# Neural measures of subsequent memory reflect endogenous variability in cognitive function

**Christoph T. Weidemann**[1*] **and Michael J. Kahana**[2]

[1]Swansea University; [2]University of Pennsylvania

***For correspondence:** ctw@cogsci.info (CTW)

**Abstract** Humans cognition exhibits a striking degree of variability: Sometimes we rapidly forge new associations whereas at others 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.

## Introduction

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 ***Paller and Wagner*** (***2002***); ***Kim*** (***2011***); ***Hanslmayr and Staudigl*** (***2014***).

A key question, however, is whether the observed SMEs actually 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 ***Kahana et al.*** (***2018***). Despite the large number of studies that have documented and characterized SMEs across a wide range of memory tasks and encoding manipulations, the relative contributions of endogenous and external factors have yet to be established.

Free recall studies of SMEs typically compare brain activity associated with the encoding of subsequently recalled and non-recalled items within a given list. Some of the strongest predictors of recall performance are characteristics of individual items (e.g., how familiar they are or their position in the study list) ***DeLosh and McDaniel*** (***1996***); ***Merritt et al.*** (***2006***); ***Murdock*** (***1962***). Such idiosyncratic item-level effects are therefore serious confounds in item-level SME analyses and difficult to control, because repetition of items across lists would produce carry-over effects. To limit these item-level effects in our examination of broader external factors that also affect recall performance (such as session-level time-of-day effects or list-level proactive interference effects), we computed list-level SMEs. Specifically, we analyzed EEG recordings from 97 individuals who each studied and recalled 24 word lists in each of at least 20 experimental sessions that took place over the course of several weeks. We trained ridge regression models to predict the (logit-transformed) proportion of recalled items for each list, $p(\mathrm{rec})$, on the basis of spectral EEG features that we averaged over recordings during all encoding periods in that list. Additionally, we leveraged a prior statistical model of memory performance which identified several critical variables predicting recall performance across both lists and sessions ***Kahana et al.*** (***2018***). By removing linear effects of these variables, we uncovered the components of neural activity that predict the residual recallability of studied items. Comparing SMEs for these residuals with those obtained for raw recall performance thus allowed us to estimate the relative contributions of endogenous neural variability and external factors to the SME. Throughout this paper we assessed our ability to predict recall performance with a leave-one-session-out cross-validation procedure (see methods for details).

## Results

***Figure 1*** shows the mean proportion of recall as the function of several external variables that affect recall performance for entire sessions (intersession predictors) and for individual lists within each session (interlist predictors). Specifically, we considered sleep duration in the night prior to the free recall test, time of day, and self-rated alertness at the beginning of the experimental session as intersession predictors and experimental block within each session, the list number within each block, and the average "recallability" of items within each list as interlist predictors (***Kahana et al., 2018***). We are showing the effects of these variables across all participants (discretized into two bins for each of the intersession predictors and into ten bins for recallability) for illustrative purposes, but we applied all of our analyses separately to the full data from each individual. Additionally, we also considered the effect of session number (which was heterogeneous across participants with some showing increased performance with increasing practice and some showing a decline in performance) as an additional predictor in our intersession and interlist regression models (described below). Detailed analyses of the effects of these variables on recall performance in a large subset of this data set were the focus of a previous study (***Kahana et al., 2018***).

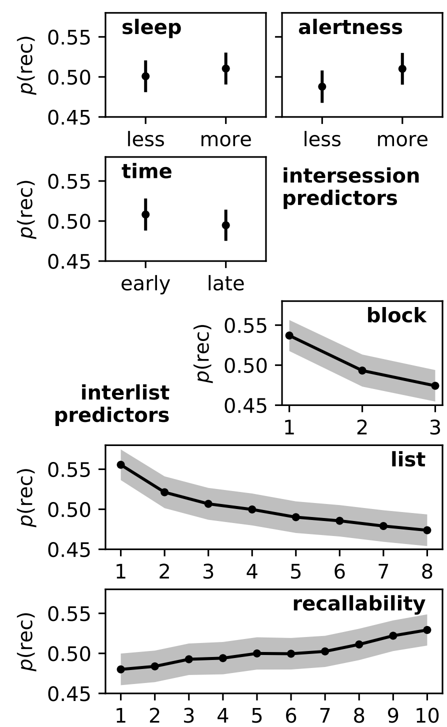


**Figure 1.** Mean probability of recall (and associated 95% confidence intervals) as a function of intersession (amount of sleep, rated alertness, and time of day) and interlist (block number within a session, list number within a block, and mean recallability of items within a list) predictors. For the purpose of this visualization we discretized each individual's intersession predictors into two bins and mean recallability scores into ten bins, but our analyses applied separately to the full data set from each individual.

