

Supporting Information

Category-specific neural oscillations predict recall organization during memory search

Neal W Morton ^{*}, Michael J. Kahana [†], Emily A. Rosenberg [†], Gordon H. Baltuch [‡], Brian Litt [‡], Ashwini D. Sharan [§], Michael R. Sperling [§], and Sean M. Polyn ^{*}

^{*}Department of Psychology, Vanderbilt University, PMB 407814, 2301 Vanderbilt Place, Nashville, TN 37240, [†]Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104, [‡]Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, and [§]Comprehensive Epilepsy Center, Thomas Jefferson University Hospital, 900 Walnut Street, Philadelphia, PA 19107

Submitted to Proceedings of the National Academy of Sciences of the United States of America

SI Materials and Methods

Scalp Electroencephalography (EEG) Experiment

Participants. Forty-one paid volunteers (15 female, age 18–30 years) were recruited; 3 participants were excluded due to technical problems with the EEG recording apparatus, and 9 participants were excluded due to excessive eye movements, leaving 29 participants presented here. The research protocol was approved by the Institutional Review Board of the University of Pennsylvania.

Experimental paradigm. Stimuli consisted of color and black and white photographs of famous landmarks, celebrity faces, and common objects, with the name of the stimulus presented in text above the picture. There were 256 stimuli for each category. Stimuli were presented using pyEPL [1].

Prior to the first free-recall session, participants rated their familiarity with each stimulus used in the experiment. This was done both to assess participants' pre-experimental familiarity with each stimulus, and to provide participants at least a minimal familiarity with each stimulus. Stimuli were presented pseudorandomly, with the constraints that every group of three contained stimuli from each of the three categories, and that no two adjacent items were of the same category. Each stimulus was presented for 3500 ms, during which participants rated their familiarity with the stimulus' referent on a four-point scale. Each stimulus was followed by a blank interstimulus interval (ISI) of 1000±200 ms. Participants were given a chance to rest after each group of 48 items.

In the subsequent 3 sessions, participants were presented with 48 study-test lists. Each list was composed of 24 stimuli. There were two types of lists: *mixed* lists which contained 8 stimuli from each of the three categories, and *pure* lists which were composed of stimuli all drawn from the same category. In the mixed lists, items were presented in trains of same-category items, with each train containing 2–6 items. The order of category trains was pseudorandom, with the constraints that all categories appeared in each set of 3 trains, and that adjacent trains did not contain the same category. Each session contained 10 mixed lists and 6 pure lists. The pure lists were included to establish a baseline measure of temporal clustering, so this effect could be controlled for when examining category clustering [2]. The order of mixed and pure lists within each session was pseudorandom. Stimuli did not appear more than once within a session, and stimuli were chosen so that items from the same sub-category (e.g. stadiums, presidents) did not appear in the same list.

Each stimulus was presented for 3500 ms, during which participants made a category-specific four-point semantic judgment (celebrities: "How much do you love or hate this person?"; landmarks: "How much would you like to visit this place?"; objects: "How often do you come across this object in your daily life?"). Studied items where the participant didn't respond or responded faster than 300 ms were excluded from all analyses; 0–36 study epochs were excluded for each participant. Each stimulus was followed by a blank ISI of 1000±200 ms.

After presentation of the last stimulus, the screen was blank for 1300±100 ms, followed by presentation of a row of asterisks and a 300 ms tone signaling the start of a 90 s immediate free recall (IFR) period. Participants were instructed to recall items from the list in any order, without regard to stimulus category. Digital recordings of vocal recalls were scored using PyParse [3]. Intrusions of items not in the word pool were scored to determine the category if possible (e.g. "Meryl Streep" was not in the word pool but is clearly a celebrity, while "rock" may have referred to an object or a landmark and therefore had ambiguous category). Intrusions of ambiguous category were excluded from all analyses.

At the end of each session, there was a final free recall (FFR) period where participants were given 360 s to recall names of stimuli from any of the lists presented during the session.

Behavioral analysis. We assessed the degree of category clustering during IFR using the list-based semantic clustering index (LBC_{sem}) [4]. We used a relabeling procedure to establish a baseline level of clustering expected due to the temporal contiguity of same-category items during study [2]. Each pure list was assigned a set of category labels by randomly sampling with replacement from the set of mixed lists for that subject. Mean LBC_{sem} was then calculated for the relabeled pure lists. The random relabeling procedure was repeated 10000 times to establish a null distribution of mean LBC_{sem} expected in the absence of category clustering. Because LBC_{sem} varies with list length, we used a different measure, the adjusted ratio of clustering (ARC) score, to compare category clustering in IFR and FFR [5].

Scalp EEG recordings and data processing. EEG measurements were recorded using 129-channel HydroCel Geodesic Sensor Nets and a Net Amps 200 Amplifier (Electrical Geodesics, Inc.). An analog bandpass filter of 0.5–200 Hz was applied to recorded voltage, which was then digitized at 500 Hz. Recordings were initially referenced to Cz and were later converted to an average reference. To remove the influence of electrodes with poor contact, we first used multiple regression to remove signal related to vertical electrooculogram (VEOG) and horizontal electrooculogram (HEOG) measured using electrode pairs placed near the eyes, then excluded from the average reference electrodes with a mean or standard deviation (based on the signal with eye artifacts regressed out) greater than 4 standard deviations away from the mean across electrodes. Line noise was

Reserved for Publication Footnotes

removed using a Butterworth filter with zero phase distortion at 60 Hz.

We used a modified version of the eye motion correction procedure to remove blinks and eye movements [6]. In order to better discriminate between blinks and eye movements, we identified blinks by applying a threshold to the difference between a fast and slow running average of the VEOG. Before each session of the experiment, participants were instructed to make 10 voluntary blinks and 20 eye movements (5 each of up, down, left, and right saccades) while HEOG and VEOG signals were recorded. The blink detector was applied to each participant's voluntary blinks and eye movements, and the threshold was adjusted to correctly identify at least 80% of the blinks while minimizing the number of eye movements incorrectly identified. The optimized blink detector was then applied to that participant's experimental data to identify time periods containing blinks. A buffer of 150 ms before and 500 ms after was added to each time sample identified as containing a blink to capture slower changes missed by the blink detector. Multiple linear regression was used to predict the signal at each electrode using (1) VEOG not containing blinks, (2) VEOG containing blinks, (3) HEOG not containing blinks, (4) HEOG containing blinks, and an intercept as predictors. The residuals from this regression were then used as corrected EEG. When calculating propagation factors, we did not subtract the average event-related potential (ERP) from each epoch as Gratton and colleagues [6], because we found in an independent data set that correction performance was better when propagation factors were calculated on raw EEG rather than deviation scores (performance improved according to the metrics of variance after correction, and deviation from an estimate of the "true" ERP obtained from averaging events that passed a strict voltage threshold) [6]. The EEG of 9 participants was not well-corrected by this procedure, due to large eye movement artifacts that were difficult to discriminate from blinks; these participants were excluded from the present analyses.

