Neural measures of subsequent memory reflect endogenous variability in cognitive function

Christoph T. Weidemann[1] & Michael J. Kahana[2]

[1]*Swansea University*

[2]*University of Pennsylvania*

Human cognition exhibits a striking degree of variability: Sometimes we rapidly forge new associations whereas at other times new information simply does not stick. Although strong correlations between neural activity during encoding and subsequent retrieval performance have implicated such "subsequent memory effects" (SMEs) as important for understanding the neural basis of memory formation, uncontrolled variability in external factors that also predict memory performance confounds the interpretation of these effects. By controlling for a comprehensive set of external variables, we investigated the extent to which neural correlates of successful memory encoding reflect variability in endogenous brain states. We show that external variables that reliably predict memory performance have only minimal effects on electroencephalographic (EEG) correlates of successful memory encoding. Instead, the brain activity that is diagnostic of successful encoding primarily reflects fluctuations in endogenous neural activity. These findings link neural activity during learning to endogenous states that drive variability in human cognition.

The capacity to learn new information can vary considerably from moment to moment. We all recognize this variability in the frustration and embarrassment that accompanies associated

memory lapses. Researchers investigate the neural basis of this variability by analyzing brain activity during the encoding phase of a memory experiment as a function of each item's subsequent retrieval success. Across hundreds of such studies, the resulting contrasts, termed subsequent memory effects (SMEs), have revealed reliable biomarkers of successful memory encoding.[1–3]

A key question, however, is whether the observed SMEs indicate endogenously varying brain states, or whether they instead reflect variation in external stimulus- and task-related variables, such as item difficulty or proactive interference, known to strongly predict retrieval success.[4] Studies characterizing SMEs generally attribute them to endogenous factors affecting encoding processes and/or to specific experimental manipulations (such as encoding instructions) aimed at directly affecting these processes.[3,5,6] At the same time, some of the strongest predictors of recall performance are characteristics of individual items (e.g., pre-experimental familiarity or position in the study list)[7–9] which are difficult to investigate, given that the successful retrieval of individual items is not under direct experimental control. Such idiosyncratic effects are therefore serious confounds in SME analyses. In cases where encoding conditions are explicitly manipulated, it is difficult to disentangle these and other external effects from ongoing endogenous fluctuations that also affect encoding success. The relative contributions of endogenous and external factors to the SME have thus yet to be established.

Here we approach these challenges in two ways using a large free-recall data set comprising 97 individuals who each had their EEG recorded while they studied and recalled 24 word lists in each of at least 20 experimental sessions that took place over the course of several weeks. Each list

contained 24 words and the same 576 words (24 words in 24 lists) were presented in each session, but their assignment to lists varied. Our use of the term "item" refers to the presentation of words within a study list rather than to a specific word identities (i.e., we do not use this term to refer to a specific word's repeated presentations across sessions). Our first approach closely builds on standard SME analyses that compute a contrast for neural activity during each item's presentation in the study list. Rather than only predicting subsequent memory as a binary variable, however, we also statistically accounted for a comprehensive list of item-specific external factors that correlate with recall performance and computed SMEs for the corresponding residuals. Comparing SMEs for these residuals with the standard item-level SME predicting binary retrieval success thus allowed us to estimate the relative contributions of endogenous neural variability and external factors to the SME (to the extent that SMEs are driven by external factors, SMEs should be absent when the effects of these external factors are statistically removed from recall performance).

Sequential dependencies in human performance[4,10–12] as well as investigations of endogenous neural fluctuations that drive variability in evoked brain activity and overt behavior[13–19] suggest that endogenous factors operate at time scales that are slower than the time allocated to the study of individual items in standard memory tasks (many seconds or minutes rather than a few seconds or less). We leveraged this presumed feature of endogenous fluctuations of cognitive function in our second approach where we computed list-level SMEs (rather than the standard item-level SMEs). Specifically, we averaged epochs of EEG activity following the presentation of individual study items within each list and used these list-averaged epochs to predict the proportion of recalled words in each list. This approach eliminates or severely reduces the effects of item-specific

external factors (because we are averaging neural activity across all study periods in a list), but the list-level SME could still reflect other external factors that also affect recall performance (such as session-level time-of-day effects or list-level proactive interference effects).[4] We therefore also statistically removed effects of list and session number (as well as effects of the average "recallability" of the words comprising each list; see methods for details) and computed SMEs for the corresponding residuals. As with the item-level SMEs, comparing the SME predicting list-level recall to the SME predicting residuals of list-level recall after accounting for external factors associated with each list and experimental session thus allowed us to estimate the relative contributions of endogenous and external factors on the list-level SME.

Results

Figure 1A shows an excerpt of a study list, associated spectral power, and output of an item-level logistic regression classifier computing an estimated probability of subsequent recall for each word. Figure 1B illustrates how neural activity associated with individual study items can be aggregated as a function of their subsequent recall. In this paper we focus on neural activity averaged between 0.3 and 1.6 s following the presentation of each study item (or list-level averages of these epochs) and Figure 1B also shows these aggregates.

As explained above, we aim to distinguish SMEs that are due to external factors predicting recall performance from those indexing endogenous variability associated with encoding success. Figure 2 illustrates the effects of one such external factor: serial position.[9] The top row shows the

