OXFORD

Cerebral Cortex, 2023, **33**, 6891–6901
https://doi.org/10.1093/cercor/bhad008
Advance access publication date 25 January 2023
Original Article

Direct brain recordings suggest a causal subsequent-memory effect

Daniel Y. Rubinstein[1,*], Christoph T. Weidemann[2,3], Michael R. Sperling[1], Michael J. Kahana[4]

[1]Department of Neurology, Thomas Jefferson University, Philadelphia, PA 19107, USA,
[2]Department of Psychology, Swansea University, Swansea SA2 8PP, UK,
[3]Department of Bioengineering, Columbia University, New York, NY 10027, USA,
[4]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

*Corresponding author: Department of Neurology, Thomas Jefferson University, 901 Walnut St. Suite 400, Philadelphia, PA 19107, USA.
E-mail: daniel.rubinstein@jefferson.edu

Endogenous variation in brain state and stimulus-specific evoked activity can both contribute to successful encoding. Previous studies, however, have not clearly distinguished among these components. We address this question by analysing intracranial EEG recorded from epilepsy patients as they studied and subsequently recalled lists of words. We first trained classifiers to predict recall of either single items or entire lists and found that both classifiers exhibited similar performance. We found that list-level classifier output—a biomarker of successful encoding—tracked item presentation and recall events, despite having no information about the trial structure. Across widespread brain regions, decreased low- and increased high-frequency activity (HFA) marked successful encoding of both items and lists. We found regional differences in the hippocampus and prefrontal cortex, where in the hippocampus HFA correlated more strongly with item recall, whereas, in the prefrontal cortex, HFA correlated more strongly with list performance. Despite subtle differences in item- and list-level features, the similarity in overall classification performance, spectral signatures of successful recall and fluctuations of spectral activity across the encoding period argue for a shared endogenous process that causally impacts the brain's ability to learn new information.

Key words: subsequent memory effect; episodic memory; free recall; neural decoding; intracranial EEG.

Introduction

Fluctuations in neural processes during encoding contribute to the likelihood of recalling a given experience (Griffiths et al. 2016; deBettencourt et al. 2018). Studies using functional neuroimaging and electrophysiological methods have demonstrated that neural activity measured during encoding of individual items reliably predicts their subsequent retrieval, an effect termed the subsequent memory effect (SME; Wagner et al. 1998; Paller and Wagner 2002; Sederberg et al. 2003; Kim 2011). However, key questions remain unanswered regarding the relationship of the SME to memory encoding states. First, most SME findings focus on individual items, where factors such as serial position, semantic characteristics or idiosyncratic autobiographical associations may independently influence encoding success regardless of ongoing cognitive processes in the brain, leading to the possibility that the SME largely reflects these exogenous factors instead of memory-related internal states (Bainbridge et al. 2019; Aka et al. 2021; Halpern et al. 2021). Second, the extent to which neural signals associated with SMEs reflect item-specific processing versus longer time-scale fluctuations in the brain's ability to encode information, remains unknown. Thus, the question of whether the SME truly captures endogenously varying memory-related states, and how to more effectively measure such states, is unresolved.

Some studies have addressed the question of exogenous versus endogenous sources of encoding by controlling for temporal order effects known to predict recall (Serruya et al. 2014; Kahana et al. 2018; Aka et al. 2021; Weidemann and Kahana 2021). Weidemann and Kahana (2021) show that even controlling for such variables, scalp EEG-recorded activity significantly predicts recall, suggesting that we can observe meaningful internal states with neural recordings. Consistent with the finding of endogenous factors underlying recall, Kahana et al. (2018) show that accounting for external predictors of recall, such as alertness, temporal order and a general measure of recallability, still leaves a large proportion of variability in performance unexplained—variability that we may be able to partially explain if neural activity can reveal internal states.

As Weidemann and Kahana (2021) demonstrate, one way to control for some of the confounds affecting memory, such as serial position, is to average neural activity and performance over multiple items. This multi-item analysis may also reveal different, longer time-scale aspects of the brain's ability to encode information, as has been shown using fMRI in a study of state-related SME (Donaldson et al. 2001; Otten et al. 2002). Findings of pre-stimulus SMEs in the hippocampus also hint at the importance of considering longer time periods in predicting memory performance (Park and Rugg 2010; Urgolites et al. 2020). As the brain cements episodic memories over extended periods of time, examining slightly longer term SMEs may reveal unique neural mechanisms and systems that contribute to encoding and integration of new memories into existing schema (Preston and Eichenbaum 2013; Sheehan et al. 2018). Thus, investigating memory encoding across multiple studied items could give more complete insight into the

Received: October 14, 2022. **Revised:** January 4, 2023. **Accepted:** January 5, 2023

physiology underlying encoding states, leading to a more accurate assessment of such states.

To this end, we constructed multivariate classifiers based on intracranial EEG recordings to predict recall over multiple items, and compared the performance of these classifiers with those predicting single item-level recall. We then investigated the temporal dynamics of classifier output to determine if multi-item level classifiers could reveal internal encoding states, acting as biomarkers of successful encoding. Finally, we examined the important neural features of multi-item classifiers and single item-level classifiers to better understand the neural underpinnings of good encoding states.

Materials and methods

From a pool of 259 total participants who performed either random or categorised free recall, we first selected 66 patients who performed both versions, to increase the amount of data for each subject and allow our results to generalise across task manipulations, although we did not account for task version in any analyses here. We then selected patients who recalled an average of at least one item per list, contributed data from at least 10 lists per session and from at least 24 lists in total (i.e. across all sessions). This resulted in a sample of 62 patients. We trained two types of classifiers to predict either individual item recall or list-level recall performance, based on neural activity during word presentation or average activity over entire lists, respectively.

Participants

All patients, who had medication-resistant epilepsy, provided informed consent to be enrolled in the Defense Advanced Research Projects Agency Restoring Active Memory project. Patients underwent neurosurgical implantation of electrodes to identify and monitor seizure activity. During this time they also performed a variety of cognitive tasks. Data were collected across the following eight participating institutions: Columbia University Hospital (New York, NY), Dartmouth-Hitchcock Medical Center (Lebanon, NH), Emory University Hospital (Atlanta, GA), Hospital of the University of Pennsylvania (Philadelphia, PA), Mayo Clinic (Rochester, MN), National Institutes of Health (Bethesda, MD), Thomas Jefferson University Hospital (Philadelphia, PA) and University of Texas Southwestern Medical Center (Dallas, TX). Experimental protocols were approved by each Institutional Review Board.

