 72, 50	lh-superiorfrontal
	3.93	36, 72, 51	lh-superiorfrontal
	3.88	35, 34, 47	lh-precuneus
	3.88	55, 62, 29	rh-parstriangularis
	3.87	40, 76, 29	rh-medialorbitofrontal
	3.78	39, 75, 29	rh-medialorbitofrontal
	3.77	39, 72, 33	rh-superiorfrontal
	3.71	54, 61, 29	rh-insula
Negative	-6.98	44, 14, 27	rh-lateraloccipital
	-6.16	19, 18, 29	lh-lateraloccipital
	-5.96	22, 20, 28	lh-lateraloccipital
	-5.82	51, 23, 23	rh-fusiform
	-5.73	22, 20, 29	lh-lateraloccipital
	-5.63	19, 19, 29	lh-lateraloccipital
	-5.63	51, 23, 24	rh-lateraloccipital
	-5.60	21, 21, 28	lh-lateraloccipital
	-5.59	52, 23, 24	rh-fusiform
	-5.55	52, 22, 24	lh-lateraloccipital

Note: Coordinates are within MNI152NLin2009cAsym standard space. Abbreviations include *rh* for right hemisphere and *lh* for left hemisphere.

Table 4: Ten Most Positive and Negative Voxels within Scene FCI-ROIs

FCI-ROI	*t*-Statistic	Coordinates	Harvard-Oxford brain area
Positive	4.19	22, 55, 31	lh-insula
	4.10	21, 56, 31	lh-pars opercularis
	4.01	22, 54, 31	lh-insula
	3.95	38, 70, 36	lh-rostralanteriorcingulate
	3.92	22, 56, 31	lh-insula
	3.91	39, 70, 39	rh-superiorfrontal
	3.90	21, 55, 31	lh-insula
	3.87	47, 28, 59	rh-superiorparietal
	3.85	51, 72, 27	rh-lateralorbitofrontal
	3.83	39, 70, 38	rh-superiorfrontal
Negative	-8.15	25, 16, 37	lh-inferiorparietal
	-7.41	25, 17, 37	lh-inferiorparietal
	-7.17	25, 17, 36	lh-inferiorparietal
	-6.80	25, 16, 36	lh-inferiorparietal
	-6.78	30, 22, 29	lh-lingual
	-6.66	26, 16, 36	lh-lateraloccipital
	-6.56	24, 16, 37	lh-lateraloccipital
	-6.51	25, 15, 37	lh-inferiorparietal
	-6.36	28, 22, 30	lh-fusiform
	-6.34	26, 16, 37	lh-lateraloccipital

Note: Coordinates are within MNI152NLin2009cAsym standard space. Abbreviations include *rh* for right hemisphere and *lh* for left hemisphere.

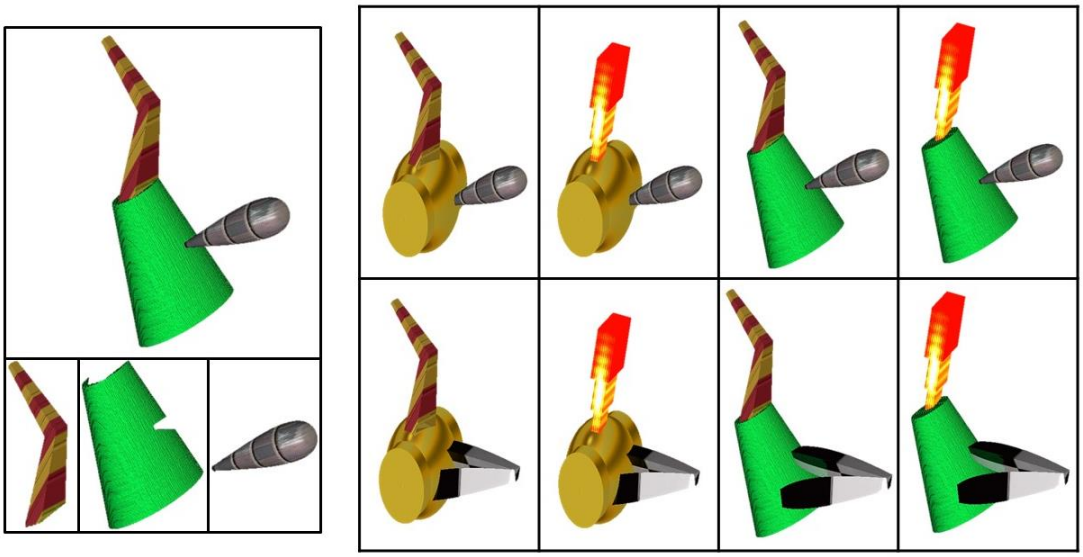
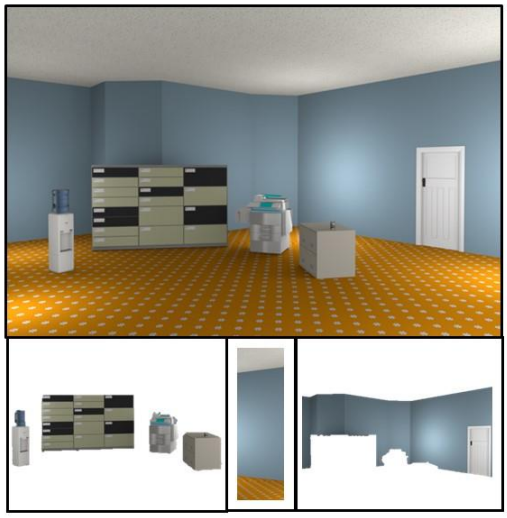