Our focus in the present article is understanding the neural correlates of category clustering; therefore, the EEG analyses reported here examine only the mixed lists. We examined EEG recorded during familiarization, study, immediate free recall, and final free recall. In order to thoroughly characterize neural activity during the free recall periods, we examined both continuous data including entire recall periods and segmented data locked to the onset of vocalized recalls.

Oscillatory analysis. We measured oscillatory power using a Morlet wavelet transform with a wave-number of 6. Oscillatory power was calculated at 34 logarithmically spaced frequencies from 2 to 100 Hz. Power values were then log-transformed and down-sampled to 20 Hz. Power was z -transformed relative to the mean and standard deviation of a baseline period, separately for each frequency, electrode, and session. For study epochs, the baseline period was 500–400 ms before stimulus onset. For recall epochs locked to vocalization onset, quiet times during recall where no vocalizations were being made were used as baseline; for each list, enough 100 ms baseline epochs were randomly chosen from quiet periods to match the number of recall events on that list. For examining entire recall periods, power was normalized relative to all samples in a given recall period.

Multivariate pattern analysis. We used multivariate pattern analysis [7] to decode stimulus category based on patterns of oscillatory power. Classification was carried out using penalized logistic regression (penalty parameter = 10), using L_2 regularization [8]. Classification analyses were carried out using the EEG Analysis Toolbox (available at: <http://code.google.com/p/eeeg-analysis-toolbox>) and the Princeton MVPA Toolbox (available at: <http://www.pni.princeton.edu/mvpa>).

Study analysis. During study, we first analyzed each time (relative to stimulus onset) and frequency separately. Classifier performance was assessed using a cross-validation procedure, where the classifier was trained on all lists except one, then tested on the remaining list. Classi-

fier performance was measured as the fraction of unseen items whose category was correctly classified. This procedure was repeated with a different list left out on each iteration, and classifier performance was averaged over iterations.

We examined changes in classifier performance over time by analyzing average power in 500 ms bins swept over the stimulus presentation period, including all electrodes and frequencies in the classified patterns. Finally, to obtain a measure of overall classifier performance for a given item presentation, average oscillatory power was calculated for 0–0.5 s post-stimulus onset (*early* time bin) and 0.5–3.5 s post-stimulus onset (*late* time bin), and the classifier was given patterns including both time bins, all frequencies, and all electrodes.

We also examined the performance of a classifier trained on the familiarization period, then tested on the study period. We trained the classifier on power during presentation of all items during the familiarization period, including both the early and late time bins, and all electrodes and frequencies in each pattern; we then applied the classifier to all study patterns, and assessed the classifier's performance by measuring the fraction of study events that were correctly classified.

Study epochs were divided based on the later recall performance of the studied items. Items were first divided into *recalled* (recalled during IFR) and *forgotten* (not recalled during IFR) groups, and we examined whether classifier performance differed between these groups. Recalled items were further divided based on whether they were *subsequently clustered* (recalled as part of a sequence of 2 or more items of the same category) or *subsequently isolated* (not recalled as part of a category cluster), to test the hypothesis that subsequently clustered items would be associated with greater classifier accuracy.

Recall analysis. To examine whether patterns of oscillatory power observed during study were reactivated during recall, we trained the classifier on average power observed during the late time bin of all of the patterns observed during item presentation, then applied the classifier to oscillatory power recorded during recall. We assessed the degree of reactivation of category-specific oscillatory patterns during the recall period using a correlation-based reactivation metric [9].

The classifier provides an estimate of the strength S_t^i of each category i at each time bin t . The record of recalls during each free recall period was sampled at 20 Hz to match the sampling rate of the oscillatory power. Each time bin was either assigned no category (if no recalls were currently being made), or to exactly one category. The 1 s preceding onset of each vocalized recall was labeled with the category of the recalled item. When there was overlap between recalls, the earlier item took precedence. This resulted in a set of 3 vectors \mathbf{R}^j , where each element R_t^j is 1 for times t when category j is active, and 0 when category j is not active. These vectors represent the *recall record* of each recall period.

We calculated a correlation-based reactivation metric to measure reactivation of category patterns during recall. We treated all recall periods as part of one record by concatenating the recall periods together. We calculated Pearson's linear correlation between \mathbf{S}^i and \mathbf{R}^j for $i = 1, 2, 3$ and $j = 1, 2, 3$ to create a cross-correlation matrix. The diagonal of the cross-correlation matrix corresponds to correlations between classifier estimates and the correct recall records, while the off-diagonal entries correspond to correlation with the incorrect categories. We calculated the mean correlation in the diagonal entries and subtracted the mean correlation in the off-diagonal entries to obtain a summary index of the classifier's ability to track each subject's recall behavior, which we refer to as the reactivation metric [9].

We used a permutation test to determine whether reactivation was significant across subjects. For each subject, the columns of the cross-correlation matrix were scrambled, and the mean reactivation metric was calculated. This process was repeated 5000 times to establish a null distribution of reactivation metrics, and reactivation was considered significant if it was greater than 95% of the null distribution. We also examined reactivation at different frequencies by training and testing the classifier at each frequency individually. In order to con-

Table S1. Patient information

ID	HOSP	AGE	SEX	HAND	ELC	SES
1	UP	18	M	A	100	6
2	UP	39	M	L	77	1
3	UP	40	M	R	38	2
4	TJ	25	M	R	35	3
5	TJ	40	F	R	82	10
6	TJ	39	M	L	52	4
7	TJ	34	F	R	92	2
7	TJ	34	F	R	86	8
8	TJ	39	F	R	85	1
9	TJ	44	M	R	124	4
10	TJ	29	M	R	36	1
11	TJ	43	M	R	57	5

This table provides the hospital (HOSP) at which each patient's data were collected, as well as each patient's age in years (AGE), sex (SEX), handedness (HAND), number of implanted electrodes (ELC), and number of testing sessions (SES). Patient 7 underwent invasive monitoring with 2 partially overlapping sets of electrodes (see text for details). A, ambidextrous; F, female; L, left; M, male; R, right; TJ, Thomas Jefferson Hospital (Philadelphia, PA); UP, Hospital of the University of Pennsylvania (Philadelphia, PA).

control Type I error rate while accounting for the correlation structure of the data, we scrambled the columns of the cross-correlation matrix in the same way for each frequency, then pooled the null distributions of each frequency together to make one null distribution accounting for familywise error. This familywise null distribution was then used to set the significance threshold for all frequencies [10].