Given the strong effects of item-level characteristics on recall performance it is possible that they explain a large proportion of the variance in item-level SMEs. Additionally, it is possible that any endogenous variability driving SMEs is relatively fast, varying on the order of seconds (i.e., the time devoted to the study of individual items in typical memory experiments) rather than minutes (i.e., the time encompassing a full study list) or longer. It was therefore not clear that brain activity averaged over the individual study periods would be similarly informative about

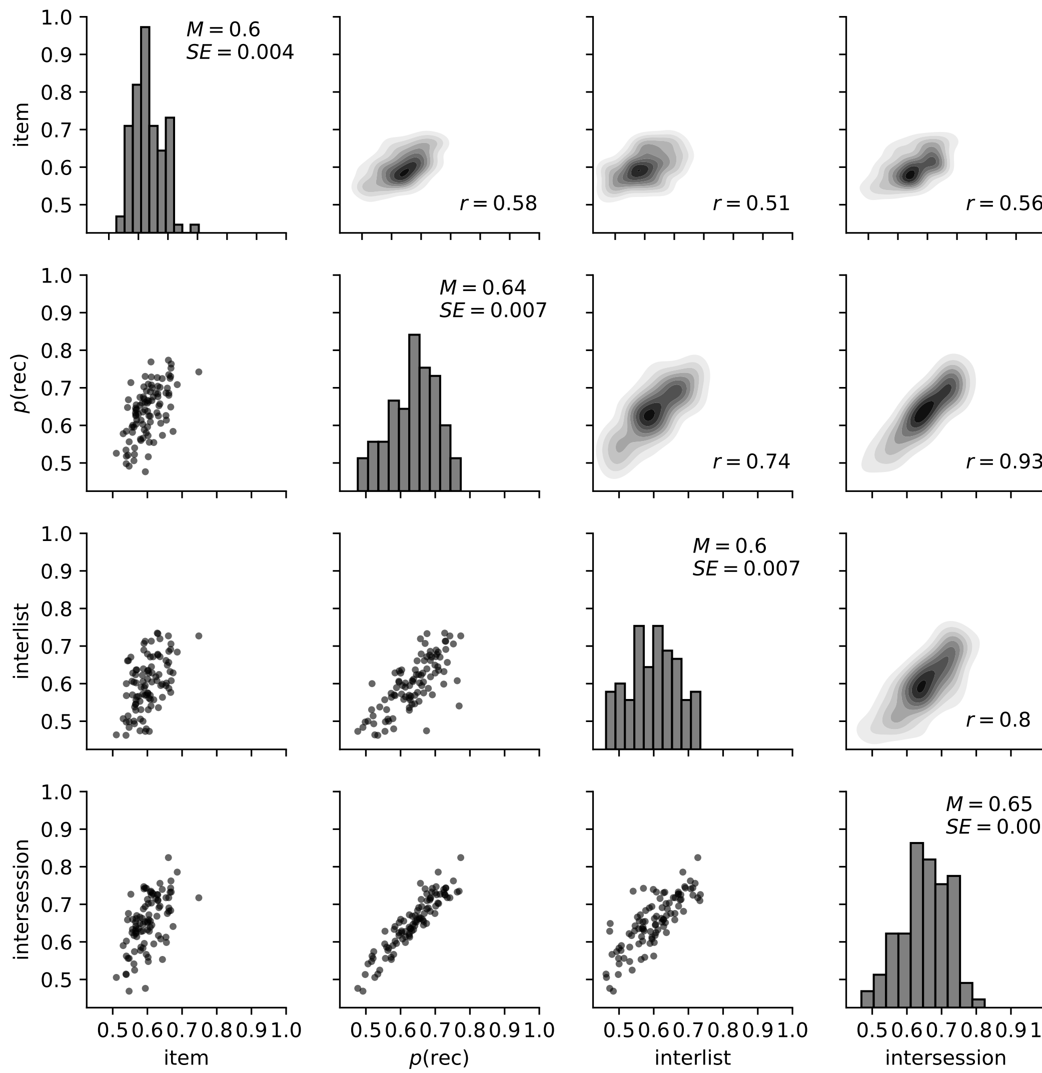


**Figure 2.** Areas under the ROC functions (AUCs) for classifier performance predicting subsequent memory for individual items (using a logistic regression classifier; item), lists of items (predicting the probability of recall across list items; $p$(rec)), as well as for residuals of list-level recall performance after regressing out interlist and intersession predictors. The lower triangle shows scatter plots for each pair of AUCs across participants. The upper triangle shows bivariate kernel density estimates of these same data with the corresponding correlations. The main diagonal shows histograms of each classifier's AUCs with the corresponding means and standard errors.