Free recall task

Patients performed two versions of a verbal delayed free recall task in which each session consisted of up to 25 lists with 12 words each. In one version, words were drawn randomly from a pool of 300 commonly used nouns (http://memory.psych.upenn.edu/WordPools). In the categorised version, words were drawn from a separate pool such that each list consisted of four words from three semantic categories (data previously published in Weidemann et al. 2019). In each list, words were displayed for 1.6 s each, with randomly jittered inter-stimulus intervals of 0.75–1 s. Each word list thus lasted ~30 s. Following list presentation, patients performed a 20-s arithmetic distractor task of simple addition problems. Finally, patients had 30 s to recall as many words as possible. Patients completed as many sessions as was comfortable. The number of total sessions completed ranged from 2 to 13; 40 patients completed between two and four sessions, and 22 completed five or more sessions.

iEEG recording and localisation

We recorded neural activity using depth and surface electrode contacts. We constructed virtual bipolar contacts by subtracting the signal between adjacent monopolar contacts, and localised them to the midpoints of the two monopolar contacts. Monopolar contacts of a given bipolar pair were located within 20 mm of each other, outside of any clinician-determined seizure onset zone or region showing epileptic spikes. We registered post- and pre-implantation imaging using Advanced Neuroimaging Tools (Avants et al. 2008). We localised surface contacts based on MRI segmentation using FreeSurfer (Desikan et al. 2006), and clinical neurophysiologists localised subcortical sources.

iEEG spectral preprocessing

We aggregated recording segments from 0.3 to 1.6 s post-word onset, for each list. We used Morlet wavelets (# cycles = 5) implemented in MNE-Python (Gramfort et al. 2013) to calculate spectral power at eight logarithmically spaced frequencies from 3 to 180 Hz (3, 5, 10, 17, 31, 56, 100, 180) with a 1.2-s buffer period on each side of each segment. For each session and frequency, we log-transformed and z-scored the power. For list-level analyses, we averaged power over all word presentation segments for each list. For analyses in which we applied the classifier to longer, continuous epochs (all 30 s before, during and after list presentation, or the 4 s surrounding each item), we averaged power over the whole list presentation period, including inter-stimulus epochs, and after calculating power, we down-sampled data to 500 Hz and averaged power over 1-s epochs incremented by 0.1 s.

List-level predictions

To predict the performance of a given list, we implemented a linear regression model using the "sklearn" "Ridge" package in Python, with an L2 regularisation parameter α of $1/(2 \times 0.0007)$, based on previously published results (Weidemann et al. 2019; Weidemann and Kahana 2021). We trained models on all but one list and predicted recall performance on the held-out, test list (with each list held out once), using features of encoding epoch power at all eight frequencies, in all contacts (eight features per contact). To normalise power, we z-scored the training set of list-level power values within session, using the mean and standard deviation of the test list's session to normalise the test list. To normalise list performance, we logit-transformed list performance, p, adjusting for performance of 0 or 100% ($P = 0$ or 1) by using: $\ln\left[\left(p + \frac{0.5}{12}\right) / \left(1 - p + \frac{0.5}{12}\right)\right]$ (Stevens et al. 2016). Otherwise, we used the standard formula: $\ln\left[p / (1 - p)\right]$. Next, we mean-subtracted list performances within each session, first omitting the held-out list's performance to protect the training data from testing data. We then subtracted the mean performance of the corresponding session from the test list's performance. Finally, we correlated predicted and observed list performance to quantify the overall performance of the model for a given subject. For parametric statistical tests, we used the Fisher transformation of the correlation values. We also performed permutation testing to obtain a distribution of 50 baseline correlation values by randomly shuffling list performances within each session and recalculating the correlation between predicted and observed list performance. Correlation values were deemed significant if greater than 95% of the baseline values.

Item-level classification

We performed item-level classification similarly to list-level prediction, except instead of using linear regression, we used logistic

regression, which is more appropriate for binary variables. We used L2-regularised logistic regression, implemented using the "sklearn" "LogisticRegression" package in Python, with a regularisation parameter C of 0.0007, to be analogous with the L2 regularisation performed in list-level predictions and to be consistent with prior work (Weidemann et al. 2019; Weidemann and Kahana 2021). We set the "balanced" class weight parameter, which adjusts weights inversely proportional to the class frequencies, to account for the imbalance of recalled and unrecalled items. For each test list, we trained the model on all items from all other lists. To evaluate the classifier, we compared observed recall with the classifier-predicted probability of recall. We also calculated classifier-predicted list performance by summing the predicted probabilities of each individual item in the held-out test list. We constructed permutation-based baseline values similarly to in list-level classification.

Recallability correction

To correct for item and list recallability in item- and list-level prediction, we first measured item recallability as in Kahana et al. (2018), by calculating the probability of item recall within each patient and averaging these probabilities across patients. For each patient, we calculated this recallability measure for each item viewed based on all other patients. During recall classification, we first used this recallability measure to predict recall for all items excluding those in the held-out list, in a linear regression model. We applied this same model to items in the held-out list. We then used the residuals from these predictions as the new recall values, which we trained the linear regression models to predict, as above, based on neural activity measures. For the list-level analysis, the individual item recallabilities were averaged over each list to generate list-level recallability measures.

Shuffled list control

To dissociate the confounding effects of item-level SME from list-level SME, we used the approach from Weidemann and Kahana (2021) and generated new lists for each subject according to the following procedure: within each session, we aggregated the recalled and unrecalled items and randomly shuffled each of them, separately. For each list within each session, we picked recalled items from the shuffled set until the list contained the same number of recalls as in the true data. We then picked the remainder from the set of unrecalled items. We excluded those picked items from being chosen for subsequent lists; that is, items were picked without replacement. Finally, we repeated the list-level performance prediction procedure described above, 50 times, to generate 50 values of correlation between observed and predicted list-level performance. We used the mean of these 50 repetitions for each subject.

Cross classification

To assess the relative performance of the classifiers, we tested the item-level classifiers on list-level recall prediction, and vice versa. First, we trained the item-level classifier as above, and tested it on the neural features of the held-out list to generate the prediction of recall. We correlated these predictions with the observed list recall performance. To test the list-level classifier on item-level prediction, we trained the list-level classifier as above and tested it on the neural features of each item of the held-out list. We correlated these predictions of recall performance with observed item-level recall (a point-biserial correlation).

Temporal analysis of encoding state

Temporal analysis of classifier output

We analysed temporal fluctuations of classifier output by first training an item- or list-level classifier as described above, except that for the training epochs of the list-level classifier, we averaged power over the whole list instead of just word presentation epochs. For each test list, we applied the trained classifier to 1-s sliding windows, from 31 s before, to 61 s after onset of the first word of the list, incrementing by 100 ms. We also applied the classifier to windows from 2 s before to 2 s after individual words, to construct a peri-stimulus time course of the classifier output. For both time courses, we averaged them for each subject to generate the subject-level mean.

Spectro-temporal analysis of classifier performance

We analysed fluctuations in classifier performance over time and frequency using a similar temporal segmentation structure as above, but focusing only on the list presentation epoch (from 1 s before, to 31 s after onset of the first word), using 10-s time windows sliding by 1 s. For each time window, we trained the list-level classifier on neural power from that time window alone and using only one of the eight frequencies, to predict performance on the entire list. We correlated these predictions of list performance with observed performances, over all lists, to quantify the classifier performance for that time-frequency cell. Similarly, we analysed fluctuations in classifier performance over time alone, by training the list-level classifier on each 10-s time window using all eight frequencies.

Correlation between spectral power and performance

For each subject, we correlated both item- and list-level performance with power at each contact for each frequency. We averaged correlations across all contacts for a given frequency to compare the item- and list-level correlations by frequency. To compare the item- and list-level correlations by region and frequency, we first aggregated contacts and averaged correlations over nine ROIs based on the grouping used in Weidemann et al. (2019): inferior frontal gyrus (IFG), middle frontal gyrus (MFG), superior frontal gyrus (SFG), temporal cortex (TC), hippocampus (HC), parahippocampal gyrus (PHG), inferior parietal cortex (IPC), superior parietal cortex (SPC) and occipital cortex (OC). We performed multiple comparisons correction for statistical tests using false discovery rate (FDR, $q < 0.05$) on permutation-based P-values (Benjamini and Hochberg 1995). We calculated P-values by randomly shuffling region labels of electrode contacts 1000 times, at the subject level, recalculating the group-level mean correlations for each permutation, and comparing these with the true mean correlation.