Figure 1. Fribble Stimulus Set Examples. *Left*: a conjunctive fribble (top) and "tail", "body" and "head" features (bottom, left to right); *Right*: A family containing eight unique conjunctive fribbles, constructed from three binary features.

A



B

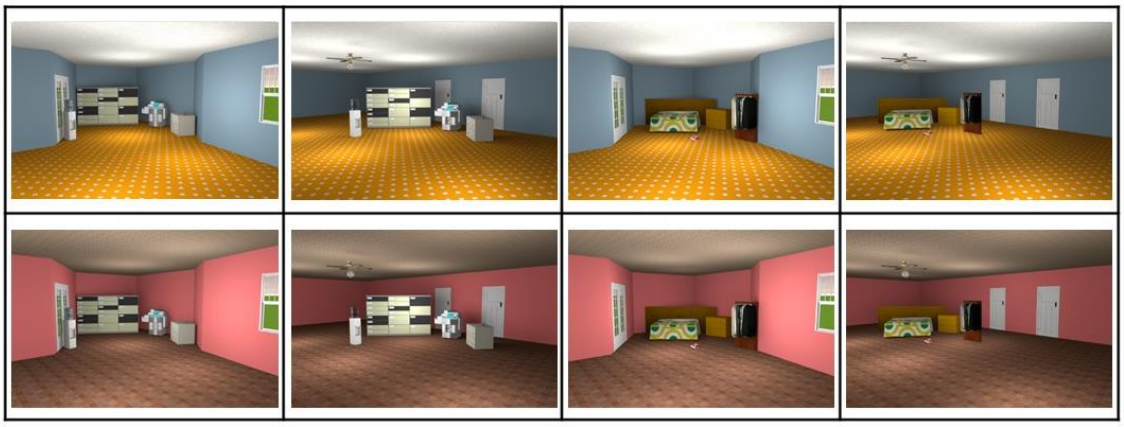


Figure 2. Scene Stimulus Set Examples. *A*: a conjunctive scene (top) and room furniture, color and shape features (bottom, left to right); *B*: A family containing eight unique conjunctive scenes composed using three binary room features.