list-level recall performance as standard item-level SMEs. To compare the sizes of our list-level SME to the classic item-level SME, we trained an L2 penalized logistic regression (LR) model to predict subsequent recall of individual items (again using a leave-one-session-out cross-validation procedure to measure classification performance). For classification problems, the area under the receiver operating characteristic (ROC) function (AUC) provides a convenient index of classification performance with an AUC of 0.5 corresponding to chance performance and an AUC of 1.0 indexing perfect classification (***Fawcett, 2006***). To allow direct comparisons between the performance of the item-level classifier and our ridge regression models predicting $p(\text{rec})$ we also calculated AUCs for our regression models by discretizing the proportion of list-level recalls. Specifically, for these analyses we treated lists whose $p(\text{rec})$ exceeded the total proportion of recalled items in a session as the target category and all other lists as the the non-target category. ***Figure 2*** shows AUCs for the item-level classifier as well as for three different list-level regression models which we will discuss in turn.

The list-level regression model predicting $p(\text{rec})$ yielded a mean AUC of 0.64 which was significantly higher than that for the item-level LR classifier ($M = 0.6$; $t(96) = 6.236$, $SE = 0.006$, $p < 0.001$). This demonstrates that spectral features averaged over encoding periods effectively predict list-level recall performance. The fact that the list-level SME not only matched, but exceeded, the item-level SME therefore decisively rules out the possibility that brain activity predicting recall performance is predominantly driven by idiosyncratic item-level characteristics or by fast endogenous variation that fluctuates on the order of seconds.

Having ruled out item-level characteristics and fast endogenous variation as significant contributors to the SME, we next consider the extent to which external variables affecting recall performance for entire sessions (intersession predictors: sleep, alertness, and time of day) and those that affect recall performance at the list-level (interlist predictors: block, list, recallability) (***Kahana et al., 2018***) are driving differences in brain activity that predict recall success. To the extent that either set of variables can explain the SME, we can conclude that it also does not reflect slow endogenous variability at the level of minutes (i.e., lists) or days (i.e., sessions). We constructed interlist and intersession regression models (both models also included session number as a predictor) to remove linear effects of the respective external variables on $p(\text{rec})$. We then predicted the resulting list-level residuals with ridge regression models using the same spectral EEG features as for our list-level regression model predicting $p(\text{rec})$. ***Figure 2*** shows that AUCs for the interlist and intersession regression models respectively matched ($M = 0.6$) or exceeded ($M = 0.65$, $t(96) = 8.499$, $SE = 0.006$, $p < 0.001$) those for the item level classifier, demonstrating that spectral features effectively predict list-level performance even after accounting for linear effects from several external variables that affect recall. These results thus rule out these factors as major contributors to the SME, suggesting that SMEs predominantly reflect slow endogenous variability in cognitive function.

Whereas the AUCs for the interlist regression models were significantly lower than those of the regression models predicting $p(\text{rec})$ ($t(96) = 7.379$, $SE = 0.005$, $p < 0.001$), the AUCs for the intersession regression models exceeded those for the other list-level regression models ($t(96) = 5.812$ and $11.682$, $SE = 0.003$ and $0.005$, $p$s $< 0.001$, for comparisons with the $p(\text{rec})$ and interlist models, respectively; ***Figure 2***). This pattern of results indicates some effects of interlist factors on our measures of brain activity predicting recall performance, leading to a reduction in model performance when linear effects of interlist predictors were removed. The fact that the intersession models were better able to generalize across sessions indicates that relevant brain activity varying across sessions was not effectively captured by our models (because we used a leave-one-session-out cross-validation procedure to measure model performance, AUCs index the ability of our models to generalize across sessions). Thus, removing linear effects of intersession predictors removed variability that the models could not account for, leading to increased performance. These results establish a small role for list-level effects due to external factors (e.g., proactive interference) in the SME in addition to strong effects of endogenous variability in encoding processes.

***Figure 2*** also highlights substantial correlations between AUCs for the different models. This suggests that the different models use brain activity similarly to predict (residuals of) recall performance. It is difficult, however, to interpret the levels of these correlations in light of the fact that the dependent measures also correlate substantially—a previous analysis (***Kahana et al., 2018***) showed a reduction of variability of the residuals for the interlist and intersession models relative to $p(\text{rec})$ of only around 11% on average, leaving most of the variability in recall performance unaccounted for by external variables.