We also examined recall-specific category-related activity by training the classifier on recall epochs, to identify the category of upcoming recalls. The classifier used oscillatory power at all frequencies, averaged from 3 to 0.5 s before onset of vocalization (the 0.5 s immediately before each vocalization was excluded in order to limit the influence of vocal response preparation artifacts). Recall epochs were excluded if they overlapped with vocalizations of previous recalls (immediate free recall: 68.5% (SEM 1.3%) of epochs were excluded, leaving 44–146 epochs for each subject; final free recall: 78.0% (SEM 1.3%) of epochs were excluded, leaving 31–93 epochs for each subject). Performance was assessed using cross-validation, with one list left out on each iteration, and performance was averaged over iterations.

Electrocorticography (ECoG) Experiment

Participants. We tested 11 patients (3 female; age 18–44, mean 35.5, SD 8.2) with medication-resistant epilepsy who were undergoing invasive EEG monitoring to determine the location of epileptogenic foci for subsequent resection. See Table S1 for detailed patient information. The patients had a total of 864 surface and depth electrodes (Fig. 1a); electrode placement was determined by the clinical team.

The research protocol was approved by the relevant institutional review boards, and informed consent was obtained from all participants. To limit the effects of seizures and medication on task performance and brain activity, we refrained from testing when patients were on high doses of pain medications or anti-epileptic drugs, and during the 6 hour period following any clinically significant seizure. For 2 sessions, patient 7 was implanted with one set of electrodes; before the remaining 8 sessions, some electrodes were added, and some were removed to create a second set of electrodes. In the reported classification analyses, we treated these 2 sets of sessions as coming from distinct participants. The number of electrodes that overlapped between the 2 sets of sessions was relatively small (24 electrodes; 26.1% of the first set of electrodes, and 27.9% of the second set), precluding our ability to combine these 2 sets of data.

Materials. The word pool consisted of the 216 items from the scalp EEG experiment that were the most recognizable (as judged by the experimenters). In addition to the original picture used in the scalp EEG experiment, 4 additional pictures were found for each item.

Procedure. Participants were presented with lists of 9 items, with 3 items from each category. Category order was pseudorandom within each list. Before each item, a text cue indicating the category of the upcoming item was shown for 1000 ms. There was a 200–500 ms ISI before presentation of the item, which lasted for 3500 ms. While the item was on the screen, participants made a category-specific judgment, as in the scalp EEG experiment. The ISI between each item and the next category cue was 800–1200 ms. After presentation of the last stimulus, the screen was blank for 1200–1400 ms, followed by presentation of a row of asterisks and a 300 ms tone signaling the start of a 60 s IFR period. If 60 s had not passed yet, but the participant indicated that he or she had finished recall, the experimenter pressed a button to end the recall period.

Each item had 5 distinct pictures, which all appeared during the session (but never in the same list). Participants were told that the same item might appear multiple times, but to simply focus on remembering the items of the current list.

Participants were presented with 20 lists in each session. There was a 240 s final free recall test at the end of each session. Each participant completed 1–10 sessions (see Table S1 for the number of sessions completed by each participant).

ECoG recordings and data processing. ECoG was recorded using a Grass Telefactor or Nicolet digital video-EEG system. ECoG was sampled at 400 or 512 Hz. A digital Butterworth notch filter with zero phase distortion at 60 Hz was used to remove electrical noise. Synchronization pulses controlled by the computer presenting the stimuli were sent the EEG monitoring system, and later used to align electrophysiological data to events in the experiment (precision < 4 ms).

Oscillatory power was measured at 37 logarithmically spaced frequencies from 2 to 128 Hz. Power was log-transformed and down-sampled to 16 Hz. Power was normalized as in the scalp EEG experiment, except power measured during study epochs was normalized relative to 500–400 ms before onset of the category cue. Epochs examined during the recall period consisted of data from 2000 ms before to 1000 ms after vocalization onset. Epochs were only included if they did not contain previous vocalizations (immediate free recall: 70.8% (SEM 4.5%) of epochs were excluded, leaving 5–744 epochs for each subject; final free recall: 67.6% (SEM 3.9%) of epochs were excluded, leaving 3–90 epochs for each subject). Power was normalized relative to periods with no vocalizations; for each recall period, enough 100 ms baseline epochs were randomly chosen from quiet periods to match the number of recalls during that recall period. We also examined continuous data including entire recall periods; power was z -transformed based on the mean and standard deviation of power over each recall period, separately for each electrode and frequency.

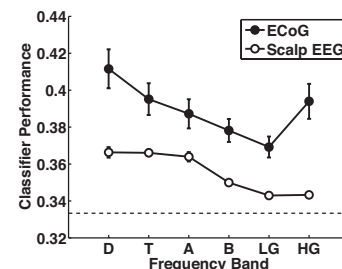


Fig. S1. Classifier performance for study period cross-validation, for the ECoG and scalp EEG experiments, as a function of frequency band. D: delta, 2–4 Hz, T: theta, 4–8 Hz, A: alpha, 10–14 Hz, B: beta, 16–25 Hz, LG: low gamma, 25–55 Hz, HG: high gamma, 65–100 Hz. The dotted line indicates chance performance (1/3). Error bars represent standard error of the mean.

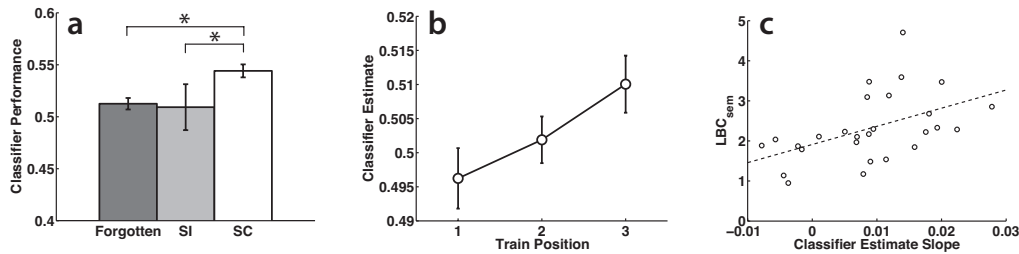


Fig. S2. Scalp EEG study period analyses with oscillatory power above 30 *Hz* excluded. **(a)** When a classifier was trained on item presentations during a familiarity judgment task and applied to the study period of free recall lists, subsequently clustered (SC) items were classified more accurately than both subsequently isolated (SI) items and forgotten items. Error bars represent 95% confidence intervals based on within-subject error [16]. **(b)** The classifier's estimate of the strength of the current category is plotted as a function of position within a train of same-category item presentations. On average, classifier estimates rose with successive same-category stimuli. Error bars represent 95% confidence intervals based on within-subject error [16]. **(c)** The slope of the regression of classifier estimate on train position was correlated with individual differences in category clustering as measured by LBC_{sem} ($r = 0.490$, $p < 0.01$). Two outliers have been removed from the plot; with them included, the correlation is still significant ($r = 0.419$, $p < 0.05$).