Results

Our investigation addressed four main questions: (i) Can we reliably classify list-level recall performance and how does the performance of these classifiers compare with standard item-level prediction? (ii) Do classifiers trained on items exhibit strong transfer to list-level recall, and vice versa? (iii) Do fluctuations in predicted recall during encoding correspond with task-relevant events, thus enabling its use as a biomarker for encoding state? (iv) Which aspects of neural activity—along the dimensions of time, frequency and region—underlie recall prediction at the item and list level?

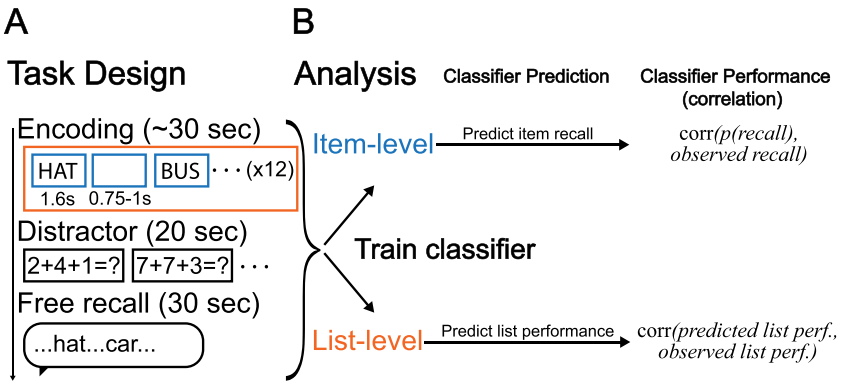

Fig. 1. Task and analytic strategy. (A) Depiction of free recall task, consisting of up to 25 repeating blocks of encoding, distractor and free recall epochs. (B) Depiction of analyses, where we generate item-level classifiers to predict probabilities of individual item recalls (*p(recall)*), and list-level classifiers to predict list-level performance. We evaluate classifier performance by correlating observed recall performance and predicted recall performance.

Item- and list-level classification of memory encoding

To evaluate list-level prediction of memory performance, we first used a leave-one-list-out scheme to predict recall performance for each list, and then correlated the classifier-generated predictions of recall performance with observed performance (Fig. 1; see Methods for details). To have a comparable measure for the evaluation of the item-level classifier, we correlated its predictions with the binary recall status of the corresponding items (i.e. a point-biserial correlation). At the individual subject level, list-level prediction was significant (permutation test, $P < 0.05$) in 40 of 62 patients. The distribution of correlations between predicted and observed list-level recall, across our 62 subjects, had a mean value of 0.25 (95% CI: [0.19, 0.31]) (t-test: $t(61) = 8.8$, SE = 0.029, $P < 0.001$). Item-level classification was significant in 60 of 62 patients, and the distribution of correlations for item-level recall had a slightly lower mean value of 0.22 (95% CI: [0.19, 0.24]) (t-test: $t(61) = 18.2$, SE = 0.012, $P < 0.001$). A statistical comparison of the item- and list-level correlations failed to detect any reliable differences ($M = 0.027$, paired t-test: $t(61) = 1.44$, SE = 0.027, $P = 0.15$). Figure 2 shows the comparable levels of performance of item- and list-level classifiers, and highlights the difference in variance.

To verify that recall predictability was not solely due to item memorability effects, we repeated the above analysis but with first measuring and correcting for item recallability (Fig. S3). After this correction, item-level classification was significant in 58 subjects, with a group mean of 0.17 (95% CI: [0.14, 0.19]) (t-test: $t(61) = 12.8$, SE = 0.013, $P < 0.001$). Although still highly significant, these predictions were significantly decreased compared with no recallability correction ($M = 0.049$, paired t-test: $t(61) = 11.3$, SE = 4.5×10^{-3}, $P < 0.001$). In contrast, list-level prediction of recall was not significantly affected by recallability correction ($M = 3.7 \times 10^{-4}$, paired t-test: $t(61) = 0.60$, SE = 3.2×10^{-3}, $P = 0.55$). List-level classification after recallability correction was significant in 39 subjects, with a group mean of 0.25 (95% CI: [0.19, 0.31]) (t-test: $t(61) = 8.7$, SE = 0.030, $P < 0.001$).

To dissociate the confounding effects of item-level SMEs from the list level, which is necessarily composed of items, we synthesised new lists while maintaining the true number of recalled items in each list. If the list-level performance predictions are significantly reduced, this would imply there is useful information at the list level that is not found at the item level. We found this to be the case that shuffling lists significantly reduced the correlation between observed and predicted list-level performance in 56 of 62 subjects, and with a group mean difference of −0.15 (95% CI: [−0.20, −0.10]) (paired t-test: $t(61) = -6.2$, SE = 0.025, $P < 0.001$).

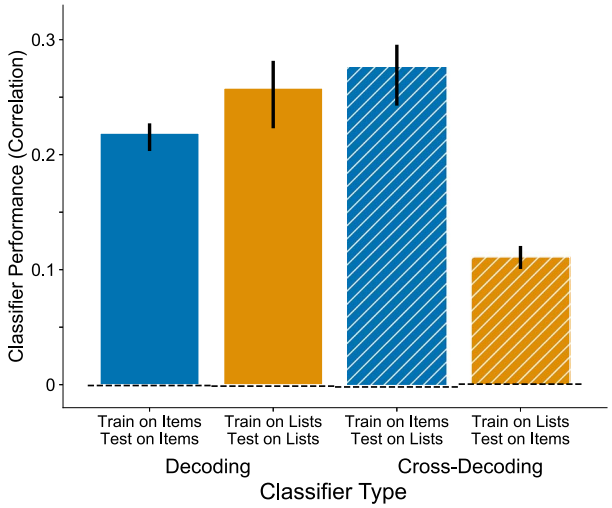

Fig. 2. Classifier performance. We quantified classifier performance as the correlation between predicted probability and observed recall (for item level) or predicted and observed list performance (for list level). We trained classifiers on either item-level recall (blue) or list-level performance (orange), and tested on either left-out items (outer bars) or lists (inner bars). For cross-decoding, we trained the classifiers on item-level recall and tested them on left-out lists (hatched blue), or vice versa (hatched orange). We used permutation testing to derive the expected null correlation (dotted baselines), in which we randomly shuffled recall performances prior to classification for each subject (# permutations = 50). Error bars indicate ±1 SEM.

Item-level classifiers exhibited significant transfer to list-level predictability, and vice versa (i.e. cross-decoding) (Fig. 2, hatched bars). Classifiers trained on item recall and tested on list performance exhibited a mean correlation of 0.27 (95% CI: [0.22, 0.32]) (t-test: $t(61) = 10.4$, SE = 0.026, $P < 0.001$). Prediction of item recall by classifiers trained on lists showed the lowest mean correlation value of 0.11 (95% CI: [0.09, 0.13]) but was still significantly positive (t-test: $t(61) = 11.1$, SE = 0.010, $P < 0.001$). Figure 2 shows that when predicting item recall, item-level classifiers performed better than list-level classifiers (paired t-test: $t(61) = 11.0$, SE = 0.010, $P < 0.001$).