A standard measure of performance for regression models is the correlation between predicted and actual values of the dependent measures. These correlations mirror the pattern of the AUCs shown in ***Figure 2*** with $r = 0.26$, $0.29$, and $0.2$ for $p(\text{rec})$, intersession residuals, and interlist residuals respectively (all pairwise differences were statistically significant, $t(96) = 8.463$–$11.533$, $SE = 0.003$–$0.008$, $p\text{s} < 0.001$). The point-biserial correlation between predictions from the item-level classifier and recall status of individual items was $0.16$. This confirms the above AUC-based analyses indicating the effectiveness of spectral features in predicting list-level performance and the ability of our models to capture some brain activity associated with interlist, but not intersession, predictors (because of the better performance for the intersession models and the reduced performance of the interlist models relative to the models predicting $p(\text{rec})$ as explained above). Likewise, as in the above analyses, none of the list-level SMEs fell short of the item-level SME suggesting that brain activity that is predictive of recall success is mainly driven by slow endogenous variability.

In addition to investigating the correlations between predictions from the different regression models and the corresponding dependent measures, we can also assess the extent to which the different models generalize to predicting the other measures.[1] This analysis reveals an advantage for models trained on intersession residuals, even when these were tested on $p(\text{rec})$ or interlist residuals. To assess the size of these differences, we removed the linear effects of the measure each model was trained on from the generalization measures and computed the (semi-partial) correlations between the model predictions and the resulting residuals. The semi-partial correlations between predictions of models trained on intersession-residuals and the other two measures were positive ($M = 0.1$ for both $p(\text{rec})$ and interlist-residuals; $t(96) = 17.324$ and $13.731$, $SE = 0.006$ and $0.008$, $p\text{s} < 0.001$, respectively). This confirms that the performance advantage for models trained on intersession residuals generalizes to the prediction of $p(\text{rec})$ and interlist residuals—a result that complements the above finding suggesting that removing linear effects of intersession predictors eliminates variability in recall performance that is not effectively captured by our measures of brain activity. The only other semi-partial correlations significantly deviating from 0 were those between predictions of the models trained on $p(\text{rec})$ and the interlist-residuals ($M = 0.07$, $t(96) = 10.158$, $SE = 0.007$, $p < 0.001$), reflecting the fact that models trained on $p(\text{rec})$ were better able to capitalize on brain activity that is relevant for predicting recall performance than models that could not make use of brain activity that reflects interlist predictors.

***Figure 2*** showed high correlations between the performances of the different models predicting item and list-level recall which suggests that there is considerable overlap between the patterns of brain activity predicting these measures. We investigated this relationship by correlating power across a range of frequencies and regions of interest (ROIs) with each of the measures of recall performance. These correlations exhibited a consistent pattern with low (negative) correlations in the $\theta$ and $\alpha$ range ($\approx 5$–$10$ Hz) which increased for higher (and lower) frequencies (***Figure 3***). For the (point-biserial) correlation of brain activity with item-level recall, we observed negative correlations in the $\theta$ and $\alpha$ range and positive correlations in the $\gamma$ ($> 30$ Hz) range, consistent with numerous findings showing that decreased power in lower frequencies and increased power in higher frequencies predicts subsequent memory (***Hanslmayr et al., 2012***; ***Burke et al., 2014***; ***Long and Kahana, 2015***; ***Weidemann et al., 2019***). As shown in ***Figure 3***, the correlations for the list-level measures of recall performance exhibited qualitatively very similar patterns, confirming that the

[1] This is conceptually similar to a cross-decoding approach where models trained on one data set are used for predictions on a different data set (***Weidemann et al., 2019***). In the current application we train models on identical features to predict different measures of recall performance rather than predicting the same dependent measure in different data sets.