The locations of the intracranial electrodes were determined using an indirect stereotactic technique based on co-registered post-operative computed tomography and pre- or post-operative magnetic resonance imaging, and converted into Montreal Neurological Institute coordinates. The Talairach Atlas was used to determine the anatomical location of each electrode [11, 12]. Electrodes were divided into 7 regions of interest (ROIs; see Fig. 1a): frontal (220 electrodes), prefrontal (188), temporal (532), medial temporal (76), hippocampus (22), occipital (56), and parietal (57). The prefrontal ROI is a subset of electrodes in the frontal ROI; similarly, the hippocampal ROI is a subset of the medial temporal ROI, which is a subset of the temporal ROI. We used brain images from the WFU Pick-Atlas for data visualization [13].

Multivariate pattern analysis. Pattern analysis methods were the same as in the scalp EEG experiment, except that classification analyses were carried out separately for each ROI. Significance of classifier performance during study was assessed using a permutation test. The labels corresponding to each category were permuted 5000 times, and the mean fraction correct over participants was calculated for each permutation. The same permutations were used across all ROIs. The permuted distribution of fraction correct scores was pooled over all ROIs to create one null distribution, which was used to establish a significance threshold that controls familywise Type I error at $\alpha < 0.05$ [10]. A similar method was used to assess significance of reactivation during recall; in this case, columns of the cross-correlation matrix for each participant were permuted to calculate a null distribution of reactivation metrics, which was pooled over all ROIs to set familywise Type I error at $\alpha < 0.05$.

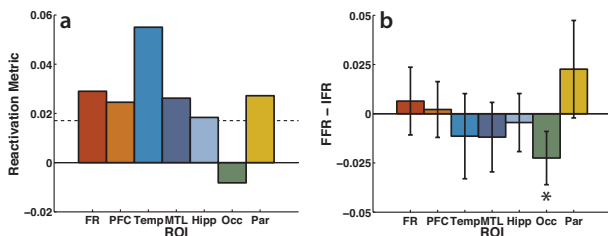


Fig. S3. **(a)** Reactivation of category-specific information during final free recall is observed in frontal, prefrontal, temporal, medial temporal, hippocampal, and parietal electrodes. The dotted line indicates the significance threshold for a permutation test comparing performance to chance (familywise Type I error rate < 0.05). **(b)** There is significantly less reactivation in occipital electrodes in FFR compared to IFR. Difference in reactivation metric between FFR and IFR is shown for each region of interest (ROI). Error bars indicate 95% confidence corresponding to a two-tailed *t*-test. * indicates $p < 0.05$, Bonferroni corrected. FR: frontal lobe, PFC: prefrontal cortex, Temp: temporal lobe, MTL: medial temporal lobe, Hipp: hippocampus, Occ: occipital lobe, Par: parietal lobe.

SI Results

Category clustering. Participants exhibited reliable category clustering, as measured by the list-based semantic clustering index (LBC_{sem}) [4]. In the scalp EEG experiment, LBC_{sem} in immediate free recall (IFR) was 3.66 (SEM 0.25); this exceeded the amount of category clustering expected given temporal influences on recall, calculated using a relabeling procedure (described in SI Materials and Methods; mean 0.808, SD 0.061; $p < 0.0002$) [2]. In the ECoG experiment, because items from each category were randomly placed in the list, LBC_{sem} expected by chance is 0. LBC_{sem} in IFR was 1.19 (SEM 0.16), which was significantly greater than chance ($t(11) = 7.39$, $p < 0.0001$).

Because LBC_{sem} varies with list length, we used a different measure of semantic clustering, the adjusted ratio of clustering (ARC) score [5], to compare category clustering in IFR and final free recall (FFR). In the scalp EEG experiment, ARC score for immediate recall was 0.60 (SEM 0.02); ARC score for final recall was 0.88 (SEM 0.02), and was significantly greater than immediate free recall ($t(28) = 14.93$; $p < 0.0001$). A similar difference was observed in the ECoG experiment (IFR: mean 0.62, SEM 0.25; FFR: mean 0.85, SEM 0.03; $t(11) = 3.65$, $p < 0.002$).

Frequency-specificity of category-specific activity during study.

To examine the category-specificity of oscillations at different frequency bands during the study period, we averaged classifier cross-validation performance over the entire stimulus presentation period (0–3500 ms post-stimulus onset) and 6 frequency bands: delta (2–4 *Hz*), theta (4–8 *Hz*), alpha (10–14 *Hz*), beta (16–25 *Hz*), low gamma (25–55 *Hz*), and high gamma (65–100 *Hz*). Here, we used 100 *Hz* as the upper bound of high gamma to allow comparison of the ECoG and scalp EEG signals. Classifier performance for the ECoG experiment was greater than scalp EEG ($F(1, 234) = 188.66$, $p < 0.0001$). There was also a main effect of frequency ($F(5, 234) = 15.45$, $p < 0.0001$), and a significant interaction ($F(5, 234) = 3.51$, $p < 0.005$). There was an interaction between ECoG and scalp EEG in the low and high gamma bands ($F(1, 78) = 10.88$, $p < 0.002$), with the ECoG data showing a greater advantage for high gamma over low gamma (Fig. S1). This may reflect attenuation of high-frequency oscillations by the skull [14].

Although gamma-band oscillations were attenuated at the scalp electrodes, classifier performance for frequencies in the gamma band was still reliably above chance during the study period. Recent findings suggest that some high-frequency EEG activity measured at scalp electrodes is related to miniature saccades, rather than brain activity [15]. Furthermore, voltage potentials related to miniature saccades do not differ in polarity on different sides of the eye, so our regression procedure for subtracting the influence of eye movements (which relied on difference potentials to measure eye movements) would not