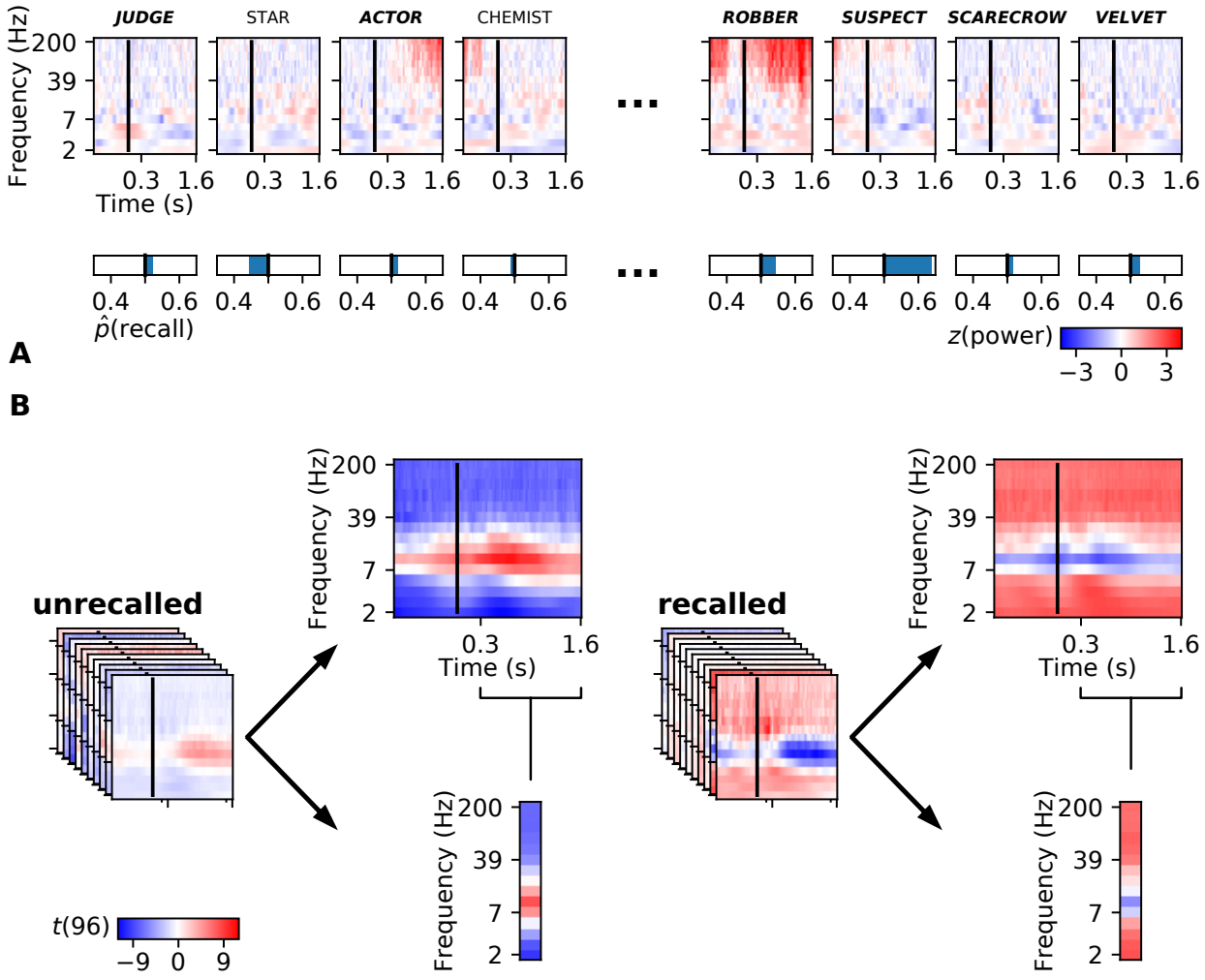

Figure 1: **A**: z-transformed power around the presentation of study words during the beginning and end of one participant's (ID: 374) 4th study list in the 16th experimental session. The study words are indicated at the top of each sub-panel with bold italic font indicating subsequent recall. The horizontal bar graphs at the bottom of each sub-panel show the output of the logistic regression classifier (see methods) centered at 0.5. **B**: Average power for subsequently unrecalled (left) and subsequently recalled (right) words during study across all lists from all participants (we averaged all data within participants and calculated the shown t-values across participants). All of our analyses were based on neural activity between 0.3 and 1.6 s following study word onset (indicated with vertical black lines) and the average power across this time interval is also illustrated. For this visualization, we aggregated EEG activity across 28 superior electrodes (see methods for details).

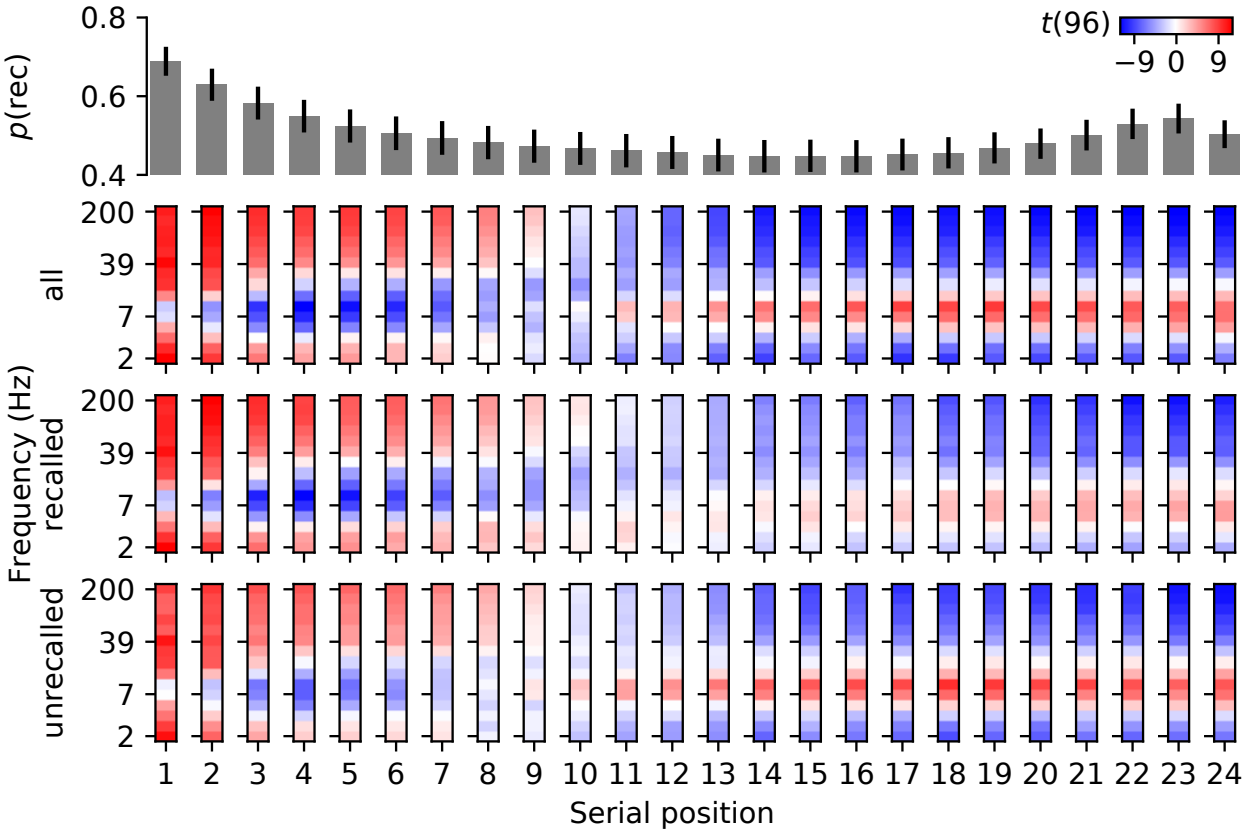

Figure 2: Mean probability of recall as a function of serial position across all participants (top row) and associated neural activity (averaged between 0.3 and 1.6 s after the onset of study items) for all, subsequently recalled, and subsequently unrecalled trials respectively (we averaged all data within participants and calculated the shown t-values across participants). Error bars indicate 95% confidence intervals. For this visualization, we aggregated EEG activity across 28 superior electrodes (see methods for details).