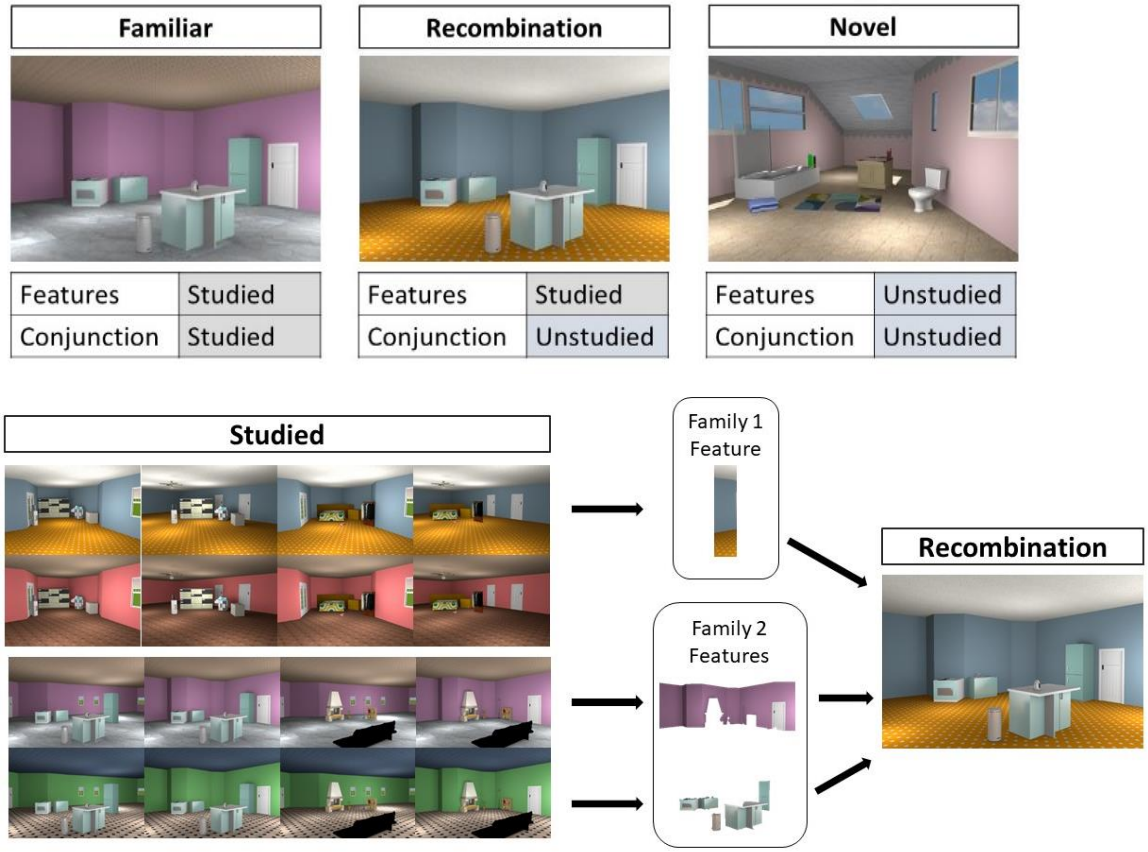


Figure 3. Scene Stimulus Set Recombination Examples. *Top*: counterbalanced across participants, two families from the scene stimulus set were designated to be studied (i.e., presented in the first scan session). In this example, a Recombination stimulus is created by combining the color feature from the first family with the room shape and furniture features from the second family. All potential 1 + 2 feature combinations within a given family pairing yielded 48 Recombination stimuli; *Bottom*: the three mnemonic stimulus classes differed on the basis of whether features and the conjunction of features were studied prior to the memory test.