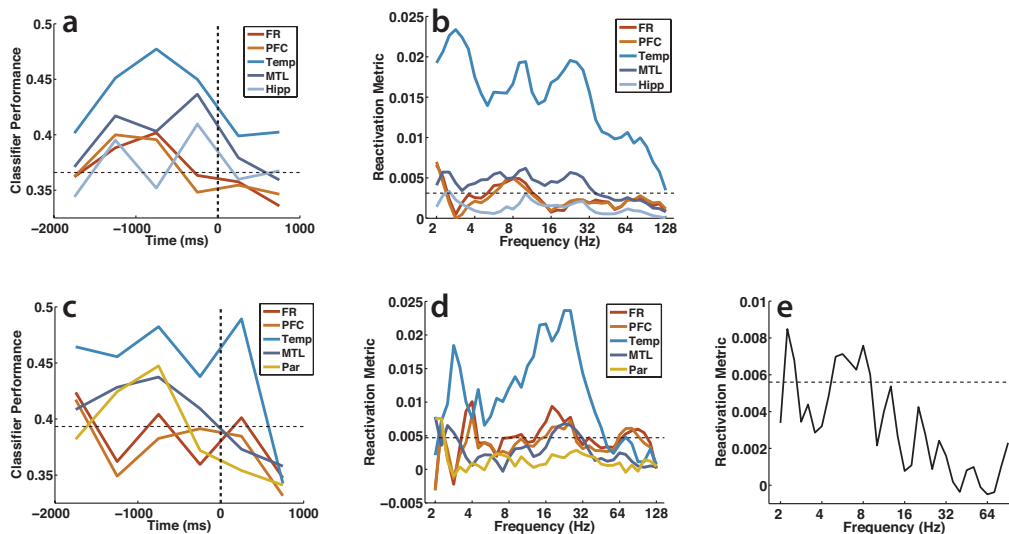


Fig. S4. Time- and frequency-specificity of reactivation during free recall. (a) In the intracranially monitored participants during immediate free recall (IFR), reactivation peaks during the 1 s before onset of vocalization, then drops during vocalization of the recalled item. Onset of vocalization is at time = 0. (b) Reactivation during IFR as a function of region of interest (ROI) and frequency. (c) Performance of a classifier trained on the study period and applied to final free recall (FFR), in 500 ms bins. (d) Reactivation during FFR as a function of ROI and frequency. (e) In the scalp-monitored participants, no individual frequencies demonstrated significant reactivation during IFR. However, during FFR, there was significant reactivation in the delta and theta frequency bands. The horizontal dotted lines indicate significance thresholds for permutation tests comparing performance to chance (familywise Type I error rate < 0.05).

be effective at removing the influence of miniature saccades on our measures of oscillatory power [15].

To ensure that our conclusions were not influenced by miniature saccades, we replicated our analyses of scalp EEG during the study period, this time excluding all frequencies greater than 30 Hz. We found similar results in all analyses. As shown in Figure S2a, subsequently clustered items were classified more accurately than both subsequently isolated ($t(28) = 2.17, p < 0.05$) and forgotten items ($t(28) = 5.00, p < 0.0001$). As shown in Figure S2b-c, classifier estimates increased significantly with multiple presentations of items of the same category (slope: mean 0.0068, SEM 0.0022, $t(28) = 3.13, p < 0.005$), and this increase was correlated with individual differences in category clustering as measured by LBC_{sem} ($r = 0.419, p < 0.05$; with 2 outliers excluded, $r = 0.490, p < 0.01$). The slope of classifier estimates also correlated with LBC_{sem} on a trial-by-trial basis (see main text for details; t for the slope of regression of LBC_{sem} on slope of classifier estimates: mean 0.442, SEM 0.160, $t(28) = 2.77, p < 0.005$).

Category-specific activity preceding stimulus onset. In the scalp EEG experiment, when all study items are included, classifier performance is significantly above chance during the 500 ms before stimulus onset (mean 35.81%, SEM 0.45%, $t(28) = 5.53, p < 0.0001$). This is likely due to persistence of category-specific activity related to the previous item (69.1% of items are preceded by an item of the same category; also see Fig. 4a for evidence of sustained category-specific activity during the study period).

In the ECoG experiment, each item was preceded by a cue indicating the category of the upcoming item. Classifier performance in the 500 ms before stimulus onset was significantly above chance (mean 37.81%, SEM 1.53%, $t(11) = 24.7, p < 0.0001$; see Fig. 1d); this may be due to activity related to anticipation of the stimulus or preparation for the category-specific judgment to be made about the item.

Time-course and frequency specificity of reactivation. We demonstrated that category-specific oscillations were reactivated during recall using a correlation-based analysis (Fig. 2b). In order to determine the time-course of category reactivation in the ECoG experiment dur-

ing IFR, we applied the classifier to recall epochs locked to the onset of vocalization (see SI Materials and Methods for details).

We divided recall epochs into 500 ms time bins, and ran a separate classification analysis for each bin. Significance of classifier performance was assessed using a permutation test. The labels corresponding to each category were permuted 5000 times, and the mean fraction correct over participants was calculated for each permutation. The same permutations were used across all time bins and ROIs. The permuted distribution of fraction correct scores was pooled over all time bins and ROIs to create one null distribution, which was used to establish a significance threshold that controls familywise Type I error at $\alpha < 0.05$ [10]. As shown in Figure S4a, classifier accuracy peaks during the 1 s before onset of vocalization, then drops during vocalization of the recalled item.

Next we examined whether reactivation of category-specific oscillations during IFR was restricted to particular frequency bands. We ran a separate classification analysis for each frequency to determine the degree of reactivation of each frequency of oscillation. Reactivation was measured using a correlation-based reactivation metric (see SI Methods and Materials for details). Significance was assessed using a permutation test where the columns of each participant's cross-correlation matrix were permuted 5000 times; the significance threshold was determined by pooling all frequencies and ROIs into one null distribution (familywise Type I error < 0.05) [10]. As shown in Figure S4b, in temporal electrodes, oscillatory power at all frequencies was reactivated; in medial temporal electrodes, delta, theta, alpha, and beta power was reactivated; and in frontal electrodes, delta, theta, and alpha power was reactivated.

We also examined the time-course and frequency specificity of reactivation in final free recall. There is a similar time-course as in IFR, with classifier performance peaking around 1 s before onset of vocalization, then decreasing after vocalization (Fig. S4c). In the scalp electrodes, classifier performance was below chance at all time bins. Temporal electrodes demonstrated reactivation in all frequency bands except high gamma; in frontal electrodes, theta, beta, and high gamma power was reactivated; and in medial temporal electrodes, beta power was reactivated (Fig. S4d). In scalp electrodes, there was significant reactivation in delta and theta power (Fig. S4e).

Classifier performance at temporal electrodes and subsequent memory. We found that, at temporal electrodes, there was no difference in classifier performance between recalled and forgotten items ($t(11) = 1.27, p = 0.23$, Bonferroni corrected). Since performance at temporal electrodes is near ceiling (mean 79.22%, SEM 2.96%), we also examined whether the category probability estimates produced by the classifier predicted subsequent memory status. Classifier estimates are free to vary continuously, so they may be more sensitive in some cases than fraction correct, which is binary for each classified item [17]. We found that classifier estimates at temporal electrodes were significantly greater for subsequently recalled than forgotten items ($t(11) = 8.50, p < 0.0005$, Bonferroni corrected).