probability of subsequent recall across each serial position in the study list with particularly good recall performance for items presented early in the study list (a primacy effect) and a smaller benefit for items presented towards the end of each list (a recency effect). Associated neural activity for the 0.3–1.6 s post-study word onset epochs is shown across all items at a given serial position and separately as a function of subsequent recall. As is evident from the figure, serial position not only had a strong effect on recall performance, but also on associated neural activity. Indeed comparing the average neural activity across all recalled and unrecalled items shown in Figure 1B with the average neural activity at each serial position shown in Figure 2 reveals strong similarities between the pattern of neural activity for recalled items and that at early serial positions and, likewise, between the pattern of neural activity for unrecalled items and that at late serial positions. This correspondence (and similar relationships between recall performance, neural activity, and other external factors) thus severely complicates the interpretation of standard item-level SMEs.

In this paper we characterize SMEs by quantifying the information that is predictive of subsequent recall performance across all electrodes and frequency bands. For this multivariate approach we trained L2 regularized logistic regression classifiers to predict the recall status of individual items and ridge regression models (i.e., L2 regularized linear regression models) to predict residuals of recall performance for individual items as well as (logit-transformed) proportions of list-level recall and corresponding list-level residuals. Throughout this paper we assessed our ability to predict recall performance with a leave-one-session-out cross-validation procedure (see methods for details). Our multivariate models used the average neural activity between 0.3 and 1.6 s following the presentation of each study item (or list-level averages of these epochs; Figure 1B) as features.