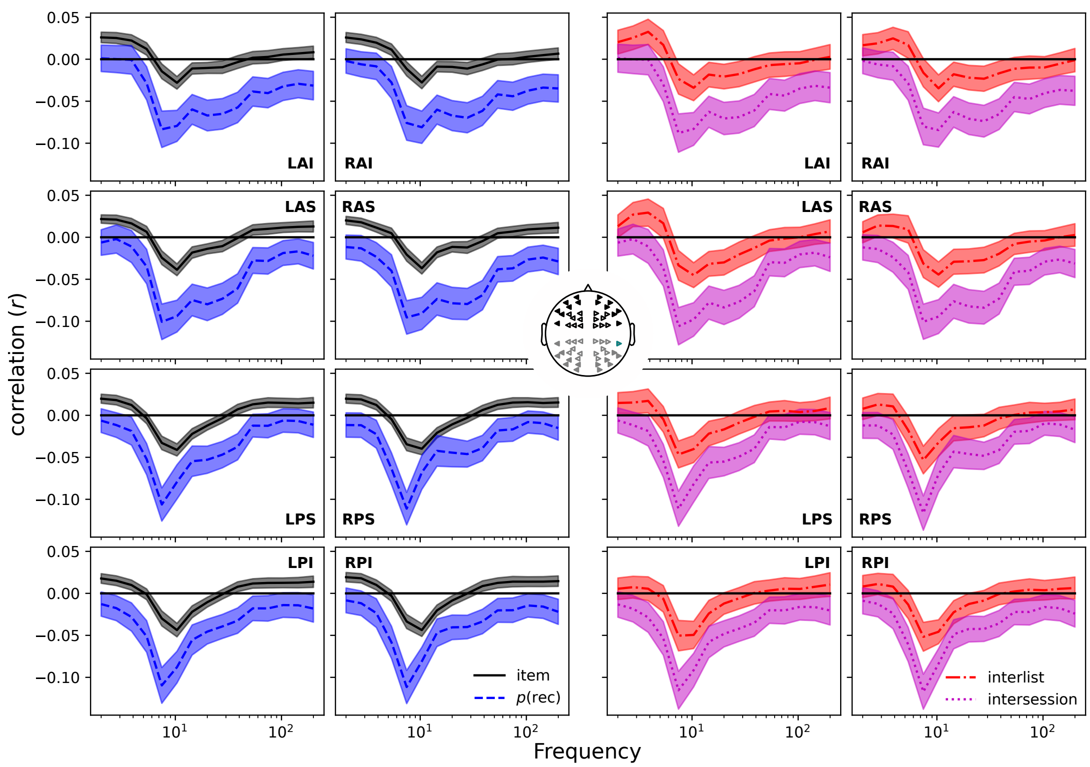


**Figure 3.** Correlations between mean power in each frequency across electrodes within each region of interest (ROI) and measures of recall performance (recall of individual items, $p$(rec), and residuals from the interlist and intersession models). The inset in the middle of the figure illustrates the locations of the ROIs and each panel includes an ROI label with the first letter indicating the hemisphere (L: left, R: right), the second letter distinguishing between anterior (A) and posterior (P) ROIs, and the last letter specifying the ROI position as either inferior (I) or superior (S). Zero is indicated as are 95% confidence intervals (shaded regions).

different ways of calculating SMEs leverage brain activity in similar ways.

The similarity in how brain activity correlates with different measures of recall performance complements our analysis of correlations between AUCs associated with different regression models (***Figure 2***). Just like that analysis, however, this similarity is difficult to interpret in light of substantial correlations between the dependent measures. To directly assess how brain activity covaries with variability that is specific to intersession and interlist predictors (removing linear effects of $p(\mathrm{rec})$), we therefore correlated brain activity with corresponding residuals (intersession$|p(\mathrm{rec})$ and interlist$|p(\mathrm{rec})$, respectively; ***Figure 4***). As is evident from ***Figure 4***, correlations of brain activity with intersession$|p(\mathrm{rec})$ residuals were close to zero and varied little across frequencies or ROIs, consistent with the above analyses indicating that our measures of brain activity did not capture much of the variability in recall performance associated with intersession predictors. The correlations of brain activity with interlist$|p(\mathrm{rec})$, however, were relatively strong, complementing the above analyses indicating that our measures of brain activity are sensitive to interlist predictors of recall performance.

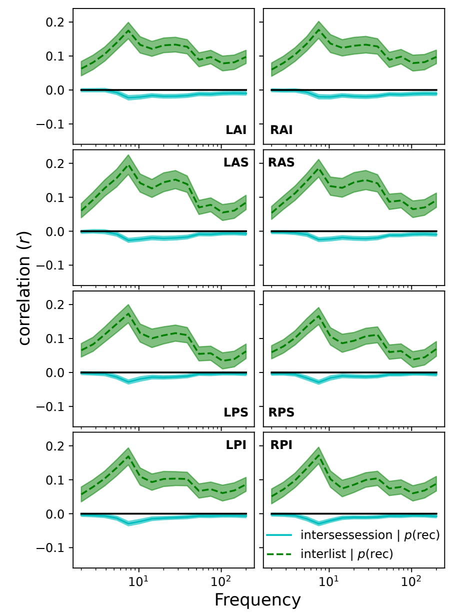


**Figure 4.** Correlations between mean power in each frequency across electrodes within each region of interest (ROI) and intersession and interlist residuals after regressing out linear effects of $p(\mathrm{rec})$ (intersession$|p(\mathrm{rec})$ and intersession$|p(\mathrm{rec})$ respectively). Each panel shows these correlations for a different ROI (labeled and arranged as in ***Figure 3***). Zero is indicated as are 95% confidence intervals (shaded regions).