Sensitivity of the classifier to subsequent recall performance. In the scalp EEG experiment, the familiarization period and the study period are associated with distinct cognitive processes. The familiarization period involves making familiarity judgements about presented stimuli, which do not need to be remembered; in contrast, the study period involves distinct encoding tasks for each category, as well as cognitive processes involved in explicitly attempting to remember the items for the free recall test. Therefore, a classifier may be sensitive to different category-specific neural representations depending on whether it was trained on the familiarization period or the study period.

We examined whether item-level fluctuations in classifier performance predicted subsequent clustering by category. When the classifier was trained on the familiarization period, then applied to study, there was a significant difference between subsequently clustered and isolated items. However, when the classifier was trained on the study period (using a cross-validation procedure), the difference in classifier performance between subsequently clustered (mean performance 60.39%, SEM 1.23%) and subsequently isolated items (mean performance 59.55%, SEM 1.99%) was not significant ($t(28) = 1.17, p = 0.25$). This suggests that a classifier trained on the familiarization period is more sensitive to changes in category-specific activity that predict subsequent recall organization. To test this hypothesis, we examined whether there was an interaction between training period (familiarization or study) and subsequent organization (isolated or clustered). There was a significant main effect of training period ($F(1, 28) = 32.42, p < 0.0001$; accuracy was better when the classifier was trained on the study period), no main effect of subsequent organization ($F(1, 28) = 2.53, p = 0.13$), and no interaction ($F(1, 28) = 1.75, p = 0.20$). Therefore, although there is a significant difference in classifier performance between clustered and isolated items when the classifier is trained on the familiarization period, and no difference when the classifier is trained on the study period, the magnitude of the difference in classifier performance between subsequently clustered and subsequently isolated items does not significantly interact with training period. Similarly, the difference in classifier performance between recalled and forgotten items does not depend on the training period. There was a significant main effect of training period ($F(1, 28) = 53.78, p < 0.0001$), a significant main effect of subsequent memory ($F(1, 28) = 16.87, p < 0.0005$), and no interaction ($F(1, 28) < 1$).

However, the training period is important for determining whether the classifier is sensitive to effects of integration of category representations over multiple item presentations. As shown in Figure S5a, when the classifier was trained on the familiarization period, there was no increase in classifier estimate with train position (slope over train positions 1–3, based on weighted least-squares regression: mean 0.0010, SEM 0.0021, $t(28) = 0.51, p = 0.31$, one-sided test compared to 0). Slope was significantly greater when the classifier was trained on the study period ($t(28) = 3.12, p < 0.005$). Furthermore, when the classifier was trained on the familiarization period, the slope of classifier estimates did not correlate with LBC_{sem} (Fig. S5b; $r = 0.00062, p = 0.997$). A dependent correlations

test showed that LBC correlates significantly better with slope for study cross-validation than for familiarization-to-study classification ($t(26) = 2.30, p < 0.03$). Similarly, the slope of classifier estimates is more sensitive to differences in the amount of clustering on individual lists when the classifier is trained on the study period, compared to when it is trained on the familiarization period (t for the slope of the regression of LBC_{sem} on slope of classifier estimates: mean 0.190, SEM 0.159). This difference is marginally significant ($t(28) = 1.96, p = 0.061$).

Therefore, when the classifier is trained on the study period, its estimates are sensitive to integration of category-specific patterns over time, and also somewhat sensitive to item-level fluctuations that predict subsequent organization. In contrast, a classifier trained on the familiarization period is sensitive only to item-level fluctuations that predict subsequent organization, but not integration of category-specific activity over time. This may be due to differences in the cognitive processing engaged during familiarization and study; during familiarization, participants presumably do not construct retrieval cues, since they are not required to memorize the items. Therefore, a classifier trained on the familiarization period may be insensitive to retrieval-cue-related signal when it is applied to the study period.

Comparing classification of immediate and final recall. As shown in Fig. 5a, classifier cross-validation performance was significantly

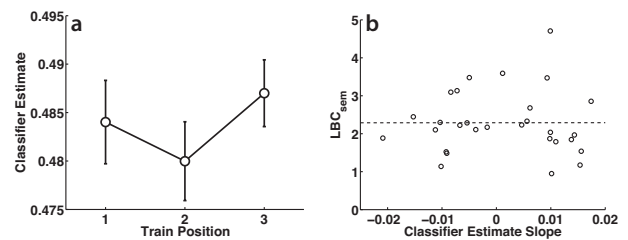


Fig. S5. (a) When a classifier is trained on the familiarization session, then applied to the study period during the free recall task, classifier estimate does not increase with train position. Across participants, the slope of the regression of classifier estimate on train position is not significantly positive. Error bars indicate 95% confidence based on within-subject error [16]. (b) Furthermore, classifier estimate slope does not correlate with individual differences in category clustering, as measured by LBC_{sem} ($r = 0.00062, p = 0.997$).

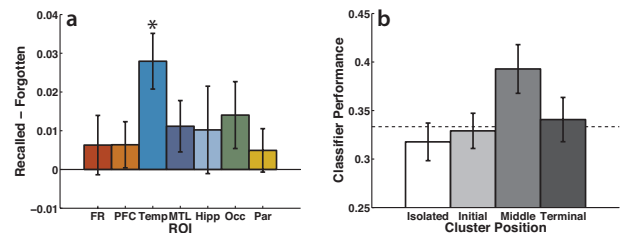


Fig. S6. (a) The fidelity of category-specific patterns in ECoG recorded from temporal electrodes still predicts subsequent memory when training set imbalances are removed: Classifier estimates are greater for subsequently recalled items than subsequently forgotten items. ROI: region of interest, FR: frontal lobe, PFC: prefrontal cortex, Temp: temporal lobe, MTL: medial temporal lobe, Hipp: hippocampus, Occ: occipital lobe, Par: parietal lobe. Error bars indicate 95% confidence intervals corresponding to a one-tailed t -test; * indicates $p < 0.05$, Bonferroni corrected. (b) In the scalp EEG-monitored participants, classifier performance during immediate free recall was related to category clustering even when the training set was balanced using random undersampling. Classifier performance was significantly higher for cluster positions where the previous item was from the same category as the current item (middle, terminal), as compared to cluster positions where the previous item was from a different category (isolated, initial). Classifier performance was also greater when the next item was the same category (initial, middle), compared to when the next item was a different category (isolated, terminal). Error bars indicate 95% confidence intervals corresponding to a one-tailed t -test vs. chance (1/3; indicated by the dotted line).

greater in FFR, compared to IFR. A possible confound is the number of epochs being used to train the classifier. However, the number of epochs was greater for IFR (mean 110.45, SEM 4.51) than for FFR (mean 62.21, SEM 2.61). This difference was significant ($t(28) = 12.11, p < 0.0001$). Because the classifier was, on average, trained on fewer epochs for the FFR classification, this difference cannot account for the increased classifier performance in FFR compared to IFR. Furthermore, a similar difference in performance is observed when the training set is held constant (the classifier is trained on the study period, and applied to IFR and FFR recall periods). Classifier performance is greater for FFR (reactivation metric: mean 0.0130, SEM 0.0041) than IFR (mean 0.0029, SEM 0.0012). This difference is significant ($t(28) = 2.45, p < 0.05$).