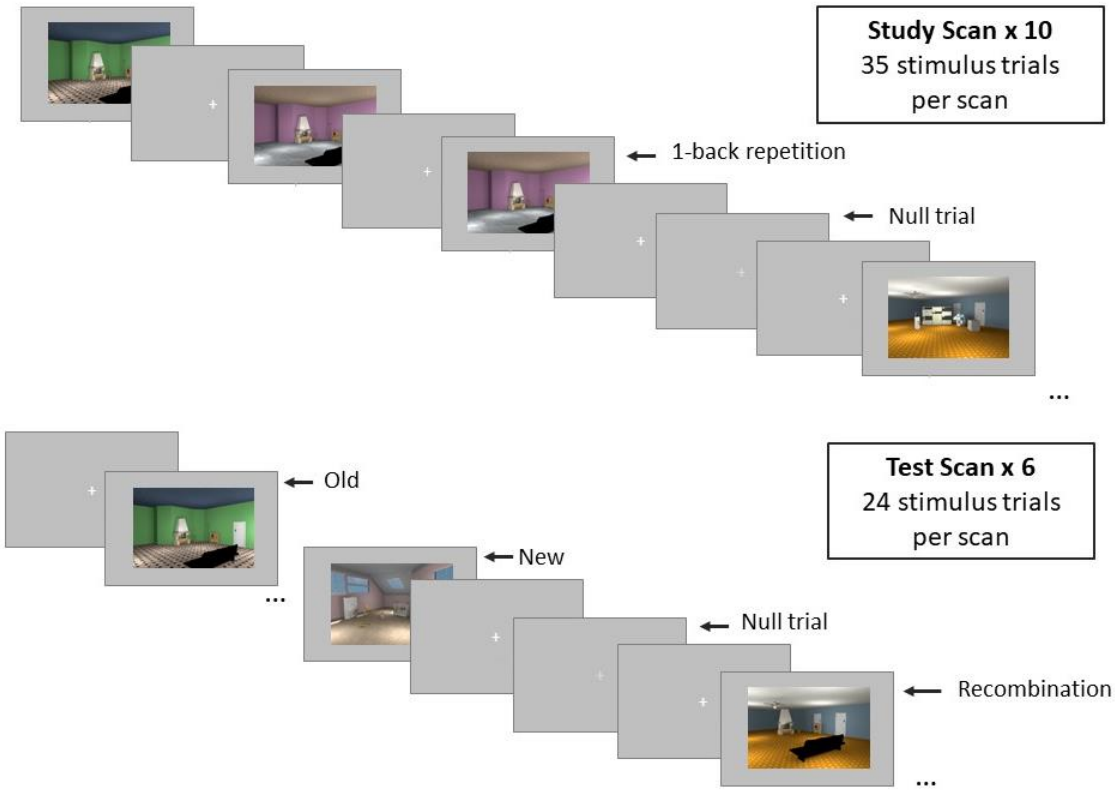


Figure 4. Experimental Design. During a study scan, participants indicated if the current image was identical to the image immediately before (i.e., 1-back repetition task). During a test scan, participants distinguished between Familiar, Novel, and Recombination stimuli. Both study and test scans featured null trials in which participants indicated when a fixation cross dimmed. Participants completed ten study runs and then exited the scanner for a self-paced break. After re-entering the scanner, participants completed six test runs. Scene stimuli are shown here, but the protocol was identical for the fribble scan sessions, completed on a separate day.

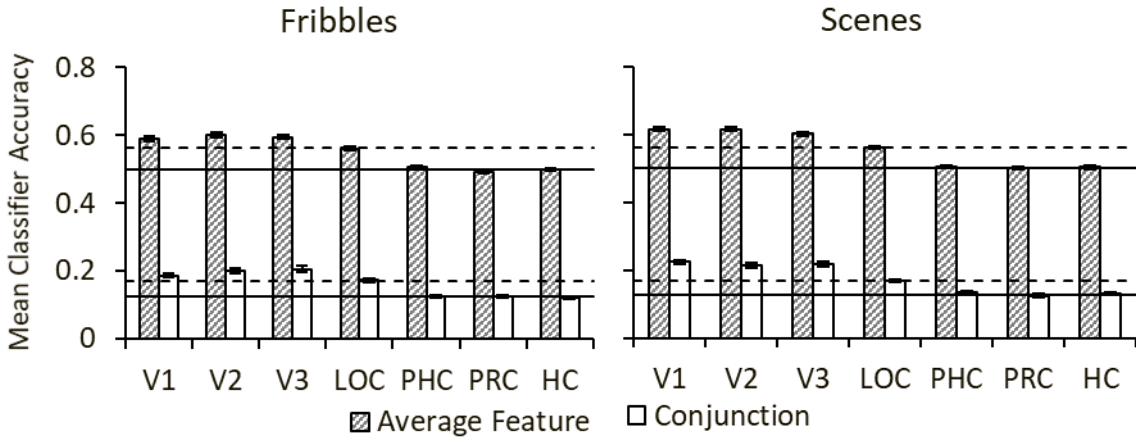


Figure 5. Study Phase Classifier Accuracies by Stimulus Set. Classifier accuracies were averaged over the two studied families from a stimulus set, for each participant. The three separate two-way feature-classifiers were also averaged (striped bars) for ease of comparison with the single eight-way conjunction-classifier (solid white bars). Accuracy levels indicating classifier performance above chance were determined heuristically with a binomial test, uncorrected for multiple comparisons (see text). This accuracy level was 0.5625 for feature-classifiers and 0.1628 for the conjunction-classifier (dashed lines). Solid lines indicate actual chance accuracy for feature-classifiers (0.125) and the conjunction classifier (0.5). Error bars are within-subject SEM. Both feature- and conjunction-classifier accuracies were not significantly greater than binomial chance performance for MTL structures, regardless of stimulus set (fribbles [left] versus scenes [right]).