## 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 (***Sweeney-Reed et al., 2016***; ***Otten et al., 2006***; ***Fellner et al., 2013***; ***Guderian et al., 2009***). 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) (***Murdock, 1962***). 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 avoided some of these confounds by assessing list-level SMEs, aggregating brain activity across the study periods of all items within a list to predict list-level recall. Our demonstration that list-level SMEs were stronger than item-level SMEs (***Figure 2***) with similar predictive patterns of brain activity (***Figure 3***), shows that item-level SMEs are not mainly driven by external variables differentiating items within a study list. This result also 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.

Even when aggregating across items within a list, a range of confounding variables remain. By studying 97 individuals who each participated in up to 23 experimental sessions, we were able to model the effects of several external variables that affect list-level recall performance. This enabled us to not only relate brain activity to the proportion of recalled items in each list, but also to residuals of recall performance after accounting for effects of these external variables. Following previous work (***Kahana et al., 2018***), we partitioned these external variables into those that varied across lists (interlist) and those that varied across sessions (intersession). Accounting for interlist variables reduced the list-level SME slightly (but not below the level of the item-level SME, Figures 2 and 3). This suggests that some, but not all, of the list-level SME reflects the effects of interlist variables. Accounting for intersession variables, on the other hand, slightly increased the size of the SME, demonstrating that the list-level SME does not include substantial contributions from these variables (Figures 2 and 3; see also ***Figure 4***).

Distinguishing between effects of external variables and endogenous processes is notoriously difficult, because it is impossible to control for effects of all possible external factors. Additionally some external factors (e.g., drug consumption or exercise) can have long-lasting and/or variable effects, making it difficult to establish their relationship with behavior. Indeed, the distinction between external and endogenous effects can be blurry, especially when external variables (such as time of day) correlate with endogenous processes (e.g., physiological changes due to circadian rhythms). In our investigation of variability in recall performance, we controlled for the major variables known to affect episodic memory. We also considered broad variables (such as recallability, time of day, and alertness) that were meant to capture the joint effects of large sets of more specific variables (e.g., features of the individual words within a study list, number of waking hours, or effects of caffeine consumption). Thus, we believe that the joint effects of external variables beyond those considered as predictors in our interlist and intersession models are likely to be too small to account for a substantial fraction of the remaining variability in recall performance or the SME.

When we controlled for the effects of sleep, alertness, and time of day, our ability to predict list-level recall from brain activity increased. This indicates that these variables did not substantially contribute to the list-level SME we observed (and hence removing their effects improved generalization of our models). Our results thus highlight the need to distinguish between variables that affect recall performance and and those whose effects manifest in our measures of brain activity. Considering additional variables that affect recall performance therefore need not reduce our estimate of the contributions of endogenous factors to the SME.

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 (***Kamp et al., 2017***; ***Summerfield and Mangels, 2006***; ***Otten and Rugg, 2001***; ***Staudigl and Hanslmayr, 2013***; ***Fellner et al., 2013***). 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, external variables do not substantially contribute to the SME even when they predict subsequent recall. 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 (***Kahana et al., 2018***; ***Gilden et al., 1995***; ***Mueller and Weidemann, 2008***; ***Verplanck et al., 1952***) as well as with those of slow endogenous neural fluctuations that drive variability in evoked brain activity and overt behavior (***Monto et al., 2008***; ***Schroeder and Lakatos, 2009***; ***Arieli et al., 1996***; ***Fox et al., 2005***, ***2007***; ***Fox and Raichle, 2007***; ***Raichle, 2015***). 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 neural fluctuations in human cognition.

# Methods and Materials

## 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. Recall performance for a large subset of the current data set was previously reported (***Kahana et al., 2018***), 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 (***Healey et al., 2014***; ***Healey and Kahana, 2014***, ***2018***; ***Lohnas and Kahana, 2013***; ***Siegel and Kahana, 2014***; ***Lohnas et al., 2015***; ***Weidemann and Kahana, 2016***, ***2019***). 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 (***Healey et al., 2014***)). 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 were selected to maximize homogeneity, by removing words that were atypical in frequency, concreteness, or emotional valence. 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. 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 (***Gramfort et al., 2013, 2014***), the Python Time-Series Analysis (PTSA) library (https://github.com/pennmem/ptsa_new), Sklearn (***Pedregosa et al., 2011***) 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 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.