Training set imbalances. In the main text, we refer to increased classifier estimates and classifier performance as reflecting stronger (or higher fidelity) category patterns. However, if there is an interaction between an effect of interest (for example, a difference between subsequently forgotten and recalled words) and category, this will complicate interpretation of classifier performance differences. For example, if there are distinct category patterns for forgotten and recalled words (this might occur if there is an interaction between category identity and subsequent memory), then the difference in classifier accuracy between forgotten and recalled items will depend on the number of each type of item in the training set. If the training set is composed mainly of recalled items, then recalled items will be classified more accurately than forgotten items (e.g. the classifier will have learned the “to-be-recalled celebrity” pattern better than the “to-be-forgotten celebrity” pattern); if there are a majority of forgotten items, then forgotten items will be classified more accurately. The observation of a difference in accuracy indicates that the fidelity of category patterns is related to subsequent memory; however, if the training set contains differing numbers of recalled and forgotten items, the direction of this relation is ambiguous.

Training set imbalances are not an issue for analyses where the classifier was trained on the familiarization period, then applied to the study period. Because the familiarization period did not include an episodic memory task, the classifier cannot develop any bias to classify either recalled or forgotten items better than the other.

In order to control for training set imbalances in other analyses, we ran a second set of classification analyses using random undersampling. When creating each training set for the classifier, we ensured that each combination of category and the conditions of interest was equally represented, by sampling randomly without replacement from the set of training patterns corresponding to each category and condition. We repeated each classification analysis 10 times to obtain a stable estimate of performance for each classified item. Classification performance was then calculated for each condition of interest.

When cross-validation was carried out for each ROI during study using random undersampling without replacement, we still observed significantly greater classifier estimates for subsequently recalled items compared to forgotten items in temporal electrodes (Fig. S6a; $t(11) = 6.98, p < 0.0005$, one-sided test).

As shown in Figure S6b, consistent with the original analysis, cross-validation classification of scalp EEG oscillations during IFR demonstrated significantly greater classifier performance for recalls preceded by an item of the same category (i.e. middle and terminal items), compared to items preceded by a recall of a different category (i.e. isolated and initial items; $F(1, 28) = 12.54, p < 0.005$). Classifier performance was also significantly greater for recalls that were followed by an item of the same category (i.e. initial and middle items), compared to items that were followed by an item of a different category (i.e. isolated and terminal items; $F(1, 28) = 8.86, p < 0.01$). There was no interaction between previous category and next category ($F(1, 28) = 2.87, p = 0.1$).

Temporal precision of power estimates. We measured oscillatory power using Morlet wavelets, which were convolved with the EEG to obtain instantaneous estimates of power. Although oscillatory power measured using wavelets is most strongly affected by oscillations at the measured time t , it will also be influenced by surrounding time points in the interval $[t - x, t + x]$, where x depends on the frequency and wavenumber of the wavelet [18]. For all frequencies, we used a wavenumber of 6, so the measured interval varied with frequency; since the lowest measured frequency was 2 Hz, the largest window for which power was affected was $t \pm 1500$ ms. Because some of the reported analyses involved contrasting classification performance for items presented or recalled near in time to items of the same or a different category, the window of influence on power calculations represents a potential confound. Below we discuss how we controlled for issues introduced by the temporal imprecision of wavelet measurements.

Interaction with list construction. The influence from surrounding items on classifier performance during study of an item depends on whether the previous and next items are the same or different categories. If 1 or more adjacent items are the same category as the current item, classifier performance may be increased, while adjacent items of different categories may decrease classifier performance. To control for this potential influence, we divided studied items into train position bins, based on whether they were presented adjacent to items of different categories (isolated), at the beginning of a train of at least two same-category items (initial), in the middle of a train (middle), or at the end of a train (terminal).

First, we examined the effect of subsequent memory status and train position bin on classifier estimates at temporal electrodes, using a two-way within-subjects ANOVA. There was a significant main effect of subsequent memory ($F(1, 11) = 12.55, p < 0.005$), no main effect of train position bin ($F(3, 33) < 1$), and no interaction ($F(3, 33) < 1$). We also tested whether individual train positions showed an effect of subsequent memory using t -tests. There was a significant effect of subsequent memory for isolated items ($t(11) = 5.24, p < 0.005$, Bonferroni corrected), a significant effect for terminal items ($t(11) = 2.98, p < 0.05$, Bonferroni corrected), and no effect for initial or middle items ($p > 0.4$). These results demonstrate that classifier estimates are greater for recalled items compared to forgotten items, even when controlling for the potential influence of adjacent items (Fig. S7a).

We next examined the effect of subsequent memory status and train position bin on classifier performance (for a classifier trained on the familiarization period and applied to the study period) at scalp electrodes (Fig. S7b). There was a main effect of subsequent memory ($F(1, 28) = 9.98, p < 0.005$), no main effect of train position bin ($F(2, 56) = 1.56, p = 0.22$), and no interaction ($F(2, 56) < 1$), indicating that the effect of subsequent memory status on classifier performance is not related to the temporal imprecision of wavelet-based power estimates. One-sided Bonferroni-corrected t -tests at individual train position bins revealed a significant effect of subsequent memory for terminal items ($t(28) = 2.20, p < 0.05$), a marginal effect for middle items ($t(28) = 2.07, p < 0.1$), and no significant difference for initial items ($t(28) = 1.82, p > 0.1$).

We also examined whether the effect of subsequent clustering (comparing classifier performance for subsequently isolated and subsequently clustered items) varied with train position bin (Fig. S7c). There was a main effect of subsequent clustering ($F(1, 28) = 5.20, p < 0.05$), no effect of train position bin ($F(2, 56) < 1$), and no interaction ($F(2, 56) = 1.42, p = 0.25$). There were no significant effects of subsequent clustering at any of the individual train position bins (all $p > 0.05$). The presence of a main effect of subsequent clustering indicates that the difference in classifier performance for clustered and isolated items is not related to influence from adjacent items.

