

Neural measures of subsequent memory reflect endogenous variability in cognitive function

Christoph T. Weidemann^{1*} and Michael J. Kahana²

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

¹Swansea University; ²University of Pennsylvania

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),

42 we computed list-level SMEs. Specifically, we analyzed EEG recordings from 97 individuals who each
43 studied and recalled 24 word lists in each of at least 20 experimental sessions that took place over
44 the course of several weeks. We trained ridge regression models to predict the (logit-transformed)
45 proportion of recalled items for each list, $p(\text{rec})$, on the basis of spectral EEG features that we
46 averaged over recordings during all encoding periods in that list. Additionally, we leveraged a prior
47 statistical model of memory performance which identified several critical variables predicting recall
48 performance across both lists and sessions *Kahana et al. (2018)*. By removing linear effects of these
49 variables, we uncovered the components of neural activity that predict the residual recallability of
50 studied items. Comparing SMEs for these residuals with those obtained for raw recall performance
51 thus allowed us to estimate the relative contributions of endogenous neural variability and external
52 factors to the SME. Throughout this paper we assessed our ability to predict recall performance
53 with a leave-one-session-out cross-validation procedure (see methods for details).

54 Results

55 **Figure 1** shows the mean proportion of recall as the function of several external variables that af-
56 fect recall performance for entire sessions (inter-session predictors) and for individual lists (within
57 each session (interlist predictors)). Specifically, we considered sleep duration in the night prior to
58 the free recall test, time of day, and self-rated alertness at the beginning of the experimental
59 session as inter-session predictors and experimental block within each session, the list number
60 within each block, and the average “recallability” of items within each list as interlist predictors (*Ka-
61 hana et al., 2018*). We are showing the effects of these variables across all participants (discretized
62 into two bins for each of the inter-session predictors and into ten bins for recallability) for illustrative
63 purposes, but we applied all of our analyses separately to the full data from each individual.
64 Additionally, we also considered the effect of session number (which was heterogeneous across
65 participants with some showing increased performance with increasing practice and some showing
66 a decline in performance) as an additional predictor in our inter-session and interlist regres-
67 sion models (described below). Detailed analyses of the effects of these variables on recall perfor-
68 mance in a large subset of this data set were the focus of a previous study (*Kahana et al., 2018*).

69 Given the strong effects of item-level characteristics on recall performance it is possible that
70 they explain a large proportion of the variance in item-level SMEs. Additionally, it is possible that
71 any endogenous variability driving SMEs is relatively fast, varying on the order of seconds (i.e.,
72 the time devoted to the study of individual items in typical memory experiments) rather than
73 minutes (i.e., the time encompassing a full study list) or longer. It was therefore not clear that
74 brain activity averaged over the individual study periods would be similarly informative about
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91

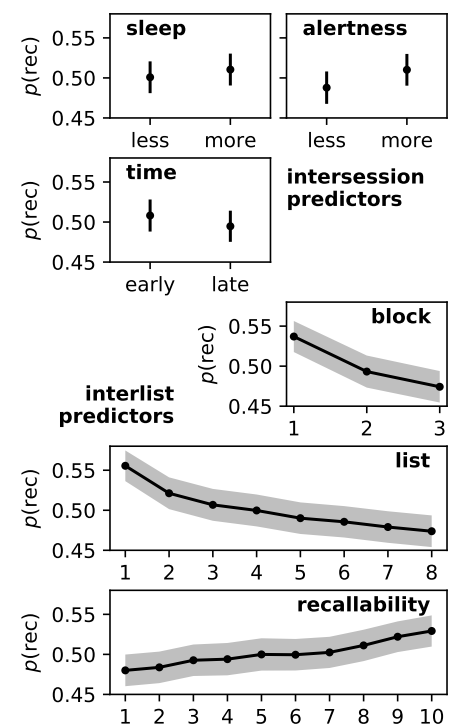


Figure 1. Mean probability of recall (and associated 95% confidence intervals) as a function of inter-session (amount of sleep, rated alertness, and time of day) and interlist (block number within a session, list number within a block, and mean recallability of items within a list) predictors. For the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