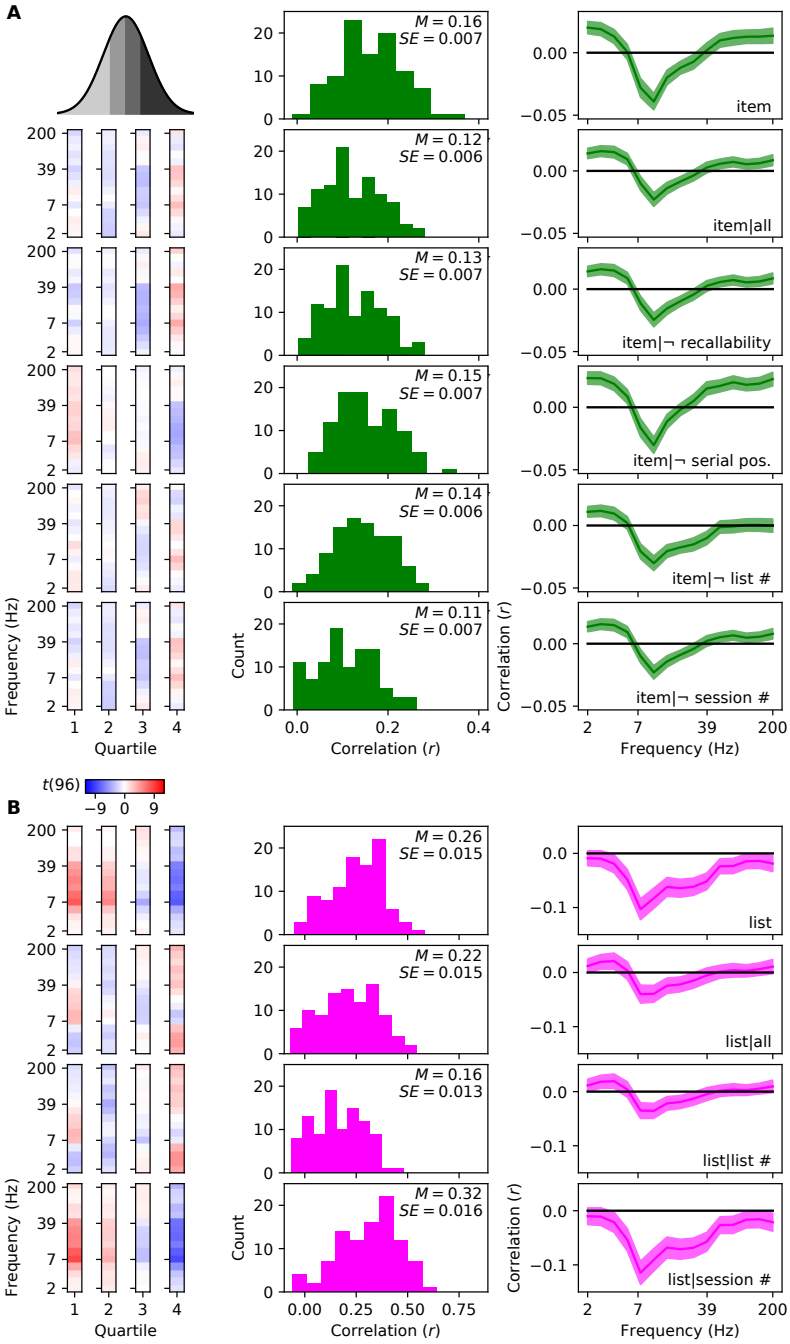

Figure 3: Item-level (**A**) and list-level (**B**) subsequent memory effects. The top row in each panel shows data for the raw SME relating to the recalled status of individual items (**A**) or the (logit-transformed) proportion of recalled items in each list (**B**). Subsequent rows show corresponding results for SMEs relating to various residuals of item- (**A**) or list-level (**B**) recall performance (indicated in the bottom right of the right sub-panel in each row). In Panel **A** "list #" only refers to the serial position of the list within the experimental session whereas in Panel **B** this label also encompasses the average recallability in each list (i.e., this label refers to both list-level external factors). The ¬ symbol in Panel **A** labels denotes that the following external factor was excluded from the regression model statistically controlling for external factors. The left column in both panels shows the neural activity associated with each quartile of (residual) recall performance averaged over the 0.3–1.6 s epochs following each study word presentation (we averaged all data within participants and calculated the shown t-values across participants). Because item-level recall performance is binary, no quartiles are shown in the top row of Panel **A**; instead a normal distribution with a different shading for each quartile illustrates the partition of the measures of (residual) recall performance in the subsequent rows (the average neural activity associated with subsequently recalled and subsequently unrecalled items is shown in Figure 1B). The middle column shows the distribution of correlations between predicted and actual recall performance indexing the performance of our classification and regression models predicting (residuals of) recall performance from neural activity. The right column shows the mean correlations between power at different frequencies and the respective (residuals of) recall performance across all participants. The black horizontal line indicates zero. Confidence regions indicate 95% CIs. For the visualizations in the left and right columns, we aggregated EEG activity across 28 superior electrodes (see methods for details).

The histogram in the middle of the top row of Figure 3A shows the (point-biserial) correlations between the output of our logistic regression classifiers predicting each item's recalled status from the corresponding brain activity during study and the actual recall performance across all participants. This correlation was significant ($M = 0.16$, $t(96) = 22.681$, $SE = 0.007$, $p < 0.001$) indicating that the different average activity patterns for recalled and unrecalled items shown in Figure 1B were indeed associated with a reliable item-level SME.

As discussed above, various external factors could be driving this SME and we therefore statistically controlled for effects of the "recallability" of individual items (which we measured by calculating how often each word was recalled by other participants in this study), the serial position of each item within the list it was presented in, the list number within each session, and the session number in the experiment. Specifically, we fit logistic regression models with these variables as predictors to classify the recalled status of each item and used the resulting residuals as a measure of recall performance for which the effects of these external factors were removed. Because our aim here is not to characterize the functional relationships between these external variables and recall performance, but to statistically remove their effects, we treated each of these variables, except for recallability, as categorical and fit separate models to each participant's data. We were thus not limited to only regressing out the linear (or higher order polynomial) effects of these variables and our approach readily accommodated inter-individual differences.

The second row of Figure 3A shows our SME analyses for the residuals controlling for the effects of recallability, serial position, list number, and session number. The first column shows

EEG activity for each quartile of these residuals suggesting that brain activity does covary with recall performance even after accounting for the effects of a comprehensive list of external factors. The corresponding histogram showing the distribution of correlations between the output of our ridge regression models predicting these residuals from brain activity and the actual residuals, indicates a robust positive correlation ($M = 0.12$, $t(96) = 19.015$, $SE = 0.006$, $p < 0.001$). This confirms a substantial SME, even after controlling for the effects of recallability, serial position, list number, and session number. The size of this SME is somewhat smaller than that for the uncorrected recall performance ($t(96) = 9.738$, $SE = 0.004$, $p < 0.001$) reflecting the fact that the uncorrected SME does include the effects of some external factors.

To better understand how the different external factors affect the SME, we repeated this analysis correcting for external factors but held out each of these factors in turn. Rows 3–6 in Figure 3A show the results of these analyses without controlling for the effects of recallability, serial position, list number, and session number respectively. All resulting correlations are positive ($M = 0.11$–0.15, $t(96) = 16.341$–22.471, $SE = 0.006$–0.007, ps < 0.001) and significantly different from the SME for uncorrected recall performance ($t(96) = 4.726$–13.438, $SE = 0.003$–0.004, ps < 0.001) as well as from that for the SME correcting for all external factors ($t(96) = 5.939$–10.790, $SE = 0.001$–0.003, ps < 0.001). This indicates that each of the external factors contributes to the difference between the size of the uncorrected and the corrected SME and that none of these factors can account for this difference in isolation. Serial position, however, explains most of this discrepancy—when controlling for all other factors, the corresponding SME is almost as large as the uncorrected SME (mean correlation of 0.15 as opposed to 0.16)

and additionally also controlling for serial position is responsible for reducing the SME to a mean correlation of 0.12.

To the extent that the uncorrected SME reflects both endogenous and external factors, we would expect that statistically removing the effects of external factors would reduce the size of the SME. Correspondingly, only partially removing effects of external factors (e.g., by holding out the removal of each of the external factors in turn like we did in the analyses described above) should result in SMEs that fall somewhere between the uncorrected SME and the SME correcting for more external factors. This is generally the pattern we found, with one notable exception: when we statistically removed the effects of all factors except for the session number, the resulting SME was slightly smaller than that for the SME also removing that effect (mean correlation of 0.11 as opposed to 0.12). This indicates that recall performance varies with session number, but that this effect of session number is not effectively captured by our measures of brain activity. Hence, when we statistically controlled for the effects of session number we removed variability in recall performance that we could not account for with our measures of brain activity leading to a slightly larger SME (and, conversely, a failure to remove the effects of session number reduced the SME).

The right column of Figure 3A shows correlations between power at different frequencies and (residual) recall performance to help illustrate the importance of different features for our regularized logistic and ridge regression models relating brain activity to (residual) recall performance. Across all measures of (residual) recall performance, correlations with spectral power were more negative in the α range (around 10 Hz) and less negative at higher and lower frequencies.

The correlations between power and uncorrected item-level recall were positive for frequencies in the γ range (> 40 Hz)—an effect that was substantially reduced for all item-level residuals, except for that not correcting for serial position. This suggests that positive correlations between γ power and recall performance largely reflect serial position effects (see also Figure 2).

As discussed above, effects of some external factors can be substantially reduced or eliminated by considering list-level rather than item-level recall performance. We thus also computed a list-level SME by relating brain activity to the (logit-transformed) proportion of words recalled in each list. The top row of Figure 3B shows the results of this analysis. The first column partitions brain activity into quartiles of list-level recall performance just like we did for the residuals of the item-level recall performance in Figure 3A. The middle column shows the performance of the ridge regression model predicting list-level recall performance on the basis of brain activity averaged over all study epochs within each list. The correlations between this model's output and the actual recall performance shown in this sub-panel indicate a substantial list-level SME ($M = 0.26$, $t(96) = 18.213$, $SE = 0.015$, $p < 0.001$). It is tempting to compare the size of this list-level SME to the item level SME shown in the top row of Figure 3A, but such direct comparisons are difficult to do sensibly. The EEG features driving the list-level SME were averaged across all study epochs within each list, whereas the item-level SME relied on features from individual epochs. Thus the neural features making up the item and list-level SMEs may differ substantially in their respective signal to noise ratios and the number of observations contributing to these different kinds of SMEs also differed considerably (in our case by a factor of 24, because each list consisted of 24 items). These differences, therefore, would need to be taken into consideration in any direct comparison.