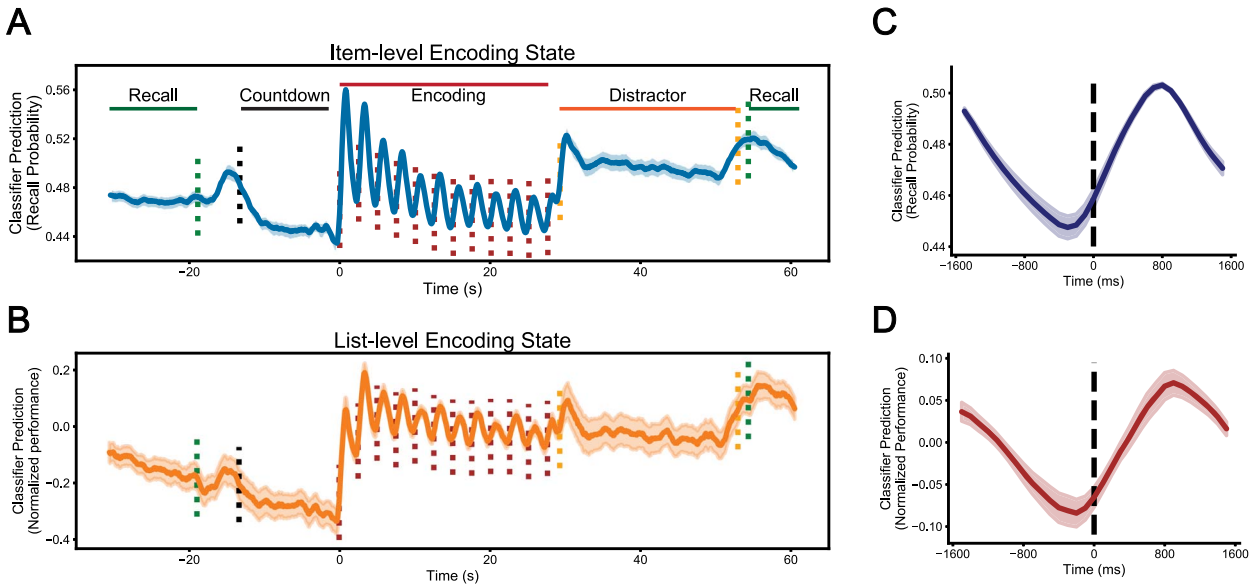

Fig. 3. Classifier predictions. (A, B) Time course of item- (A) and list-level (B) classifier predictions over the course of encoding lists, showing correspondence with task-relevant events. We normalised list-level classifier prediction by subtracting the test list session's mean performance. Dotted lines indicate the average time of task events including the start of countdown before list (black), word presentation times (brown), start/end of math distractor task (orange) and start/end of recall period (green). (C, D) Peri-stimulus time course of classifier predictions for item- (C) and list-level (D) classifier, time-locked to word presentation time, highlighting peak classifier predictions at 800 (item level) and 900 ms (list level). All shaded error regions indicate ±1 SEM.

However, when predicting list performance, list-level classifiers did not perform better than item-level classifiers (paired *t*-test: $t(61) = -1.02$, SE = 0.018, $P = 0.31$). Overall, these results demonstrate that list performance is predictable to a comparable degree as single item recall, and given the cross-decoding success, the underlying neural features of item- and list-level encoding states may largely overlap.

Temporal dynamics of classifier predictions

We next asked if classifier output could serve as a biomarker for encoding state. That is, would the temporal dynamics of classifier predictions of recall correspond to task-relevant events, and how would these dynamics differ between the two classifiers trained on item- and list-level recall performance? From the cross-decoding results showing generalisability between item- and list-level classifiers, we hypothesised that the output of the two classifiers would display similar slow but not item-level temporal patterns, as the list-level classifier has no access to this scale of neural activity. However, Fig. 3 shows that both the item- and list-level classifiers exhibit similar item-level temporal fluctuations, peaking at 800 and 900 ms after word onset, respectively (Fig. 3C and D). We also observed similarities in more slowly varying encoding state, such that both increased with onset of list presentation, distractor phase, recall period and countdown period (Fig. 3A and B). However, we also observed subtle differences at this temporal scale. Although solely descriptive and not a statistical analysis, item-level classifier predictions were highest during the first item and declined rapidly over the first half of the list, whereas list-level classifier output peaked at the second item of the list and was flatter over time.

These results show temporal fluctuations correspond to task-relevant events and phases, and suggest that classifier output may be used as a biomarker of the internal encoding state. Despite broad similarities between item- and list-level classifier outputs, especially at the single item level, minor differences suggest they reveal unique aspects of the encoding state.

Physiological substrates of classification

With evidence that the classifiers reflect internal encoding states, we next examined the neural activity supporting classification. First, were there certain times during list presentation that were more useful in predicting whole-list performance? Given that recall of early-list items better predicts overall list performance (Fig. S4), we hypothesised that early-list time windows of neural activity would also allow for better list-level recall predictions. However, list-level classifiers performed significantly better using the last time window compared with the first window (paired *t*-test: $t(61) = 2.00$, SE = 0.031 $P = 0.0498$); Fig. 4(A) shows that this increase was relatively constant over the course of the list. The pattern of increasing predictability over time was generally common across frequencies but was particularly strong in alpha and high gamma frequencies (Fig. 4B). Thus, neural activity at the end of the list may contribute more to list-level classification compared with activity at the beginning.

In addition to investigating the temporal aspects of the neural basis of multi-item encoding states, we investigated their spectral and regional aspects. Item- and list-level correlation patterns were similar, with the greatest negative correlation in the theta/alpha range, and the greatest positive correlation in the high gamma range (Fig. 5A). The greatest item-level correlations between performance and power were at 5 ($M = -0.026$, 95% CI: [−0.034, −0.018]) and 100 Hz ($M = 0.016$, 95% CI: [0.011, 0.020]). The greatest list-level correlations were at 10 ($M = -0.054$, 95% CI: [−0.067, −0.027]) and 56 Hz ($M = 0.021$, 95% CI: [0.005, 0.036]). Correlations were significant with Bonferroni correction for multiple comparisons (uncorrected p's < 0.002), except for item-level correlations at 31 Hz, and list-level correlations at 31 Hz and above.