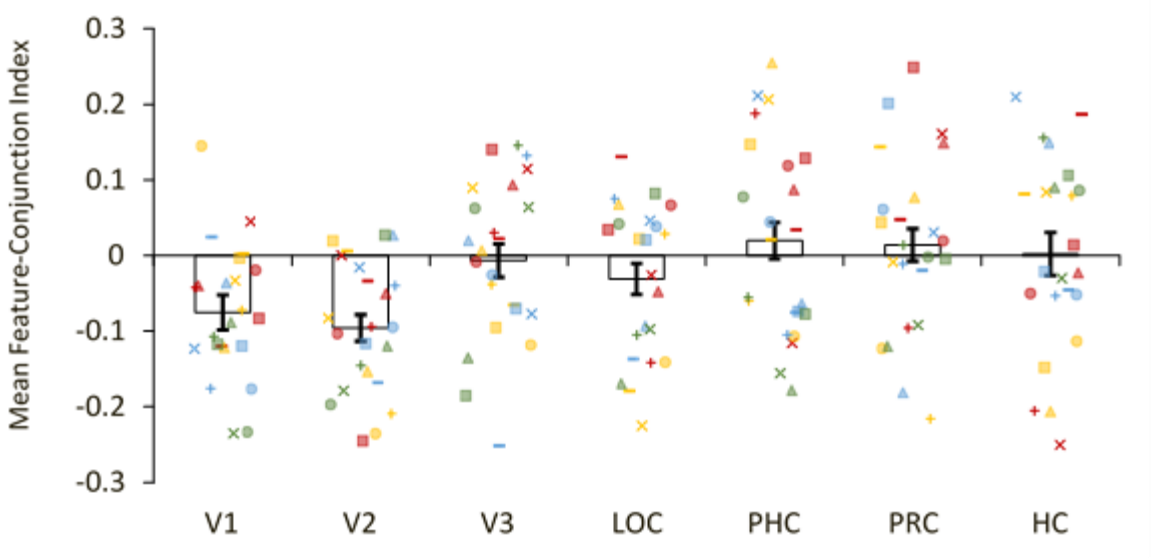


Figure 6. **Study Phase Feature-Conjunction Indexes (FCIs).** FCIs were derived by taking the natural log of the ratio of conjunction classifier accuracy (the *empirically observed* conjunction accuracy) to the product of three feature classifier accuracies (the *predicted* conjunction accuracy; see Methods). A negative FCI reflects feature-coding, while a positive FCI reflects conjunction-coding. FCIs were first averaged across the two studied families from a given stimulus set, and then across the two stimulus sets for each participant. White bars show group means; plotted points show individual participants, where each unique color-marker combination corresponds to the same participant across ROIs. Error bars are within-subject SEM. Feature-coding (negative FCIs) appeared to transition to a different form of coding — either a "hybrid" combination of feature- and conjunction-coding (zero FCIs) or conjunction-coding (positive FCIs) — when moving anteriorly along the VVS-MTL pathway. MTL FCI values should be interpreted with caution; classifier accuracies were not above chance, meaning zero FCIs could indicate either "hybrid" coding or an absence of feature- and conjunction-knowledge altogether.

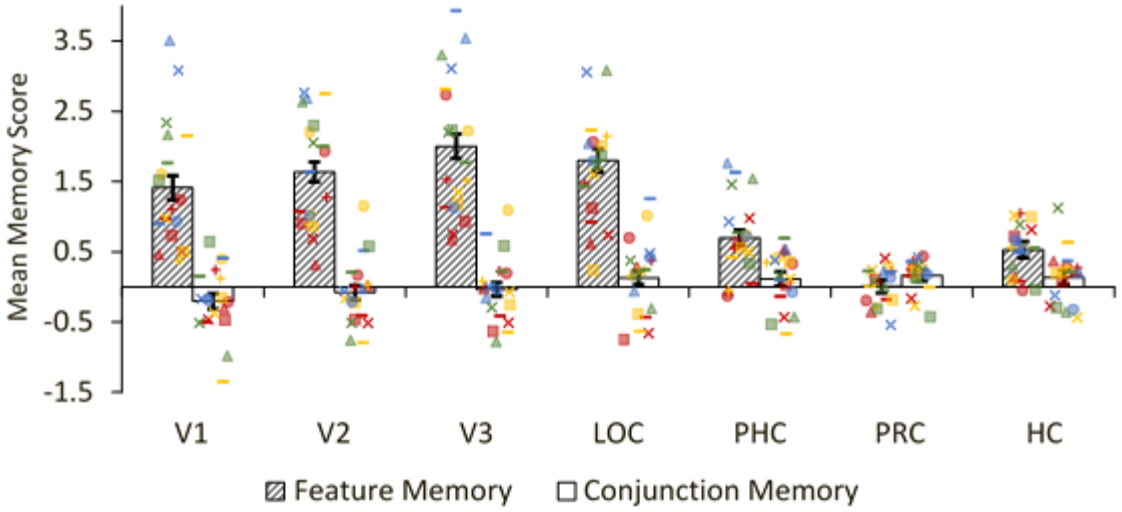


Figure 7. Test Phase Memory Scores. Memory scores were derived by calculating effect size, as measured by Cohen's D, for the contrast of Novel and Recombination trials (i.e., Feature Memory) and for the contrast of Recombination and Familiar trials (i.e., Conjunction Memory). Memory scores were averaged across fribbles and scenes for each participant. Striped and solid white bars show group means; plotted points show individual participants, where each unique color-marker combination corresponds to the same individual across ROIs. Error bars are within-subject SEM. Signals for feature memory were greatest in posterior areas along the VVS-MTL pathway, whereas signals for conjunction memory were greatest in MTL areas.