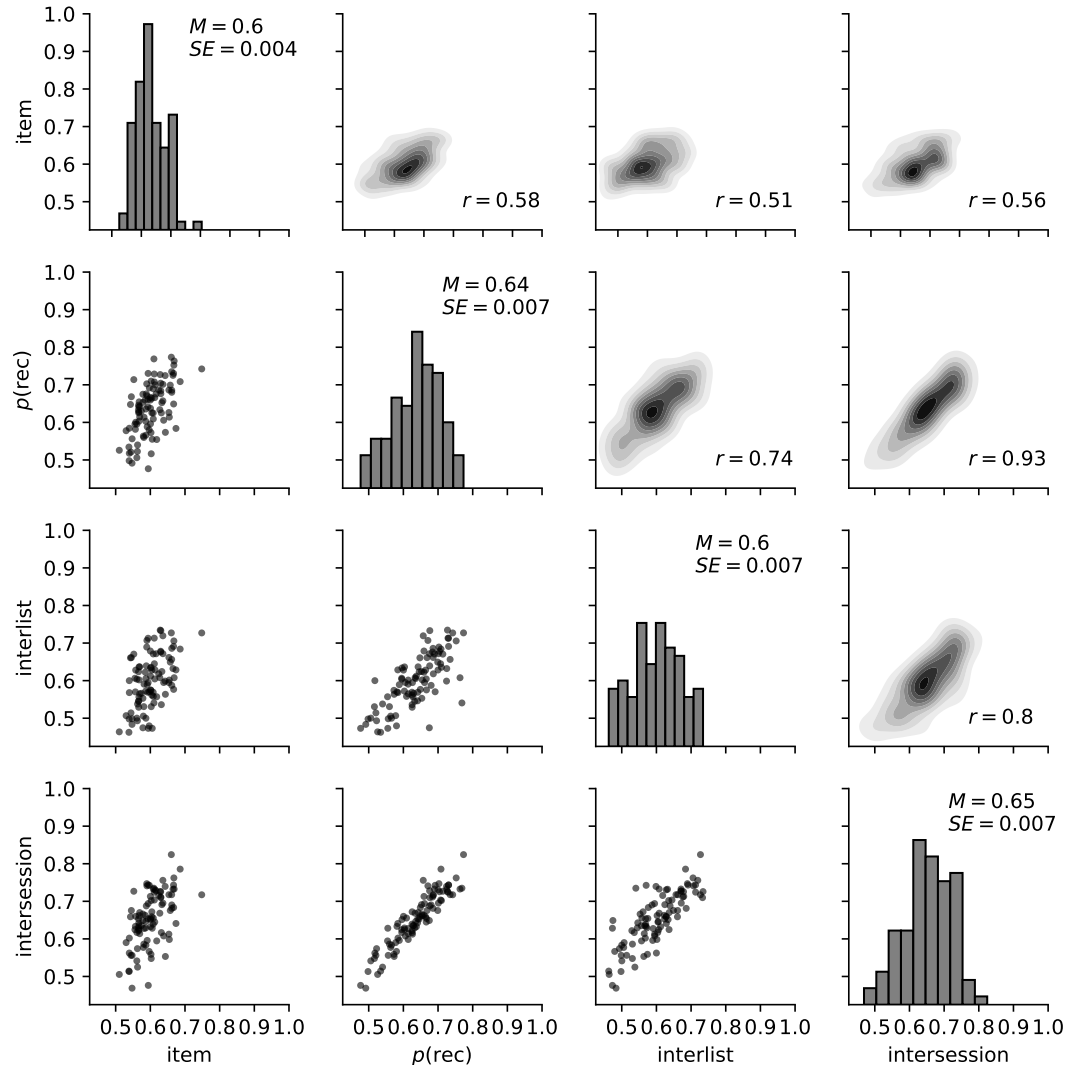


Figure 2. Areas under the ROC functions (AUCs) for classifier performance predicting subsequent memory for individual items (using a logistic regression classifier; item), lists of items (predicting the probability of recall across list items; $p(\text{rec})$), as well as for residuals of list-level 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.