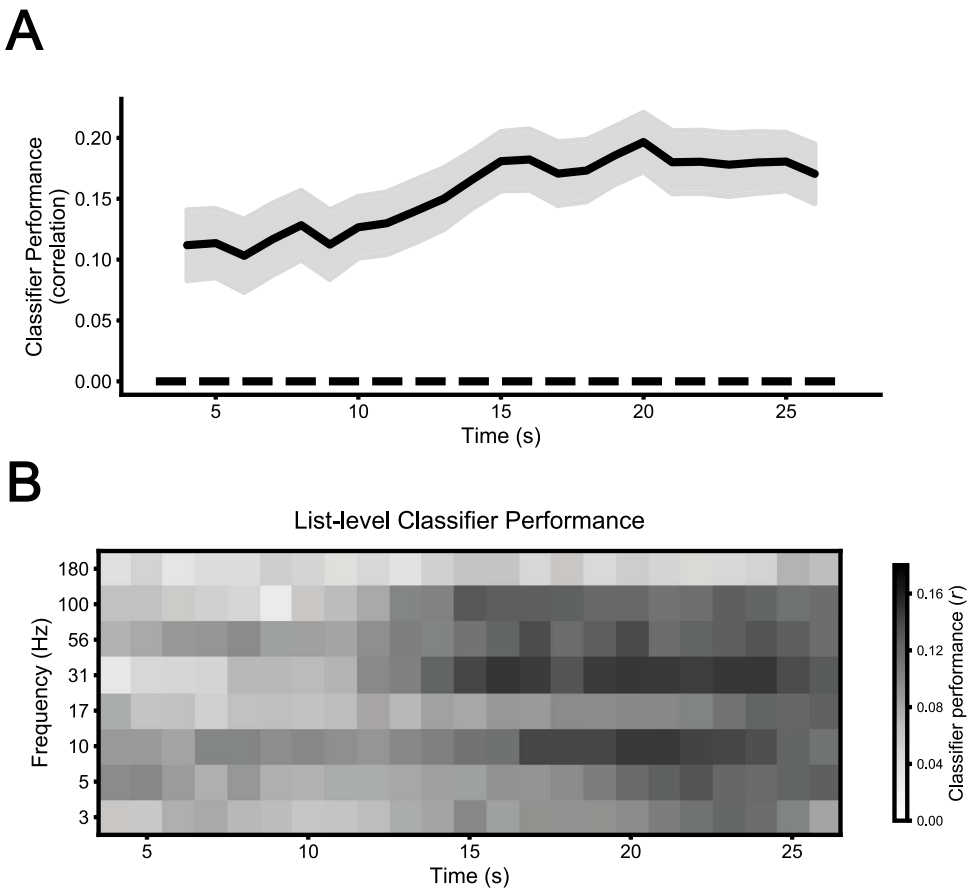

Fig. 4. Spectro-temporal analysis of list performance prediction. (A) Representation of most useful times and frequencies for prediction of list performance. For each time point, we calculated spectral power in the 10-s time window centered on the time shown, relative to list onset. We then predicted list performance using power at this time point, with leave-one-list-out cross-validation. List-level classifiers perform significantly better when using the end of the list compared with the beginning ($P < 0.05$). (B) Same as (A), except instead of using all frequencies together, we used one frequency at a time to predict list performance.

Finally, we computed correlations between performance and frequency-specific power for specific regions of interest. Item- and list-level correlation patterns by region were similar, with low-frequency negative correlations in parahippocampal regions and higher positive correlations in frontal and occipital regions. However, in contrast to item-level correlations, list-level correlations were relatively higher in frontal regions and exhibited less spectral tilt in temporal regions, especially in the hippocampus (Fig. 5B). We specifically tested the difference between item- and list-level correlations in the hippocampus and DLPFC (middle frontal gyrus) in a subset of patients who had electrode contacts in both regions ($n = 36$) and found that univariate correlations between hippocampal high-frequency activity (HFA) and recall performance were higher at the item level compared with the list level (paired t-test: $t(35) = 2.28$, SE = 0.014, $P = 0.029$). Conversely, correlations between prefrontal HFA and recall were higher at the list level (paired t-test: $t(35) = 2.52$, SE = 0.011, $P = 0.016$) (Fig. 6). Correlations between hippocampal HFA and recall performance were significantly positive at the item level with a mean of 0.010 (95% CI: [0.0008, 0.019]) (t-test: $t(35) = 2.21$, SE = 0.0044, $P = 0.033$), and although numerically negative at the list level with a mean of −0.021 (95% CI: [−0.052, 0.0097]), they were not significantly different from 0 (t-test: $t(35) = -1.39$, SE = 0.015, $P = 0.17$). Correlations between prefrontal HFA and recall were significantly positive at both the item (t-test: $t(35) = 2.29$, SE = 0.0047, $P = 0.028$) and the list level (t-test: $t(35) = 3.05$, SE = 0.013, $P = 0.0044$) with means of 0.011 (95% CI: [0.0012, 0.020]) and 0.040 (95% CI: [0.013, 0.066]), respectively. Thus, although the broad physiological phenomenon of spectral tilt was apparent in both item- and list-level SMEs, some specific patterns of regional contributions differed between the two temporal scales.

Discussion

The question of whether SMEs reflect causal internal encoding state has been unresolved due to confounding factors of item-specific characteristics. We approached answering this question by examining a longer term SME over the multiple items in a list, and examining the time course of predicted encoding, using multivariate classifiers based on intracranial recordings. We demonstrate that this approach not only controls for some item-specific aspects, but also establishes a biomarker of encoding that reveals temporal dynamics of endogenous state. We also provide a more complete understanding of the SME in general by highlighting the similar and complementary regional contributions to SMEs at different time scales.

Whether using a short window of time to predict encoding success of a single item, or a long time window to predict encoding of multiple items, we found similar performance of encoding prediction. Performance was qualitatively higher using lists compared with single items, but not significantly so.

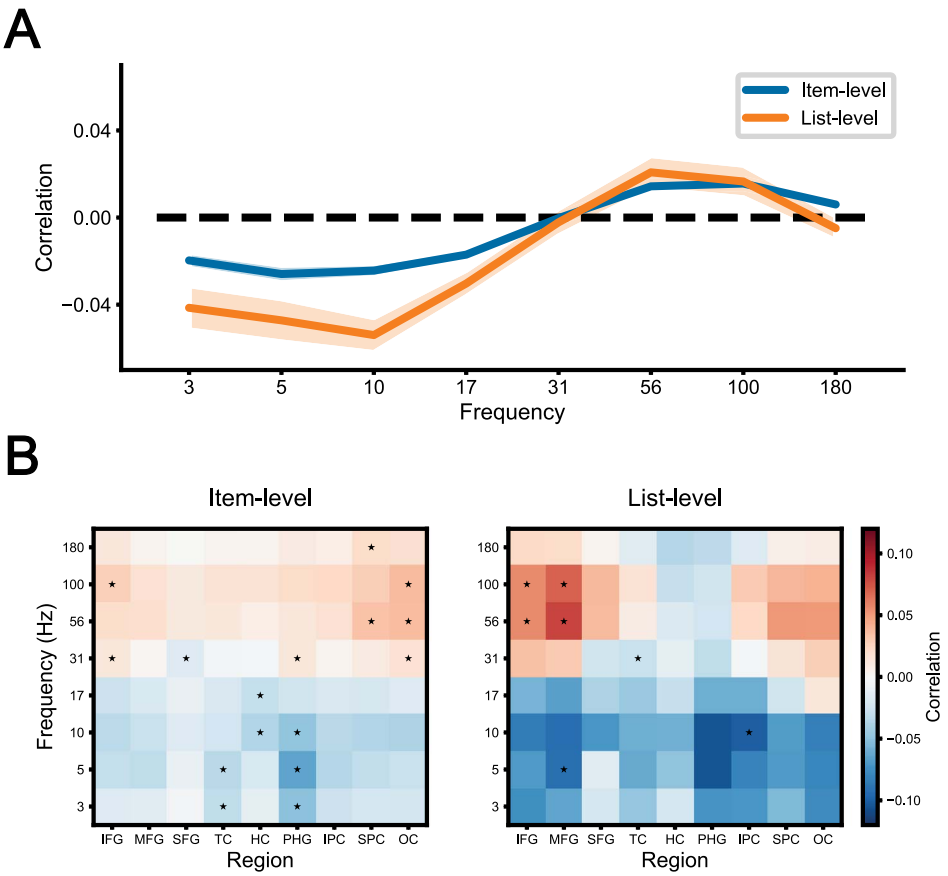

Fig. 5. Univariate correlations between power and performance. (A) We demonstrate a spectral tilt effect for item- (blue) and list-level (orange) correlations between brain-wide power and recall performance at various frequencies. Shaded area around lines indicates ±1 SEM. (B) Same as (A), except here we report correlations by region of interest (IFG: inferior frontal gyrus; MFG: middle frontal gyrus; SFG: superior frontal gyrus; TC: temporal cortex; HC: hippocampus; PHG: parahippocampal gyrus; IPC: inferior parietal cortex; SPC: superior parietal cortex; OC: occipital cortex). Asterisks in cells denote statistical significance as determined by FDR correction at $Q < 0.05$ after regional permutation analysis.