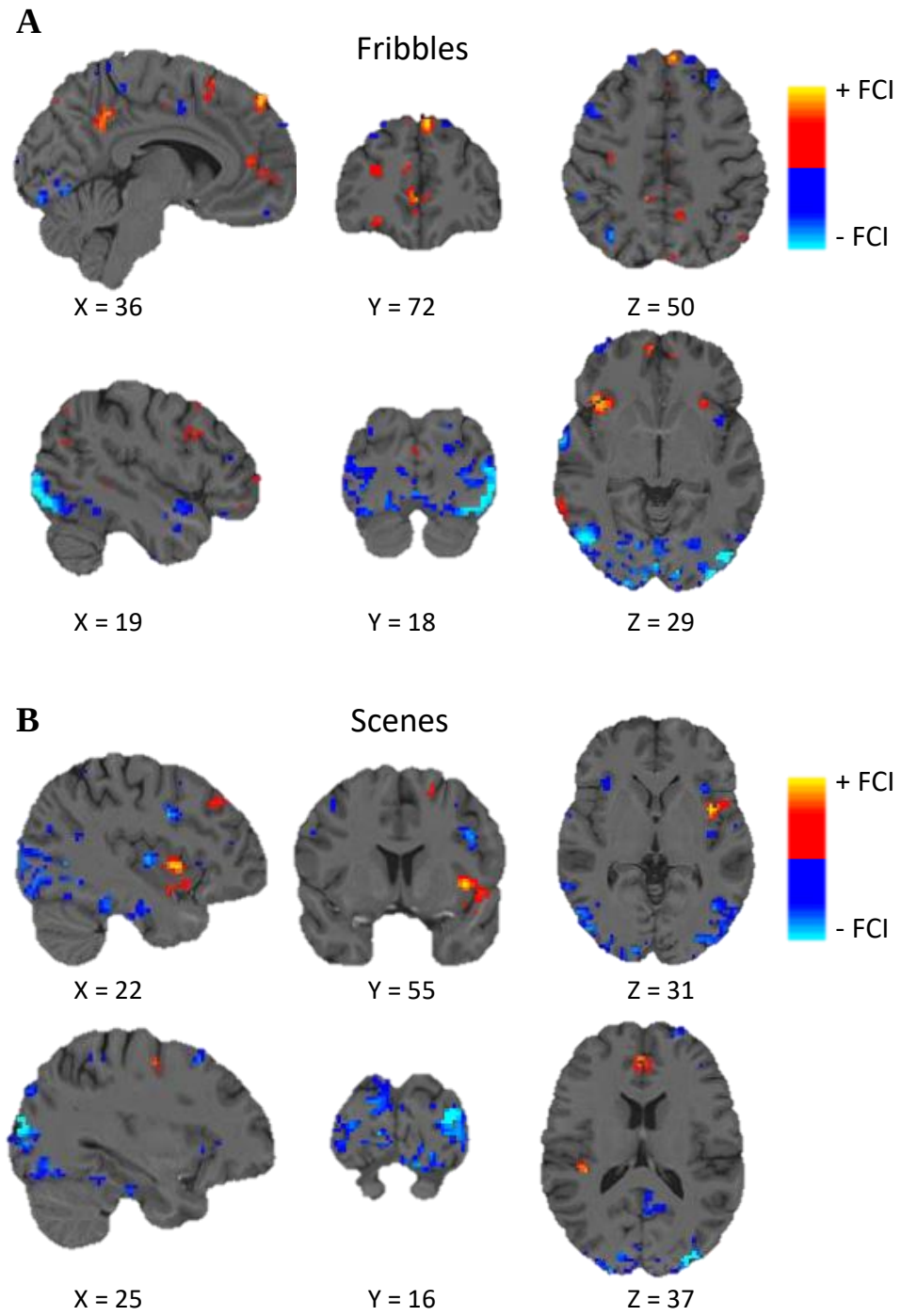


Figure 8. FCI-ROIs Defined via Searchlight Analysis of Study Phase Data. For each stimulus set and participant separately, FCI was recorded at each center voxel of a spherical ROI swept throughout the brain. Voxels that demonstrated FCIs significantly less than or greater than zero via a group-level *t*-test were grouped into negative and positive FCI-ROIs, respectively. Negative FCI-ROIs (blue) indicate regions that demonstrated extremes of feature-coding and positive FCI-ROIs (red) indicate regions that demonstrated extremes of conjunction-coding. For both fribbles (*A*) and scenes (*B*), top and bottom rows show the same FCI-ROIs, but from different views to highlight positive and negative areas.