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

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

112 Having ruled out item-level characteristics and fast endogenous variation as significant contribu-
113 tors to the SME, we next consider the extent to which external variables affecting recall performance
114 for entire sessions (intersession predictors: sleep, alertness, and time of day) and those that affect
115 recall performance at the list-level (interlist predictors: block, list, recallability) (*Kahana et al., 2018*)
116 are driving differences in brain activity that predict recall success. To the extent that either set
117 of variables can explain the SME, we can conclude that it also does not reflect slow endogenous
118 variability at the level of minutes (i.e., lists) or days (i.e., sessions). We constructed interlist and inter-
119 session regression models (both models also included session number as a predictor) to remove
120 linear effects of the respective external variables on $p(\text{rec})$. We then predicted the resulting list-level
121 residuals with ridge regression models using the same spectral EEG features as for our list-level
122 regression model predicting $p(\text{rec})$. **Figure 2** shows that AUCs for the interlist and intersession
123 regression models respectively matched ($M = 0.6$) or exceeded ($M = 0.65$, $t(96) = 8.499$, $SE = 0.006$,
124 $p < 0.001$) those for the item level classifier, demonstrating that spectral features effectively predict
125 list-level performance even after accounting for linear effects from several external variables that
126 affect recall. These results thus rule out these factors as major contributors to the SME, suggesting
127 that SMEs predominantly reflect slow endogenous variability in cognitive function.

128 Whereas the AUCs for the interlist regression models were significantly lower than those of
129 the regression models predicting $p(\text{rec})$ ($t(96) = 7.379$, $SE = 0.005$, $p < 0.001$), the AUCs for the
130 intersession regression models exceeded those for the other list-level regression models ($t(96) =$
131 5.812 and 11.682 , $SE = 0.003$ and 0.005 , $ps < 0.001$, for comparisons with the $p(\text{rec})$ and interlist
132 models, respectively; **Figure 2**). This pattern of results indicates some effects of interlist factors
133 on our measures of brain activity predicting recall performance, leading to a reduction in model
134 performance when linear effects of interlist predictors were removed. The fact that the intersession
135 models were better able to generalize across sessions indicates that relevant brain activity varying
136 across sessions was not effectively captured by our models (because we used a leave-one-session-
137 out cross-validation procedure to measure model performance, AUCs index the ability of our
138 models to generalize across sessions). Thus, removing linear effects of intersession predictors
139 removed variability that the models could not account for, leading to increased performance. These
140 results establish a small role for list-level effects due to external factors (e.g., proactive interference)
141 in the SME in addition to strong effects of endogenous variability in encoding processes.

142 **Figure 2** also highlights substantial correlations between AUCs for the different models. This

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

149 A standard measure of performance for regression models is the correlation between predicted
150 and actual values of the dependent measures. These correlations mirror the pattern of the
151 AUCs shown in *Figure 2* with $r = 0.26, 0.29,$ and 0.2 for $p(\text{rec})$, intersession residuals, and interlist
152 residuals respectively (all pairwise differences were statistically significant, $t(96) = 8.463\text{--}11.533,$
153 $SE = 0.003\text{--}0.008,$ $ps < 0.001$). The point-biserial correlation between predictions from the item-level
154 classifier and recall status of individual items was 0.16 . This confirms the above AUC-based analyses
155 indicating the effectiveness of spectral features in predicting list-level performance and the ability of
156 our models to capture some brain activity associated with interlist, but not intersession, predictors
157 (because of the better performance for the intersession models and the reduced performance of
158 the interlist models relative to the models predicting $p(\text{rec})$ as explained above). Likewise, as in the
159 above analyses, none of the list-level SMEs fell short of the item-level SME suggesting that brain
160 activity that is predictive of recall success is mainly driven by slow endogenous variability.

161 In addition to investigating the correlations between predictions from the different regression
162 models and the corresponding dependent measures, we can also assess the extent to which the
163 different models generalize to predicting the other measures.¹ This analysis reveals an advantage
164 for models trained on intersession residuals, even when these were tested on $p(\text{rec})$ or interlist
165 residuals. To assess the size of these differences, we removed the linear effects of the measure
166 each model was trained on from the generalization measures and computed the (semi-partial) cor-
167 relations between the model predictions and the resulting residuals. The semi-partial correlations
168 between predictions of models trained on intersession-residuals and the other two measures were
169 positive ($M = 0.1$ for both $p(\text{rec})$ and interlist-residuals; $t(96) = 17.324$ and $13.731,$ $SE = 0.006$ and
170 $0.008,$ $ps < 0.001,$ respectively). This confirms that the performance advantage for models trained
171 on intersession residuals generalizes to the prediction of $p(\text{rec})$ and interlist residuals—a result that
172 complements the above finding suggesting that removing linear effects of intersession predictors
173 eliminates variability in recall performance that is not effectively captured by our measures of brain
174 activity. The only other semi-partial correlations significantly deviating from 0 were those between
175 predictions of the models trained on $p(\text{rec})$ and the interlist-residuals ($M = 0.07,$ $t(96) = 10.158,$
176 $SE = 0.007,$ $p < 0.001$), reflecting the fact that models trained on $p(\text{rec})$ were better able to capitalize
177 on brain activity that is relevant for predicting recall performance than models that could not make
178 use of brain activity that reflects interlist predictors.