The fact that we could demonstrate a sizable list-level SME, however, confirms our previous result that item-specific external factors are not critical drivers of the SME.

List-level analyses, however, are not immune to all external factors. Factors that are specific to a given list or session could be important sources of list-level SMEs and we thus statistically controlled for these factors to assess their contributions to the list-level SME. Specifically, we fit linear regression models to predict list-level recall performance on the basis of average recallability of items in that list, list number, and session number. As for our item-level analyses, our focus here was not on characterizing the functional relationships between these variables and recall performance, but to regress out their effects. We therefore treated list number and session number as categorical variables to allow us to account for non-linear effects of these variables and fit models separately to each participant's data to accommodate any inter-individual differences. The second row of Figure 3B shows the results for this list-level SME after statistically correcting for the effects of external factors. Even after taking into account external factors, a substantial list-level SME remained ($M = 0.22$, $t(96) = 14.332$, $SE = 0.015$, $p < 0.001$), but it was smaller than the uncorrected list-level SME ($t(96) = 5.548$, $SE = 0.008$, $p < 0.001$) reflecting the fact that external factors do contribute to the uncorrected list-level SME.

To better understand the extent to which list and session-level external factors contribute to the list-level SME, we statistically controlled for average recallability of items within each list and list number (list-level effects; third row of Figure 3B) and, separately, for session number (session-level effects; fourth row of Figure 3B). The corresponding SMEs were significant ($M = 0.16$ and

0.32, $t(96) = 12.668$ and 20.132, $SE = 0.013$ and 0.016, respectively, both ps < 0.001). Their sizes, however, fell outside the range spanned by the SME controlling for all external factors and the uncorrected SME. The SME correcting for list-level factors was smaller than that correcting for all external factors and the uncorrected SME ($t(96) = 11.606$ and 12.466, $SE = 0.005$ and 0.008, respectively, both ps < 0.001), whereas the SME correcting for session was larger than both ($t(96) = 13.134$ and 13.950, $SE = 0.009$ and 0.005, respectively, both ps < 0.001). This pattern confirms our previous finding that session-level external factors that affect recall performance are not effectively captured by our measures of brain activity. Hence, statistically controlling for their effects enhances our ability to predict residual recall performance from brain activity whereas a failure to remove that variability from recall performance reduces the SME.

Just as in Figure 3A, the right column of Figure 3B shows the correlations between power in different frequencies and the (residuals of) recall performance. The qualitative pattern of these correlations aligned with the pattern for item-level SMEs with more negative correlations in the α range and less negative correlations at lower and higher frequencies. Positive correlations between γ power and (residuals of) list-level recall performance were absent, supporting our previous interpretation that these positive correlations in item-level SMEs are largely driven by serial position effects (which are averaged out in the list-level analyses).

The presence of a robust list-level SME is compatible with endogenous factors that vary slowly (over many seconds or minutes) rather than with the presentation of individual items during the study list. Indeed, to the extent that factors driving the SME are closely linked to the pre-

sentation of individual items, characterizing these factors as "endogenous" would be problematic. To investigate the extent to which factors predicting subsequent recall are tied to individual items rather than varying more slowly over the study periods we constructed shuffled lists that mirrored the distribution of recall performance, but synthesized lists from randomly selected items within each session. This shuffling procedure practically eliminated the SME. High statistical power resulted in statistically significant deviations from zero, but the largest shuffled SME corresponded to a mean correlation of 0.03 with the residual recall performance after accounting for session effects which was an order of magnitude smaller than the corresponding unshuffled SME. All shuffled SMEs were significantly smaller than the corresponding unshuffled ones ($t(96) = 14.286$–20.361, $SE = 0.013$–0.016, ps < 0.001), supporting our previous result that endogenous factors (rather than item-specific external factors) are the main drivers of the SME.

Discussion

Whether and how a studied item is encoded and subsequently retrieved during a free recall task is, by design, not subject to complete experimental control. Indeed, recalled and not-recalled items tend to differ on a number of dimensions. Prior work has shown that neural activity just before the presentation of individual items predicts subsequent memory performance, demonstrating SMEs that are independent of specific item characteristics.[5,20–22] Nevertheless, task-related variables also strongly predict memory performance and could be driving SMEs even when they are not linked to specific item characteristics (e.g., recalled items tend to disproportionally come from early list positions, a "primacy" effect; see Figure 2).[9] Thus, any comparison of brain activity during the study

of items as a function of their subsequent recall is fraught with confounds, complicating the interpretation of the diagnostic neural signals. We addressed this issue by statistically controlling for a comprehensive list of external factors and by computing list-level SMEs aggregating brain activity over the study periods of all items within a list to predict list-level recall. Our demonstration of substantial (and similar) item- and list-level SMEs, even after accounting for a comprehensive list of external factors, shows that even though external factors contribute to SMEs, the main drivers of SMEs are endogenous fluctuations associated with encoding success. Our ability to demonstrate substantial list-level SMEs, but only for intact (rather than shuffled) lists, suggests the presence of endogenous neural variation at slow time scales (items in a study list were presented over the course of about a minute) that predicts subsequent memory.

We were able to show that a wide range of external factors were unable to account for much of the variability in EEG activity during encoding that is predictive of recall performance. To the extent that external factors in our study were correlated, however, they could have canceled each other out, rendering them invisible to our analyses. Given the large number of participants (97), sessions (20–23), and lists (24) and in light of the substantial SMEs we did observe, we believe it is unlikely that attempts to separately account for list and session-level external factors (rather than modeling their combined effects) would have yielded substantially different results. Indeed, in analyses not presented here, we have separately regressed out specific effects that were the focus of a previous study[4] with very similar results. Nevertheless, our ability to exclude external factors as drivers of the SME is limited by our ability to measure these factors.