For the analyses shown in Figures 3 and 4, we partitioned electrodes into the 6 regions of interest (ROIs) illustrated in ***Figure 3***. This choice of ROIs follows a range of studies that used these or very similar ROIs to characterize the spatial distribution of EEG effects (***Weidemann et al., 2009***). All of our classification and regression models, however, used measures from individual electrodes as input without any averaging into ROIs.

## 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. 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 calculated the area under the ROC function on the basis of the predictions from these held-out sessions.

## List-based regression models

For the 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) list-level recall performance. Specifically, we used the same nested cross-validation procedure described above to determine optimal values for $\alpha$ (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, $p(\text{rec})$, as well as to the residuals from the interlist and intersession models as described in the results section (***Kahana et al., 2018***).

## Data availability

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

## Acknowledgements

This work was supported by Grant MH55687 to MJK. We thank Ada Aka, Effie Li, Nicole Kratz, Adam Broitman, Isaac Pedesich, 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 Nora Herweg and Ethan Solomon for helpful comments on a draft of this manuscript.

## References


**Arieli A**, Sterkin A, Grinvald A, Aertsen A. Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses. Science. 1996; 273:1868–1871. doi: 10.1126/science.273.5283.1868.

**Burke JF**, Long NM, Zaghloul KA, Sharan AD, Sperling MR, Kahana MJ. Human intracranial high-frequency activity maps episodic memory formation in space and time. NeuroImage. 2014; 85 Pt. 2:834–843. doi: 0.1016/j.neuroimage.2013.06.067.

**DeLosh EL**, McDaniel MA. The role of order information in free recall: Application to the word-frequency effect. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22:1136–1146. doi: 10.1037/0278-7393.22.5.1136.

**Fawcett T**. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27:861–874. doi: 10.1016/j.patrec.2005.10.010.

**Fellner MC**, Bäuml KHT, Hanslmayr S. Brain oscillatory subsequent memory effects differ in power and long-range synchronization between semantic and survival processing. NeuroImage. 2013; 79:361–370. doi: 10.1016/j.neuroimage.2013.04.121.

**Fox MD**, Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature Reviews Neuroscience. 2007; 8:700–711. doi: 10.1038/nrn2201.

**Fox MD**, Snyder AZ, Vincent JL, Raichle ME. Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior. Neuron. 2007; 56:171–184. doi: 10.1016/j.neuron.2007.08.023.

**Fox MD**, Snyder AZ, Zacks JM, Raichle ME. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. Nature Neuroscience. 2005; 9:23–25. doi: 10.1038/nn1616.

**Gilden D**, Thornton T, Mallon M. 1/f noise in human cognition. Science. 1995; 267:1837–1839. doi: 10.1126/science.7892611.

**Gramfort A**, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience. 2013; 7:267. doi: 10.3389/fnins.2013.00267.

**Gramfort A**, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS. MNE software for processing MEG and EEG data. NeuroImage. 2014; 86:446–460. doi: 10.1016/j.neuroimage.2013.10.027.

**Guderian S**, Schott BH, 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. 2009; 106(13):5365–5370. doi: 10.1073/pnas.0900289106.

**Hanslmayr S**, Staudigl T. How brain oscillations form memories — A processing based perspective on oscillatory subsequent memory effects. NeuroImage. 2014; 85:648–655. doi: 10.1016/j.neuroimage.2013.05.121.

**Hanslmayr S**, Staudigl T, Fellner MC. Oscillatory power decreases and long-term memory: the information via desynchronization hypothesis. Frontiers in Human Neuroscience. 2012; 6. doi: 10.3389/fnhum.2012.00074.

**Healey MK**, Kahana MJ. Is memory Search Governed by Universal Principles or Idiosyncratic Strategies? Journal of Experimental Psychology: General. 2014; 143:575–596. doi: 10.1037/a0033715.

**Healey MK**, Kahana MJ. Age-Related Changes in the Dynamics of Memory Encoding Processes Provide a Biomarker of Successful Aging. Manuscript Submitted for publication. 2018; .

**Healey MK**, Crutchley P, Kahana MJ. Individual differences in memory search and their relation to intelligence. Journal of Experimental Psychology: General. 2014; 143:1553–1569. doi: 10.1037/a0036306.

**Kahana MJ**, Aggarwal EV, Phan TD. The variability puzzle in human memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2018; doi: 10.1037/xlm0000553.

**Kamp SM**, Bader R, Mecklinger A. ERP Subsequent Memory Effects Differ between Inter-Item and Unitization Encoding Tasks. Frontiers in Human Neuroscience. 2017; 11. doi: 10.3389/fnhum.2017.00030.