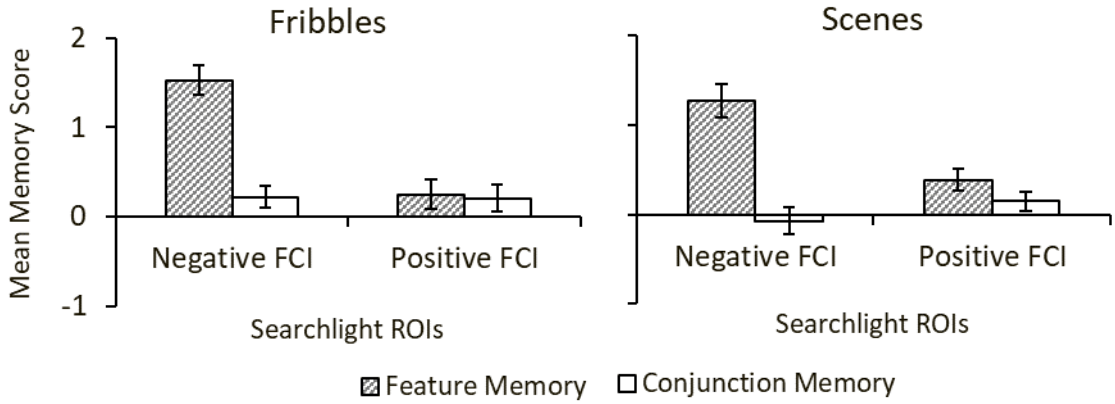


Figure 9. Test Phase Memory Scores for FCI-defined ROIs. For both fribbles (left) and scenes (right), negative FCI-ROIs included brain areas that contain more information about individual features than conjunctions of features (i.e., feature-coding), based on study phase data collected while participants viewed the stimuli. In contrast, positive FCI-ROIs included brain areas that contain more information about conjunctions of features than about individual features (i.e., conjunction-coding). Memory scores for stimulus-specific, FCI-defined ROIs were derived from corresponding fribble/scene memory test phase data in an identical manner to the memory scores for *a priori* ROIs. Feature memory was measured by the effect size (Cohen's D) of the contrast between Novel and Recombination trials; conjunction memory was measured by the effect size of the contrast between Recombination and Familiar trials. Error bars are within-subject SEM. For both fribbles and scenes, signals for feature memory were greater in feature-coding ROIs than in conjunction-coding ROIs.

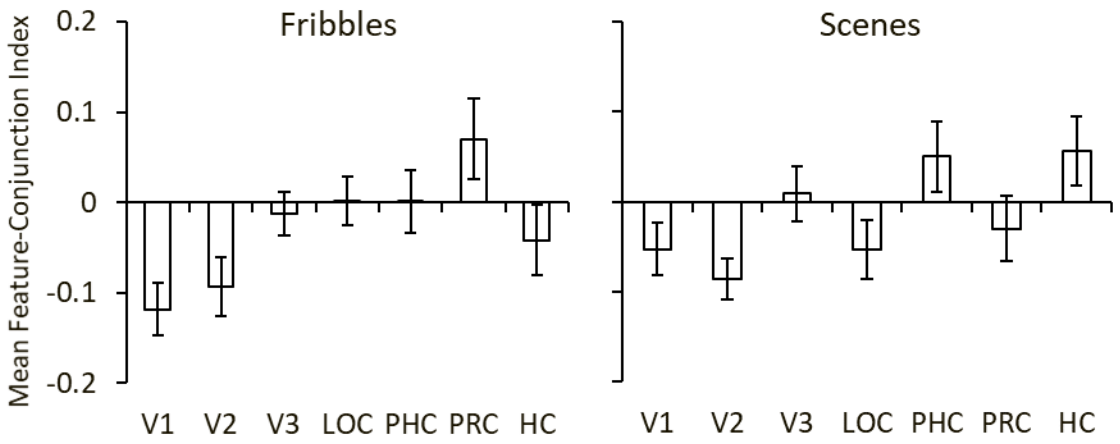


Figure 6-1. Study Phase Feature-Conjunction Indexes (FCIs) by Stimulus Set Type. FCIs were derived by taking the natural log of the ratio of conjunction classifier accuracy (*empirically observed* conjunction accuracy) to the product of three feature classifier accuracies (*predicted* conjunction accuracy; see Methods). A negative FCI reflects feature-coding, while a positive FCI reflects conjunction-coding. FCIs were averaged across the two studied families from a given stimulus set for each subject. Error bars are within-subject SEM. We found no evidence that FCI values differed between fribbles (*left*) and scenes (*right*), so FCI values were averaged across stimulus sets for all further analyses.