Our analyses have thus mainly shown what SMEs are not. This naturally raises questions about the nature of endogenous factors that are associated with recall success. The prominent negative correlation between recall performance and α power (shown in Figure 3) could suggest that the endogenous factors that drive the SMEs are related to actively engaging with / attending to the encoding task.[23] Without additional measures (e.g., by recording psychophysiological measures linked to arousal or eye fixations) or direct manipulation of task engagement, however, it is difficult to conclusively characterize the endogenous factors driving the SME.

Even though we have been referring to "the SME," we have only measured it in a free recall task using EEG. Other memory tasks, such as recognition memory tasks,[3] and modalities for recording brain activity, such as functional magnetic resonance imaging,[24] also readily produce SMEs and these contrasts may well reflect different factors than those that dominate spectral power in free recall tasks. Our approach can be readily applied to other tasks and brain activity recording modalities— such extensions would be necessary to establish the extend to which our conclusions generalize to other ways of measuring memory and associated brain activity.

Previous work, has characterized specific effects of external variables affecting entire experimental sessions (such as time-of-day or sleep duration) or individual lists (such as average recallability of items within each list) on list-level recall performance.[4] Our results suggest that this distinction between session-level and list-level external factors is useful, because we could only detect contributions of list-level and item-level (but no session-level) external factors on the SME. Indeed the size of the SME, increased when we only controlled for session-level effects (and

decreased when we selectively failed to control for session-level effects) suggesting that effects of session-level external factors were not effectively captured in our measures of brain activity. Given that EEG electrodes are removed and reapplied between experimental sessions, any variability in how electrodes are applied from session to session could make such session-level effects harder to detect in the EEG signal. Likewise, EEG signals are usually normalized within session, potentially removing signal that could be attributed to session-level external factors. These or similar methodological considerations apply generally to most multi-session experiments measuring brain activity. Whereas these limits on recording brain activity over multiple sessions complicate investigations of the neural basis for session-level effects on recall performance, they also appear to effectively remove any session-level external factors from the SME.

Distinguishing between effects of external variables and endogenous processes is notoriously difficult, because it is impossible to perfectly control for effects of all possible external factors. We approached this challenge by treating list and session number as categorical predictors, effectively modeling the joint effects of external factors associated with each list and session. Whereas this approach does not guarantee that the effects of all external factors are completely removed from our residual measures of recall performance, it seems unlikely that substantial variability due to external factors remained. Indeed, we attributed any variability in recall performance that covaried with one of our external factors to that factor, even though it is likely that some of that variability could reasonably be classified as "endogenous" (e.g., sessions could be administered at different times from day to day, and effects of circadian rhythms would have been classified as an external session effect).

The fact that substantial SMEs remained after accounting for a comprehensive set of external variables may appear in conflict with findings that task context can affect the specific form of SMEs, at least for recognition memory.[5,25–28] Task context manipulations in these studies were designed to directly affect encoding processes (e.g., by asking participants to perform different tasks on the study items) and their effects on SMEs suggest that they were successful. Here we show that in the absence of direct manipulations of how study items are presented or processed, SMEs mainly reflect endogenous factors with relatively modest contributions from external factors, at least for EEG activity in a free recall task. These findings indicate that SMEs are not only effective measures of memory formation, but that they reflect endogenous states that drive variability in cognitive function.

Our findings align well with reports of sequential dependencies in human performance[4,10–12] as well as with those of slow endogenous neural fluctuations that drive variability in evoked brain activity and overt behavior.[13–19]. Previous investigations of endogenous variability in neural activity and performance have relied on exact repetitions of stimuli across many experimental trials to limit variability in external factors. In order to study the effects of endogenous variability on recall performance, we took a complementary approach by statistically removing the effects of a comprehensive set of external factors. Despite the differences in methodologies and tasks, the conclusions are remarkably consistent in establishing an important role for slowly varying fluctuations in neural activity as drivers of variability in human cognition.

Methods

Participants We analyzed data from 97 young adults (18–35) who completed at least 20 sessions in Experiment 4 of the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) in exchange for monetary compensation. This study was approved by the Institutional Review Board at the University of Pennsylvania and we obtained informed consent from all participants. Recall performance for a large subset of the current data set was previously reported,[4] but this is the first report of electrophysiological data from this experiment. Data from PEERS experiments are freely available at `http://memory.psych.upenn.edu` and have been reported in several previous publications.[29–36] Our analyses included data from all participants with at least 20 sessions.

Experimental task Each of up to 23 experimental sessions consisted of 24 study lists that each were followed by a delayed free recall test. Specifically, each study list presented 24 session-unique English words sequentially for 1,600 ms each with a blank inter-stimulus interval that was randomly jittered (following a uniform distribution) between 800 and 1,200 ms. After the last word in each list, participants were asked to solve a series of arithmetic problems of the form $A + B + C =?$ where, A, B, and C were integers in $[1, 9]$. Participants responded to each problem by typing the result and were rewarded with a monetary bonus for each correctly solved equation. These arithmetic problems were displayed until 24 s had elapsed and were then followed by a blank screen randomly jittered (following a uniform distribution) to last between 1,200 and 1,400 ms. Following this delay, a row of asterisks and a tone signaled the beginning of a 75 s free recall period. A random half of the study lists (except for the first list in each session) were also preceded by the same arithmetic distractor task which was separated from the first study-item presentation

by a random delay jittered (following a uniform distribution) to last between 800 and 1,200 ms. Each session was partitioned into 3 blocks of 8 lists each and blocks were separated by short (approximately 5 min) breaks. At each session participants were asked to rate their alertness and indicate the number of hours they had slept in the previous night.

Stimuli Across all lists in each session the same 576 common English words (24 words in each of 24 lists) were presented for study, but their arrangement into lists differed from session to session (subject to constraints on semantic similarity[29]). These 576 words were selected from a larger word pool (comprising 1,638 words) used in other PEERS experiments. The 576-word subset of this pool used in the current experiment is included as supplementary material and ranged in arousal (2.24–7.45, $M = 4.04$) and valence (1.71–8.05, $M = 5.52$) according to independent ratings on these dimensions on scales between 1 and 9.[37] Many participants also returned for a 24th session that used words from the entire 1,638-word pool, but we are not reporting data from that session here. We estimated the mean recallability of items in a list from the proportion of times each word within the list was recalled by other participants in this study.