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

¹This 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.

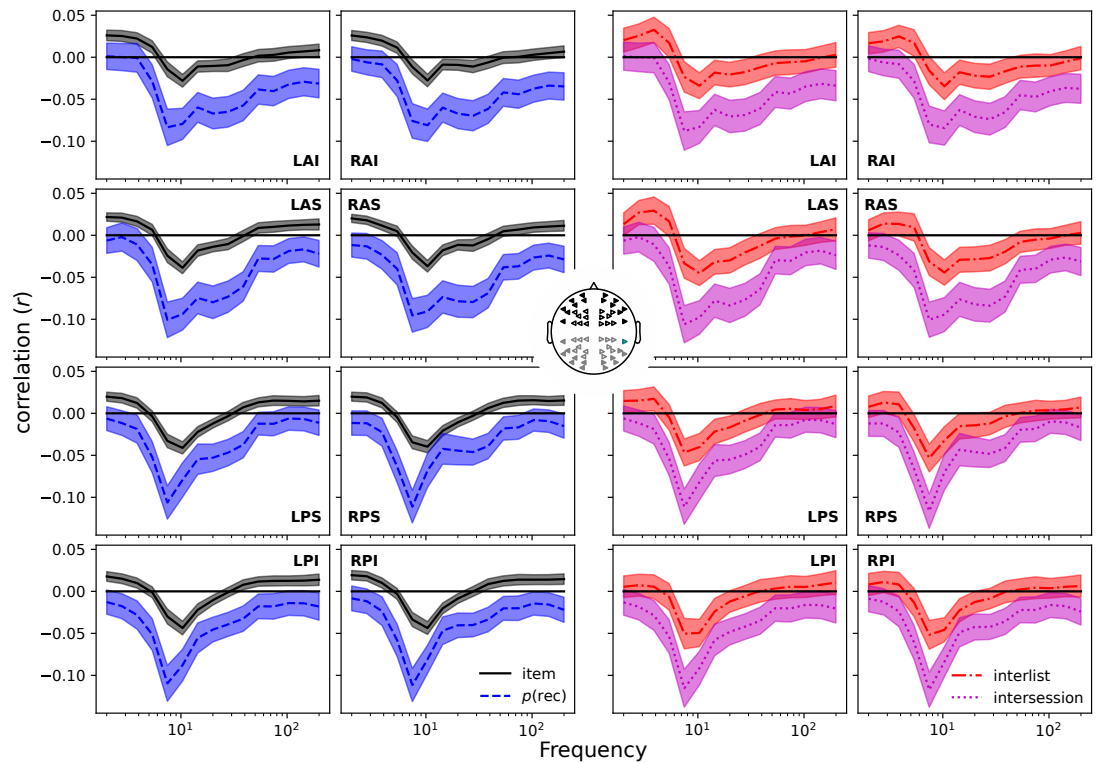


Figure 3. Correlations between mean power in each frequency across electrodes within each region of interest (ROI) and measures of recall performance (recall of individual items, $p(\text{rec})$, and residuals from the interlist and intersession models). The inset in the middle of the figure illustrates 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).