Previous work employing a nearly identical free recall task with scalp EEG also found that predictions of recall from list-based classifiers correlated slightly better with observed performance than item-based classifiers, although a direct comparison was not performed (Weidemann and Kahana 2021). Although that study sampled healthy volunteers instead of patients with epilepsy, SMEs in both populations follow similar patterns (Hill et al. 2020).

One factor that complicates our direct comparisons of item- and list-level classification is the large disparity in data quantity: item-level classifiers have 12 times the observations as the list-level classifiers, which affects the variance of the classification performance. We can somewhat account for the data disparity by repeating item-level classification analyses with smaller subsets of serial positions. Dividing the item-level classification analyses into four subgroups of serial positions (1–3, 4–6, 7–9 and 10–12) we find that the average correlations between predicted and observed recall drops from 0.22 to 0.14 (Fig. S1). There was no significant difference between the average correlation of any of the individual subgroups, which ranged from 0.13 to 0.15 (p's > 0.10). Although those subgroup-based classifiers have multiple times more data than the list-level classifiers, list-level classification correlations are likely higher due to the higher signal-to-noise ratio resulting from averaging features over longer time periods. With more data, list-level classifiers might be yet more accurate—when recalculating list-level classification performance using variable numbers of sessions in a subset of patients with the greatest number of sessions recorded, we find that classifier performance steadily increases from two to five sessions (Fig. S2).

The success of list-level classification of memory performance is consistent with, and related to, previous studies of pre-stimulus and state-related SMEs (Donaldson et al. 2001; Otten et al. 2002), especially in the hippocampus (Park and Rugg 2010; Urgolites et al. 2020), although some suggest that this signal is only relevant for recognition memory and not free recall (Merkow et al. 2014). Nevertheless, both pre-stimulus and list-level classification support the notion that neural activity outside of item presentation can still reliably predict memory encoding. With respect to list-level classification, previous work has shown that the effect is not driven solely by the predictability of the constituent items, as rearranging the items into new lists abolishes the ability to predict list-level performance (Weidemann and Kahana 2021). To dissociate the confounding effects of item- and list-level SMEs, we replicated this analysis from Weidemann and Kahana (2021) and also found that shuffling lists significantly reduced the correlation between observed and predicted list-level performance in 56 of 62 subjects, with a group mean difference of −0.15 (95% CI: [−0.20, −0.10]) (paired t-test: $t(61) = -6.2$, SE = 0.025, $P < 0.001$). Relatedly, we retrained the list-level predictions model using only the inter-stimulus intervals (700 ms preceding word presentations) instead of word presentation times, and recalculated the

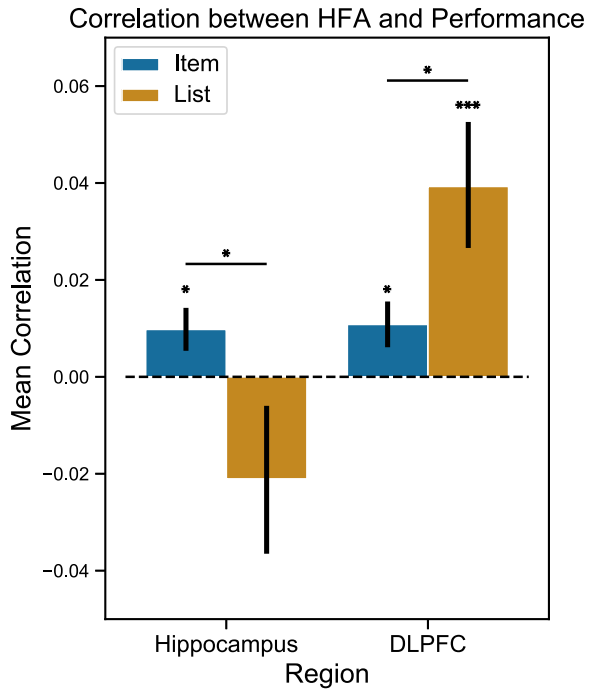

Fig. 6. DLPFC and hippocampal contribution to item- versus list-level encoding. We averaged correlations between power and recall performance over all contacts within a given region (hippocampus: left; DLPFC or middle frontal gyrus: right) and all high gamma frequencies (56, 100, 180 Hz) for both item- (blue) and list-level (orange) correlations, for each subject with contacts in both regions ($n = 36$). Hippocampal list-level correlations were not significantly different from 0 ($P = 0.17$). Paired t-tests reveal significantly greater item-level correlation with performance in the hippocampus, and greater list-level correlation in the DLPFC. Error bars indicate ±1 SEM. *$P < 0.05$, ***$P < 0.005$.

time series of encoding prediction over time. Although the resulting time series is noisier than when trained on the presentation time windows, the dynamics are overall the same (Fig. S6). This suggests that the temporal dynamics of list-level predictions in Fig. 3(B) follow item presentation timing not simply due to being trained on item-presentation windows. We also found that item-level classifiers generalised to predict list-level performance and vice versa (Fig. 2), further suggesting that neural predictors of encoding success vary slowly.

Causal versus non-causal SME

One major question arising from the SME studies has been whether correlations between neural activity and recall are causal, or rather represent correlations with external factors (Halpern et al. 2021) such as item memorability (Bainbridge et al. 2019), serial position (Murdock 1962) or semantic characteristics across items in a list (Aka et al. 2021). We argue here that external characteristics of the words cannot be the only factor underlying prediction of recall, for multiple reasons. First, even when correcting for recallability, we find significant predictive success of the classifiers (Fig. S3). Item-level classification success was significantly reduced, suggesting some contribution of item-specific effects to the item-level SME; however, average correlations between predicted and observed recall was still highly significant. Furthermore, controlling for recallability did not affect list-level classification, confirming our hypothesis that averaging over individual items would abolish some effects of item-specific characteristics.

A second confound that may theoretically underlie item-level classification is serial position, but this cannot account for the list-level classification because the training data contains no serial position information. Furthermore, despite the stronger association of early-item recall with whole-list performance, compared with later-item recall (Fig. S4), time segments later in the list provide the list-level classifier with better predictive information than time segments early in the list. This implies that predictors of list-level performance are not simply making use of aggregate item-level signals.

The inference that the SME identifies neural signals that specifically impact successful encoding of long-term memories, has recently been called into question. Specifically, Halpern et al. (2021) argue that well-known covariates of successful memory, such as serial position effects, item difficulty or linguistic properties that make certain items more memorable than others, may actually underlie previous studies claiming to identify neural correlates of success memory encoding. They report an fMRI study that controls for these variables that finds no evidence for SME in regions where it had previously been found. Here, by establishing list-level correlates of successful memory that remove any effects of serial position, and by showing that these effects remain robust even after controlling for item-level differences in word memorability, we see evidence for a causal SME.

However, as our study is inherently correlational, we cannot draw strong causal inferences about the specific brain states that lead to better memory. For example, variables that we have not controlled may lead to the observed brain states that are correlated with subsequent memory. Similarly, the observed SMEs may lead to other states that more directly cause memory formation. Although absolute causality is therefore impossible to prove, we have eliminated item-specific characteristics as the entire explanation of the SME, and therefore conclude that the neural SME used by the classifier here is likely at least partly causal.