EEG data collection and processing Electroencephalogram (EEG) data were recorded with either a 129 channel Geodesic Sensor net using the Netstation acquisition environment (Electrical Geodesics, Inc.; EGI) or with a 128 channel Biosemi Active Two system. EEG recordings were re-referenced offline to the average reference. Because our regression models weighted neural features with respect to their ability to predict (residuals of) recall performance in held out sessions, we did not try to separately eliminate artifacts in our EEG data. Data from each participant were recorded with the same EEG system throughout all sessions and for those sessions recorded with

the Geodesic Sensor net, we excluded 26 electrodes that were placed on the face and neck, rather than the scalp, from further analyses. For the visualization of EEG activity in the figures, we aggregated over electrodes 4, 5, 12, 13, 19, 20, 24, 28, 29, 37, 42, 52, 53, 54, 60, 61, 78, 79, 85, 86, 87, 92, 93, 111, 112, 117, 118, and 124 for the EGI system and electrodes A5, A6, A7, A18, A31, A32, B2, B3, B4, B18, B19, B31, B32, C2, C3, C4, C11, C12, C24, C25, D2, D3, D4, D12, D13, D16, D17, and D28 for the Biosemi system. These correspond to the superior regions of interest used we used previously.[38] All of our classification and regression models, however, used measures from all individual electrodes (with the exception of those covering the face and neck for the EGI system) as input without any averaging across electrodes. The EGI system recorded data with a 0.1 Hz high-pass filter and we applied a corresponding high-pass filter to the data collected with the Biosemi system. We used MNE,[39,40], the Python Time-Series Analysis (PTSA) library (`https://github.com/pennmem/ptsa_new`), Sklearn[41] and custom code for all analyses.

We first partitioned EEG data into epochs starting 800 ms before the onset of each word in the study lists and ending with its offset (i.e., 1,600 ms after word onset). We also included an additional 1,200 ms buffer on each end of each epoch to eliminate edge effects in the wavelet transform. We calculated power in 15 logarithmically spaced frequencies between 2 and 200 Hz, applied a log-transform, and down-sampled the resulting time series of log-power values to 50 Hz. We then truncated each epoch to 300–1,600 ms after word onset. For the item-based classifier we used each item's z-transformed mean power in each frequency across this 1,300 ms interval as features to predict subsequent recall. For the list-based regression models we averaged these values across all items in each list to predict (residuals of) list-level recall.

Removing effects of external factors For the item based analyses we fit logistic regression models separately for each participant to predict each item's recall from its average recallability (i.e., it's average probability of recall calculated from all other participants' recall data), its serial position within the study list, the list number within the current session, and the session number within the experiment. We treated all of these predictors, except for recallability, as categorical to accommodate any functional relationship between them and recall performance. This allowed us to use list and session number as predictors to model the combined effects of list and session-specific external factors rather than attempting to capture each of them separately. Furthermore, fitting these models separately to each participant's data allowed us to accommodate potentially idiosyncratic relationships between external factors and the predictors in our model as well as those between external factors and recall performance. We then calculated residuals from the full model including all item-level predictors as well as from nested models including all but one of the predictors as described in the main text. Figure 3A shows visualizations of EEG activity as a function of these residuals and histograms showing the performance of models trained to predict these residuals of item-level recall on the basis of EEG activity.

For the list-based analyses we proceeded similarly, fitting linear regression models separately for each participant to predict the logit transformed probability of recall for each list (i.e., the proportion of words that were recalled in each list). We used the average recallability of words within each list, list number within each session, and session number within the experiment as predictors (treating list and session number as categorical predictors). We again calculated residuals for the full model and also for two nested models: one including average recallability for each list and

list number (list-level predictors) and one only including session number (session-level predictor). Figure 3B shows visualizations of EEG activity as a function of these residuals and histograms of the performance of models trained to predict these residuals of list-level recall on the basis of EEG activity.

Item-based classifier For the item-based classifier we used a nested cross-validation procedure to simultaneously determine the regularization parameter and performance of L2-regularized logistic regression models predicting each item's subsequent recall. We applied this nested cross-validation approach separately to the data from each participant to accommodate idiosyncratic relationships between brain activity and recall performance and inter-individual differences in signal quality. At the top level of the nested cross-validation procedure we held out each session once—these held out sessions were used to assess the performance of the models. Within the remaining sessions, we again held out each session once—these held-out sessions from within each top-level cross-validation fold were used to determine the optimal regularization parameter, C, for Sklearn's LogisticRegression class. We fit models with 9 different C values between 0.00002 and 1 to the remaining sessions within each cross-validation fold and evaluated their performance as a function of C on the basis of the held out sessions within this fold. We then fit another logistic regression model using the best-performing C value to all sessions within each cross-validation fold and determined the model predictions on the sessions that were held-out at the top level. We determined the performance of our models solely on the basis of the predictions from these held-out sessions. There are many reasonable alternatives to for setting up these models. Our choice of L2 regularization was motivated by good performance of these models in similar data sets,[36,42] and not informed

by the current results.

Item and list-based regression models For the item- and list-based regression models we followed the same procedure as for the item-based classifier to determine the optimal level of regularization for ridge regression models predicting residuals of item-level recall or (residuals of) list-level recall performance. Specifically, we used the same nested cross-validation procedure described above to determine optimal values for α (corresponding to $1/C$), the regularization parameter in Sklearn's Ridge class, testing 9 values between 1 and 65536. We applied these models to the (logit-transformed) proportion of items recalled for each list and to the residuals from the various item- and list-level models as described in the results section.

Shuffled control lists For our list-level analyses we also computed SMEs for shuffled control lists to investigate the extent to which SMEs were linked to individual item properties or instead relied on slowly varying endogenous factors. For this approach, we separated all recalled and unrecalled items in each session, shuffled both sets of items separately, and then synthesized new lists with the original proportions of recalled and unrecalled items from the shuffled pools of recalled and unrecalled items. We repeated this procedure 20 times for each participant and concatenated the resulting shuffled lists. This shuffled session thus consisted of 20 copies of each item synthesized into 480 lists that matched the recall performance of the 24 original lists (the performance of each original list was represented 20 times in the shuffled session). We then applied all of our list-level SME analyses to these shuffled lists.

Data availability Data from this experiment are freely available at `http://memory.psych.upenn.edu`.

Code availability Data analysis code from this manuscript is freely available at `http://memory.psych.upenn.edu`.