**Kim H**. Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. NeuroImage. 2011; 54(3):2446–2461. doi: 10.1016/j.neuroimage.2010.09.045.

**Lohnas LJ**, Kahana MJ. Parametric Effects of Word Frequency in Memory for Mixed Frequency Lists. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2013; 39:1943–1946. doi: 10.1037/a0033669.

**Lohnas LJ**, Polyn SM, Kahana MJ. Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. Psychological Review. 2015; 122:337–363. doi: 10.1037/a0039036.

**Long NM**, Kahana MJ. Successful memory formation is driven by contextual encoding in the core memory network. NeuroImage. 2015; 119:332–337. doi: 10.1016/j.neuroimage.2015.06.073.

**Merritt PS**, DeLosh EL, McDaniel MA. Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. Memory & Cognition. 2006; 34:1615–1627. doi: 10.3758/bf03195924.

**Monto S**, Palva S, Voipio J, Palva JM. Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and Oscillation Amplitudes in Humans. Journal of Neuroscience. 2008; 28:8268–8272. doi: 10.1523/jneurosci.1910-08.2008.

**Mueller ST**, Weidemann CT. Decision noise: An explanation for observed violations of signal detection theory. Psychonomic Bulletin & Review. 2008; 15:465–494. doi: 10.3758/pbr.15.3.465.

**Murdock BB Jr**. The serial position effect of free recall. Journal of Experimental Psychology. 1962; 64:482–488. doi: 10.1037/h0045106.

**Otten LJ**, Quayle AH, Akram S, Ditewig TA, Rugg MD. Brain activity before an event predicts later recollection. Nature Neuroscience. 2006; 9(4):489–491. doi: 10.1038/nn1663.

**Otten LJ**, Rugg MD. Electrophysiological correlates of memory encoding are task-dependent. Cognitive Brain Research. 2001; 12:11–18. doi: 10.1016/s0926-6410(01)00015-5.

**Paller KA**, Wagner AD. Observing the transformation of experience into memory. Trends in Cognitive Sciences. 2002; 6:93–102. doi: 10.1016/s1364-6613(00)01845-3.

**Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

**Raichle ME**. The restless brain: how intrinsic activity organizes brain function. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015; 370:20140172. doi: 10.1098/rstb.2014.0172.

**Schroeder CE**, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences. 2009; 32:9–18. doi: 10.1016/j.tins.2008.09.012.

**Siegel LL**, Kahana MJ. A retrieved context account of spacing and repetition effects in free recall. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2014; 40:755–764. doi: 10.1037/a0035585.

**Staudigl T**, Hanslmayr S. Theta Oscillations at Encoding Mediate the Context-Dependent Nature of Human Episodic Memory. Current Biology. 2013; 23(12):1101–1106. doi: 10.1016/j.cub.2013.04.074.

**Summerfield C**, Mangels JA. Dissociable Neural Mechanisms for Encoding Predictable and Unpredictable Events. Journal of Cognitive Neuroscience. 2006; 18:1120–1132. doi: 10.1162/jocn.2006.18.7.1120.

**Sweeney-Reed CM**, Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Richardson-Klavehn A, Hinrichs H, Heinze HJ, Knight RT, Rugg MD. Pre-stimulus thalamic theta power predicts human memory formation. NeuroImage. 2016; 138:100–108. doi: 10.1016/j.neuroimage.2016.05.042.

**Verplanck WS**, Collier GH, Cotton JW. Nonindependence of successive responses in measurements of the visual threshold. Journal of Experimental Psychology. 1952; 44:273–282. doi: 10.1037/h0054948.

**Weidemann CT**, Kahana MJ. Assessing recognition memory using confidence ratings and response times. Royal Society Open Science. 2016; 3:150670. doi: 10.1098/rsos.150670.

**Weidemann CT**, Kahana MJ. Dynamics of brain activity reveal a unitary recognition signal. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2019; 45:440–451. doi: 10.1037/xlm0000593.

**Weidemann CT**, Kragel JE, Lega BC, Worrell GA, Sperling MR, Sharan AD, Jobst BC, Khadjevand F, Davis KA, Wanda PA, Kadel A, Rizzuto DS, Kahana MJ. Neural activity reveals interactions between episodic and semantic memory systems during retrieval. Journal of Experimental Psychology: General. 2019; 148:1–12. doi: 10.1037/xge0000480.

**Weidemann CT**, Mollison MV, Kahana MJ. Electrophysiological correlates of high-level perception during spatial navigation. Psychonomic Bulletin & Review. 2009; 16:313–319. doi: 10.3758/pbr.16.2.313.