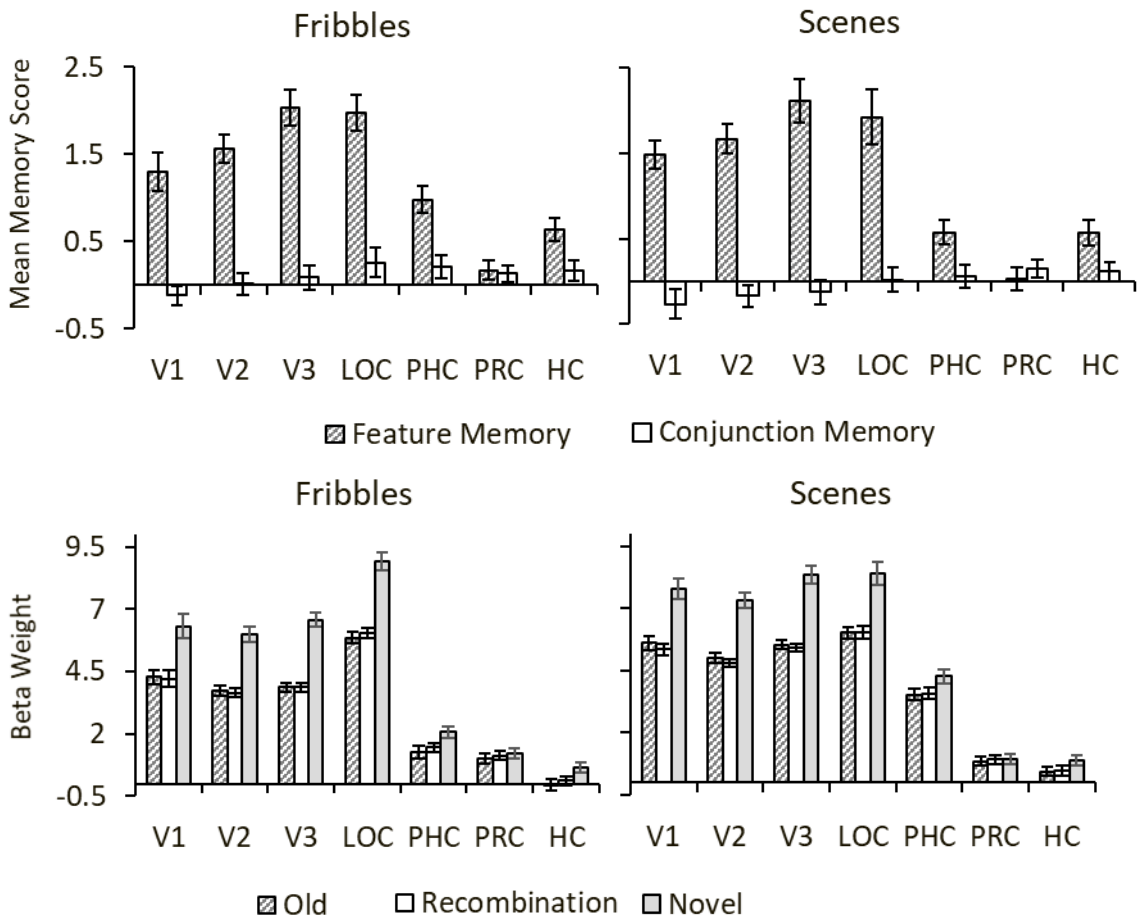


Figure 7-1. Test Phase Memory Scores by Stimulus Set Type. Memory scores (*top row*) were derived by calculating the effect size, measured by Cohen's D, for the contrast of Novel versus Recombination trials (Feature Memory) and for the contrast of Recombination versus Familiar trials (Conjunction Memory). Specifically, for Feature Memory, beta weight estimates from the GLM for correct Recombination trials were subtracted from beta weight estimates for correct Novel trials. Similarly, for Conjunction Memory, beta weight estimates for correct Familiar trials were subtracted from beta weight estimates for correct Recombination trials. Here, we display the mean of the beta weights used in the calculation of memory scores for fribbles (*bottom left*) and scenes (*bottom right*), separately. Error bars are within-subject SEM. There was insufficient evidence of stimulus set-related differences in memory scores to justify treating the fribbles (*top left*) and scenes (*top right*) separately in this analysis of memory. Consequently, memory scores were averaged across stimulus sets for all further analyses.