1. Paller, K. A. & Wagner, A. D. Observing the transformation of experience into memory. *Trends in Cognitive Sciences* **6**, 93–102 (2002).

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

3. Hanslmayr, S. & Staudigl, T. How brain oscillations form memories — a processing based perspective on oscillatory subsequent memory effects. *NeuroImage* **85**, 648–655 (2014).

4. Kahana, M. J., Aggarwal, E. V. & Phan, T. D. The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2018).

5. Fellner, M.-C., Bäuml, K.-H. T. & Hanslmayr, S. Brain oscillatory subsequent memory effects differ in power and long-range synchronization between semantic and survival processing. *NeuroImage* **79**, 361–370 (2013).

6. Hanslmayr, S., Spitzer, B. & Bäuml, K.-H. Brain oscillations dissociate between semantic and nonsemantic encoding of episodic memories. *Cerebral Cortex* **19**, 1631–1640 (2009).

7. DeLosh, E. L. & McDaniel, M. A. The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**, 1136–1146 (1996).

8. Merritt, P. S., DeLosh, E. L. & McDaniel, M. A. Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. *Memory & Cognition* **34**, 1615–1627 (2006).

9. Murdock, B. B., Jr. The serial position effect of free recall. *Journal of Experimental Psychology* **64**, 482–488 (1962).

10. Gilden, D., Thornton, T. & Mallon, M. 1/f noise in human cognition. *Science* **267**, 1837–1839 (1995).

11. Mueller, S. T. & Weidemann, C. T. Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review* **15**, 465–494 (2008).

12. Verplanck, W. S., Collier, G. H. & Cotton, J. W. Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology* **44**, 273–282 (1952).

13. Monto, S., Palva, S., Voipio, J. & Palva, J. M. Very slow EEG fluctuations predict the dynamics of stimulus detection and oscillation amplitudes in humans. *Journal of Neuroscience* **28**, 8268–8272 (2008).

14. Schroeder, C. E. & Lakatos, P. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences* **32**, 9–18 (2009).

15. Arieli, A., Sterkin, A., Grinvald, A. & Aertsen, A. Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).

16. Fox, M. D., Snyder, A. Z., Zacks, J. M. & Raichle, M. E. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience* **9**, 23–25 (2005).

17. Fox, M. D., Snyder, A. Z., Vincent, J. L. & Raichle, M. E. Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron* **56**, 171–184 (2007).

18. Fox, M. D. & Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience* **8**, 700–711 (2007).

19. Raichle, M. E. The restless brain: how intrinsic activity organizes brain function. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140172 (2015).

20. Sweeney-Reed, C. M. *et al.* Pre-stimulus thalamic theta power predicts human memory formation. *NeuroImage* **138**, 100–108 (2016).

21. Otten, L. J., Quayle, A. H., Akram, S., Ditewig, T. A. & Rugg, M. D. Brain activity before an event predicts later recollection. *Nature Neuroscience* **9**, 489–491 (2006).

22. Guderian, S., Schott, B. H., Richardson-Klavehn, A. & Duzel, E. Medial temporal theta state before an event predicts episodic encoding success in humans. *Proceedings of the National Academy of Sciences* **106**, 5365–5370 (2009).

23. Sadaghiani, S. & Kleinschmidt, A. Brain networks and α-oscillations: Structural and functional foundations of cognitive control. *Trends in Cognitive Sciences* **20**, 805–817 (2016).

24. Schott, B. H. *et al.* The relationship between level of processing and hippocampal-cortical functional connectivity during episodic memory formation in humans. *Human Brain Mapping* **34**, 407–424 (2011).

25. Kamp, S.-M., Bader, R. & Mecklinger, A. ERP subsequent memory effects differ between inter-item and unitization encoding tasks. *Frontiers in Human Neuroscience* **11** (2017).

26. Summerfield, C. & Mangels, J. A. Dissociable neural mechanisms for encoding predictable and unpredictable events. *Journal of Cognitive Neuroscience* **18**, 1120–1132 (2006).

27. Otten, L. J. & Rugg, M. D. Electrophysiological correlates of memory encoding are task-dependent. *Cognitive Brain Research* **12**, 11–18 (2001).

28. Staudigl, T. & Hanslmayr, S. Theta oscillations at encoding mediate the context-dependent nature of human episodic memory. *Current Biology* **23**, 1101–1106 (2013).

29. Healey, M. K., Crutchley, P. & Kahana, M. J. Individual differences in memory search and their relation to intelligence. *Journal of Experimental Psychology: General* **143**, 1553–1569 (2014).

30. Healey, M. K. & Kahana, M. J. Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General* **143**, 575–596 (2014).

31. Healey, M. K. & Kahana, M. J. Age-related changes in the dynamics of memory encoding processes provide a biomarker of successful aging. *Manuscript Submitted for publication* (2018).

32. Lohnas, L. J. & Kahana, M. J. Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **39**, 1943–1946 (2013).

33. Siegel, L. L. & Kahana, M. J. A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **40**, 755–764 (2014).

34. Lohnas, L. J., Polyn, S. M. & Kahana, M. J. Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review* **122**, 337–363 (2015).

35. Weidemann, C. T. & Kahana, M. J. Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science* **3**, 150670 (2016).

36. Weidemann, C. T. & Kahana, M. J. Dynamics of brain activity reveal a unitary recognition signal. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **45**, 440–451 (2019).

37. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* **45**, 1191–1207 (2013).

38. Weidemann, C. T., Mollison, M. V. & Kahana, M. J. Electrophysiological correlates of high-level perception during spatial navigation. *Psychonomic Bulletin & Review* **16**, 313–319 (2009).

39. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience* **7**, 267 (2013).

40. Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446–460 (2014).

41. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

42. Weidemann, C. T. *et al.* Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *Journal of Experimental Psychology: General* **148**, 1–12 (2019).

Acknowledgements his work was supported by Grant MH55687 to MJK. We thank Ada Aka, Yuxuan Li, Nicole Kratz, Adam Broitman, Isaac Pedisich, Karl Healey, Patrick Crutchley and Elizabeth Crutchley and other members of the Computational Memory Laboratory at the University of Pennsylvania for their assistance with data collection and preprocessing and Eric Maris, Nora Herweg and Ethan Solomon for helpful comments on a previous version of this manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to C.T.W. (email: ctw@cogsci.info).