Encoding state dynamics

Our finding that the temporal dynamics of the classifier output clearly coincide with task phases, supports the notion that the classifier output reveals internal encoding state. While the oscillations of classifier prediction with word presentation may possibly be related to item-level characteristics, the rise in prediction at the beginning of the distractor period, and especially during the recall period where there is no external stimulus, strongly suggest association with internal states untied to exogenous semantic measures. Furthermore, classifier predictions during recall increase more in high-performing lists (Fig. S5). This is not merely a continuation of higher predictions during the encoding phase, as predictions during the distractor phase are similar between low- and high-performing lists. The oscillations with word presentation were particularly notable for the list-level classifier. One expects the item-level classifier to exhibit such dynamics as the training window is limited to the item presentation window; however, the list-level classifier is trained only on the average power over the whole list, including the inter-stimulus windows, and still displays the identical pattern of encoding state peaks at 800–900 ms post word onset. This time course is consistent with previous findings that HFA-based SME peaks at around 700 ms, depending on the region (Sederberg et al. 2003; Burke et al. 2014). The two classifiers differed, however, regarding the serial position effect. The item-level encoding state exhibited a dramatic serial position effect, reminiscent of the shifts in power observed by

(Serruya et al. 2014) that also predicted subsequent recall. However, the list-level classifier exhibited a more subtle serial position effect, where encoding state peaked at the second item instead of the first, and declined to a lesser extent.

The rapid decline of the item-level classifier prediction may reflect attention-related processes or neural resources that gradually fatigue over time, and are renewed after a short break (Tulving and Rosenbaum 2006; Serruya et al. 2014). During high-performing lists as well as low-performing lists, classifier prediction of item-level recall increases at the start of the distractor period (Fig. S5). In contrast, only in low-performing lists does the list-level predicted recall increase at the start of distractor periods, suggesting that during higher memory performance (perhaps related to greater contextual clustering), list-level predictions are more memory-specific. More concrete modeling studies will further quantify these descriptive interpretations. Overall though, the time courses of item- and list-level encoding states both appear to reflect times of increased task engagement, with transient rises at task-relevant moments.

We also examined the list-level classifier performance over time, to test if certain times during list presentation were more useful in predicting whole-list performance, and thus understand how individuals encode over time. Figure 4 suggests that items are not simply encoded sequentially, with each item's encoding ceasing with the end of its presentation window. Instead, encoding of early items may continue in time such that by the end of the list, all preceding items are experiencing some degree of simultaneous encoding. Although speculative, this notion is consistent with the phenomenon of rehearsal. Specifically, when instructed to overtly rehearse during a free recall task, individuals generally rehearsed items cumulatively (Corballis 1969). This may mirror covert, silent rehearsal strategies, such that, by the end of the list, the subject is rehearsing more items than at the beginning, and therefore there may be more list-level relevant information in neural signals at the end of the list than at the beginning.

A previous investigation of encoding state based on pupil size produced a remarkably similar time course to that shown in Fig. 3(A) and (B) (Kucewicz et al. 2018). The pupil size increased more during words that were recalled than unrecalled, and also during the retrieval period, demonstrating that pupil size was sensitive to the internal state rather than simply reflecting visual stimulation. Similar to our list-level analyses (Fig. 3B), the pupil size peaked at the second word presentation of the list and similar to our item-level analyses (Fig. 3A), relative pupil size rose dramatically at the start of the distractor period. However, in contrast to the timing of our classifier predictions, which peak at 800–900 ms after word onset, the pupil size peaks at 1–2 s after word onset. Further work is needed to establish a direct correspondence between these measures of brain activity and pupil size, and to establish the extent to which they reflect specific encoding processes or more general task engagement.

Regional contributions

Having found evidence that we can gage encoding state from both single-item and list-level classifiers, we investigated how different patterns of activity contribute to successful encoding states. We found that spectral tilt across most regions was associated with better recall, similar to prior findings (Ezzyat et al. 2017). Interestingly though, the general brain-wide pattern of spectral tilt observed in item-level correlations was altered slightly in the list-level correlations, highlighting two key regions underlying memory encoding—the prefrontal cortex and hippocampus (Kuhl et al. 2012; Preston and Eichenbaum 2013). In the list-level correlations, although we observed spectral tilt patterns in association areas and especially in prefrontal cortex, we did not observe the same pattern in medial temporal areas, especially in hippocampus. This lack of spectral tilt in the hippocampus suggests that this region does not sustain its contribution to encoding state over multiple items.

Previous work has probed the idea of neural fatigue during encoding in the context of free recall, and in the process also illuminated regional differences in the dynamics of encoding state. Lohnas et al. (2020) tested the hypothesis that the hippocampus, which contributes to good encoding state through HFA (Sederberg et al. 2003; Long et al. 2014), can experience a depletion in neural resources that may then contribute to a poor encoding state. They compared hippocampal HFA during presentation of subsequently unrecalled items that followed a good encoding state (subsequently recalled items), to those that followed a similarly poor encoding state (subsequently unrecalled items), and found reduced HFA during items that followed good encoding states. In contrast to the hippocampus, which exhibited this evidence of neural fatigue, the DLPFC showed an opposite pattern of more persistent encoding state, where items following good encoding states had greater HFA than those following poor encoding states.

Building on these results, we hypothesised that these regional differences in neural fatigue would also relate to the temporal scale of hippocampal and DLPFC contributions to encoding state. Specifically, as the hippocampus fatigues at a fast rate of single items (as indexed by HFA), then hippocampal HFA should contribute to the item-level encoding state more than the list-level one. Conversely, as DLPFC HFA exhibits a more persistent contribution to encoding, there HFA should contribute more to a list- than item-level state. Indeed, our results show exactly this pattern (Fig. 6); univariate correlations between hippocampal HFA and recall performance are stronger at the item level compared with the list level, whereas correlations between DLPFC HFA and recall are stronger at the list level. Although our results align with previous findings (Lohnas et al. 2020), elucidating their relation to neural fatigue would necessitate further analyses of the temporal dynamics of encoding-related activity.

Our findings and those from Lohnas et al. (2020) support the idea of complementary roles of the prefrontal cortex and hippocampus, in which new memories may shift from their reliance on hippocampal systems to prefrontal systems, over time (Preston and Eichenbaum 2013). They also support previous work suggesting a greater role for the prefrontal cortex in coding for coarse, compared with fine, temporal context (Jenkins and Ranganath 2010). Besides contributing more to longer-term memory state, the prefrontal cortex is likely involved in modulation of other attention and memory systems during encoding (Reinhart et al. 2015). Thus, although the neural basis of the list-level encoding state overlaps significantly with the item-level state, they are also complementary, comprising different parts of a neural basis for memory encoding over longer time scales.

Conclusions

The study of memory requires (often significant) delays between encoding of the memoranda and the memory assessment. The lack of explicit responses reflecting encoding success as it happens thus presents a challenge that researchers have sought to overcome through the use of implicit measures of brain activity. Our characterisation of the temporal dynamics of brain states predicting subsequent memory performance complements previous work that strongly suggested that SMEs reflect endogenous

processes related to encoding success rather than external factors that are correlated with recall performance (Urgolites et al. 2020; Weidemann and Kahana 2021). We found that iEEG-based list-level classifiers of successful encoding perform on par with item-level classifiers, and likely reflect meaningful internal states of encoding and/or task engagement. Our work thus confirms the value of subsequent-memory analyses for the study of encoding processes, and opens up new possibilities for studying the associated dynamics.