Influence of nearby recalls. During recall, we found that items recalled during periods of category clustering were classified more

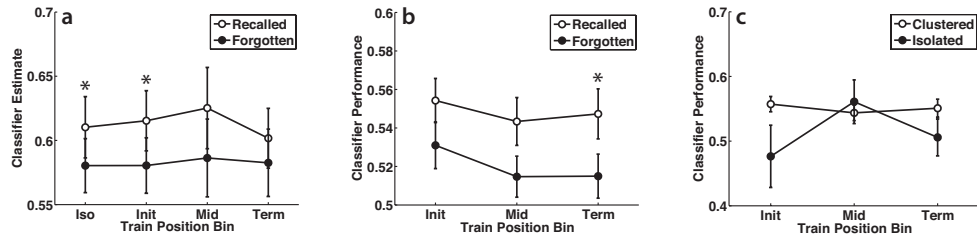


Fig. S7. Classifier performance predicts subsequent memory performance and organization, even when controlling for the category of adjacent items. (a) In temporal electrodes, there is a significant main effect of subsequent memory on classifier estimates, when train position bin is controlled. Classifier performance is plotted for recalled and forgotten items, separately for each train position bin (Iso: isolated, both the preceding and following items are of a different category; Init: initial, the preceding item is a different category, and the next item is the same category; Mid: middle, both adjacent items are of the same category; Term: terminal, the preceding item is the same category, and the next item is a different category). There is a significant effect for items in isolated or initial train position bins. (b) In the scalp EEG-monitored participants, there is a significant main effect of subsequent memory on classifier performance, when train position bin is controlled. There is a significant effect for items in the terminal train position bin. Note that, since all trains were of length 2–6, there were no isolated study items in the scalp EEG experiment. (c) In the scalp EEG-monitored participants, classifier performance is significantly greater for subsequently clustered items, compared to subsequently isolated items, even when train position bin is controlled for. The interaction between subsequent clustering and train position bin is not significant. * indicates $p < 0.05$, Bonferroni corrected. Error bars represent standard error of the mean.

accurately. Since our wavelet estimates of instantaneous oscillatory power are influenced by oscillations within an extended interval, classification of items recalled as part of a cluster may be improved by the influence of oscillatory power related to adjacent recalls of same-category items. To investigate this possibility, we examined classifier performance at 500 ms time bins relative to vocalization onset (Fig. S8a). If clustered items are better classified due to influence of nearby recalls on power estimates, this difference should only appear for time bins that are less than 1500 ms from the closest recall event. Therefore, we focused on the period from 1500 to 500 ms before vocalization onset, which cannot be influenced by adjacent recalls. We averaged classifier performance over this interval. Final free recall (FFR) performance was significantly greater than immediate free recall (IFR) performance during this critical interval ($t(28) = 1.98, p < 0.03$, one-sided test).

We next focused on IFR, and examined the difference between recalls following recall of a same-category item, and recalls not following a same-category item, as a function of time (Fig. S8b). Again, we examined the critical time interval of 1500 to 500 ms before vocalization onset, and found a significant difference between middle/terminal

items and isolated/initial items ($t(28) = 2.82, p < 0.005$, one-sided test).

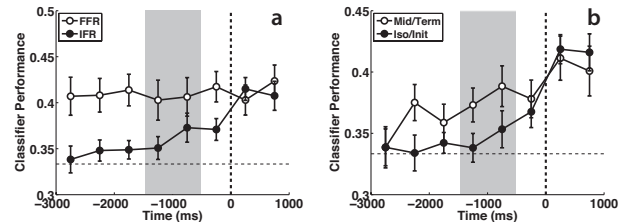


Fig. S8. (a) In the scalp-monitored participants, classifier cross-validation performance was greater in final free recall (FFR) than in immediate free recall (IFR). Classifier performance is shown as a function of recall period and time bin, relative to vocalization onset at time = 0. In the time of interest, -1500 to -500 ms (indicated by the shaded area), classifier performance was significantly greater in FFR compared to IFR. (b) In the scalp-monitored participants, recalls during IFR following a recalled item of the same category (i.e. middle and terminal items) are classified significantly more accurately than recalls following an item of a different category (i.e. isolated and initial items) during the -1500 to -500 ms time bin, indicated by the shaded region. The dotted horizontal lines indicate chance performance. Error bars represent standard error of the mean.

1. Geller AS, Schleifer IK, Sederberg PB, Jacobs J, Kahana MJ (2007) PyEPL: A cross-platform experiment-programming library. *Behavior Research Methods* 39:950–958.
2. Polyn SM, Norman KA, Kahana MJ (2009) A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review* 116:129–156.
3. Solway A, Geller AS, Sederberg PB, Kahana MJ (2010) Pyparse: A semiautomated system for scoring spoken recall data. *Behavior Research Methods* 42:141–147.
4. Stricker JL, Brown GG, Wixted JT, Baldo JV, Delis DC (2002) New semantic and serial clustering indices for the california verbal learning test—second edition: Background, rationale, and formulae. *Journal of the International Neuropsychological Society* 8:425–435.
5. Roenker DL, Thompson CP, Brown SC (1971) Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin* 76:45–48.
6. Gratton G, Coles MGH, Donchin E (1983) A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology* 55:468–484.
7. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10:424–430.
8. Duda RO, Hart PE, Stork DG (2001) *Pattern classification* (Wiley, New York), 2nd edn.
9. Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310:1963–1966.
10. Sederberg PB, Kahana MJ, Howard MW, Donner EJ, Madsen JR (2003) Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience* 23:10809–10814.
11. Talairach J, Tournoux P (1988) *Co-planar stereotaxic atlas of the human brain* (Verlag, Stuttgart).
12. Lancaster JL, et al. (2000) Automated Talairach atlas labels for functional brain mapping. *Hum Brain Mapp* 10:120–131.
13. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19:1233–1239.
14. Nunez PL, Srinivasan R (2006) *Electric fields of the brain: The neurophysics of EEG* (Oxford University Press), 2nd edn.
15. Yuval-Greenberg S, Tomer O, Keren AS, Nelken I, Deouell LY (2008) Transient induced gamma-band response in EEG as a manifestation of miniature saccades. *Neuron* 58:429–441.
16. Loftus GR, Masson MEJ (1994) Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review* 1:476–490.
17. Kuhl BA, Rissman J, Wagner AD (2011) Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia* doi:10.1016/j.neuropsychologia.2011.09.002.
18. Herrmann CS, Grigutsch M, Busch NA (2005) *Event-Related Potentials: A Methods Handbook* (The MIT Press), chap. 11.