Acknowledgments

We thank members of the Computational Memory Lab at the University of Pennsylvania for valuable advice on analyses. Most of all, we thank the patients who generously devoted their time and energy to this research. D.R. analysed data; C.W. advised on analyses; D.R., C.W. and M.K. wrote the manuscript; M.S. performed clinical duties related to data collection; M.S. and M.K. provided support. The authors declare no competing financial interests.

Author contributions

Daniel Y. Rubinstein (Conceptualisation, Formal analysis, Methodology, Software, Visualisation, Writing—original draft, Writing—review and editing), Christoph T. Weidemann (Conceptualisation, Formal analysis, Methodology, Writing—original draft, Writing—review and editing), Michael R. Sperling (Funding acquisition, Investigation, Resources, Supervision, Writing—review and editing), Michael J. Kahana (Conceptualisation, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing—original draft, Writing—review and editing).

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

This work was supported by the Defense Advanced Research Projects Agency (DARPA) Restoring Active Memory (RAM) program (cooperative agreement N66001-14-2-4032), as well as National Institutes of Health (NIH)/National Institute of Neurological Disorders and Stroke (NINDS) grants U01-NS113198 and R01-NS106611.

Conflict of interest statement: None declared.

Data availability

This manuscript has been published as a preprint on bioRxiv. Data and associated analysis code are available at http://memory.psych.upenn.edu.

References

Aka A, Phan TD, Kahana MJ. Predicting recall of words and lists. *J Exp Psychol Learn Mem Cogn*. 2021:47(5):765–784.

Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008:12(1):26–41.

Bainbridge WA, Hall EH, Baker CI. Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nat Commun*. 2019:10(1):5.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995:57(1):289–300.

Burke JF, Sharan AD, Sperling MR, Ramayya AG, Evans JJ, Healey MK, Beck EN, Davis KA, Lucas TH II, Kahana MJ. Theta and high-frequency activity mark spontaneous recall of episodic memories. *J Neurosci*. 2014:34(34):11355–11365.

Corballis MC. Patterns of rehearsal in immediate memory. *Br J Psychol*. 1969:60(1):41–49.

deBettencourt MT, Norman KA, Turk-Browne NB. Forgetting from lapses of sustained attention. *Psychon Bull Rev*. 2018:25(2):605–611.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*. 2006:31(3):968–980.

Donaldson DI, Petersen SE, Ollinger JM, Buckner RL. Dissociating state and item components of recognition memory using fMRI. *NeuroImage*. 2001:13(1):129–142.

Ezzyat Y, Kragel JE, Burke JF, Levy DF, Lyalenko A, Wanda P, O'Sullivan L, Hurley KB, Busygin S, Pedisich I, et al. Direct brain stimulation modulates encoding states and memory performance in humans. *Curr Biol*. 2017:27(9):1251–1258.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, et al. MEG and EEG data analysis with MNE-Python. *Front Neurosci*. 2013:7:267.

Griffiths B, Mazaheri A, Debener S, Hanslmayr S. Brain oscillations track the formation of episodic memories in the real world. *NeuroImage*. 2016:143:256–266.

Halpern DJ, Tubridy S, Davachi L, Gureckis TM. Identifying causal subsequent memory effects. bioRxiv. 2021:1–31. https://doi.org/10.1101/2021.11.08.467782. https://www.biorxiv.org/content/early/2021/11/10/2021.11.08.467782.

Hill PF, King DR, Lega BC, Rugg MD. Comparison of fMRI correlates of successful episodic memory encoding in temporal lobe epilepsy patients and healthy controls. *NeuroImage*. 2020:207:116397.

Jenkins LJ, Ranganath C. Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. *J Neurosci*. 2010:30(46):15558–15565.

Kahana MJ, Aggarwal EV, Phan TD. The variability puzzle in human memory. *J Exp Psychol Learn Mem Cogn*. 2018:44(12):1857–1863.

Kim H. Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. *NeuroImage*. 2011:54(3):2446–2461.

Kucewicz MT, Dolezal J, Kremen V, Berry BM, Miller LR, Magee AL, Fabian V, Worrell GA. Pupil size reflects successful encoding and recall of memory in humans. *Sci Rep*. 2018:8(1):4949.

Kuhl BA, Rissman J, Wagner AD. Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia*. 2012:50(4):458–469.

Lohnas LJ, Davachi L, Kahana MJ. Neural fatigue influences memory encoding in the human hippocampus. *Neuropsychologia*. 2020:143:107471.

Long NM, Burke JF, Kahana MJ. Subsequent memory effect in intracranial and scalp EEG. *NeuroImage*. 2014:84:488–494.

Merkow MB, Burke JF, Stein JM, Kahana MJ. Prestimulus theta in the human hippocampus predicts subsequent recognition but not recall. *Hippocampus*. 2014:24(12):1562–1569.

Murdock BB. The serial position effect of free recall. *J Exp Psychol*. 1962:64(5):482–488.

Otten LJ, Henson RNA, Rugg MD. State-related and item-related neural correlates of successful memory encoding. *Nat Neurosci*. 2002:5(12):1339–1344.

Paller KA, Wagner AD. Observing the transformation of experience into memory. *Trends Cogn Sci*. 2002:6(2):93–102.

Park H, Rugg MD. Prestimulus hippocampal activity predicts later recollection. *Hippocampus*. 2010:20(1):24–28.

Preston AR, Eichenbaum H. Interplay of hippocampus and prefrontal cortex in memory. *Curr Biol*. 2013:23(17):R764–R773.

Reinhart RMG, Woodman GF, Posner MI. Enhancing long-term memory with stimulation tunes visual attention in one trial. *Proc Natl Acad Sci USA*. 2015:112(2):625–630.

Sederberg PB, Kahana MJ, Howard MW, Donner EJ, Madsen JR. Theta and gamma oscillations during encoding predict subsequent recall. *J Neurosci*. 2003:23(34): 10809–10814.

Serruya MD, Sederberg PB, Kahana MJ. Power shifts track serial position and modulate encoding in human episodic memory. *Cereb Cortex*. 2014:24(2):403–413.

Sheehan TC, Sreekumar V, Inati SK, Zaghloul KA. Signal complexity of human intracranial EEG tracks successful associative-memory formation across individuals. *J Neurosci*. 2018:38(7): 1744–1755.

Stevens S, Valderas JM, Doran T, Perera R, Kontopantelis E. Analysing indicators of performance, satisfaction, or safety using empirical logit transformation. *BMJ*. 2016:352:i1114.

Tulving E, Rosenbaum RS. What do explanations of the distinctiveness effect need to explain? In: Hunt RR, Worthen JB, editors. *Distinctiveness and memory*. New York, NY: Oxford University Press; 2006. pp. 406–423

Urgolites ZJ, Wixted JT, Goldinger SD, Papesh MH, Treiman DM. Spiking activity in the human hippocampus prior to encoding predicts subsequent memory. *Proc Natl Acad Sci USA*. 2020:117(24): 13767–13770.

Wagner AD, Schacter DL, Rotte M, Koutstaal W, Maril A, Dale AM, Rosen BR, Buckner RL. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*. 1998:281(5380):1188–1191.

Weidemann CT, Kahana MJ. Neural measures of subsequent memory reflect endogenous variability in cognitive function. *J Exp Psychol Learn Mem Cogn*. 2021:47(4):641–651.

Weidemann CT, Kragel JE, Lega BC, Worrell GA, Sperling MR, Sharan AD, Jobst BC, Khadjevand F, Davis KA, Wanda PA, et al. Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *J Exp Psychol Gen*. 2019:148(1): 1–12.