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

192 The similarity in how brain activity correlates
193 with different measures of recall performance
194 complements our analysis of correlations between
195 AUCs associated with different regression models
196 (**Figure 2**). Just like that analysis, however,
197 this similarity is difficult to interpret in light
198 of substantial correlations between the dependent
199 measures. To directly assess how brain activity
200 covaries with variability that is specific to
201 intersession and interlist predictors (removing
202 linear effects of $p(\text{rec})$), we therefore correlated
203 brain activity with corresponding residuals
204 (intersession $|p(\text{rec})$ and interlist $|p(\text{rec})$, respectively;
205 **Figure 4**). As is evident from **Figure 4**, correlations
206 of brain activity with intersession $|p(\text{rec})$ residuals
207 were close to zero and varied little across
208 frequencies or ROIs, consistent with the above
209 analyses indicating that our measures of brain
210 activity did not capture much of the variability
211 in recall performance associated with intersession
212 predictors. The correlations of brain activity
213 with interlist $|p(\text{rec})$, however, were relatively
214 strong, complementing the above analyses
215 indicating that our measures of brain activity
216 are sensitive to interlist predictors of recall
217 performance.

218 Discussion

219 Whether and how a studied item is encoded and
220 subsequently retrieved during a free recall task
221 is, by design, not subject to complete experimental
222 control. Indeed, recalled and not-recalled items
223 tend to differ on a number of dimensions. Prior
224 work has shown that neural activity just before
225 the presentation of individual items predicts
226 subsequent memory performance, demonstrating
227 SMEs that are independent of specific item
228 characteristics (*Sweeney-Reed et al., 2016; Otten et al., 2006; Fellner et al., 2013; Guderian et al., 2009*).
229 Nevertheless, task-related variables also strongly
230 predict memory performance and could be driving
231 SMEs even when they are not linked to specific
232 item characteristics (e.g., recalled items tend to
233 disproportionately come from early list positions, a
234 “primacy” effect) (*Murdock, 1962*). Thus, any
235 comparison of brain activity during the study of
236 items as a function of their subsequent recall is
237 fraught with confounds, complicating the
238 interpretation of the diagnostic neural signals. We
239 avoided some of these confounds by assessing
240 list-level SMEs, aggregating brain activity across
241 the study periods of all items within a list to
242 predict list-level recall. Our demonstration that
243 list-level SMEs were stronger than item-level
244 SMEs (**Figure 2**) with similar predictive
245 patterns of brain activity (**Figure 3**), shows
246 that item-level SMEs are not mainly driven by
247 external variables differentiating items within a
248 study list. This result also suggests the presence
249 of endogenous neural variation at slow time
250 scales (items in a study list were presented
251 over the course of about a minute) that predicts
252 subsequent memory.

253 Even when aggregating across items within a list,
254 a range of confounding variables remain. By
255 studying 97 individuals who each participated in
256 up to 23 experimental sessions, we were able
257 to model the effects of several external variables
258 that affect list-level recall performance. This
259 enabled us to not only relate brain activity to
260 the proportion of recalled items in each list, but
261 also

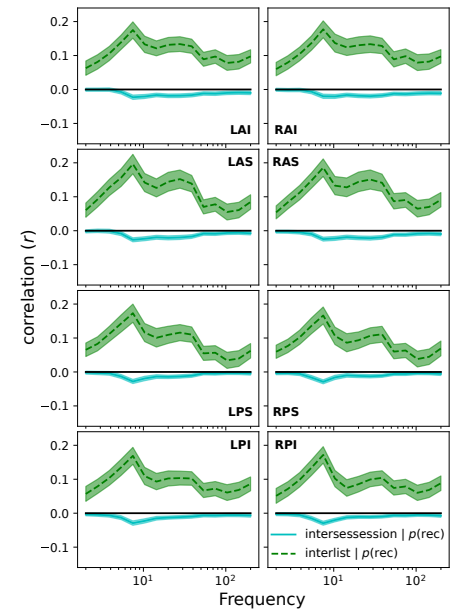


Figure 4. Correlations between mean power in each frequency across electrodes within each region of interest (ROI) and intersession and interlist residuals after regressing out linear effects of $p(\text{rec})$ (intersession $|p(\text{rec})$ and interlist $|p(\text{rec})$ respectively). Each panel shows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

241 to residuals of recall performance after accounting for effects of these external variables. Following
242 previous work (*Kahana et al., 2018*), we partitioned these external variables into those that varied
243 across lists (interlist) and those that varied across sessions (intersession). Accounting for interlist
244 variables reduced the list-level SME slightly (but not below the level of the item-level SME, Figures 2
245 and 3). This suggests that some, but not all, of the list-level SME reflects the effects of interlist
246 variables. Accounting for intersession variables, on the other hand, slightly increased the size of the
247 SME, demonstrating that the list-level SME does not include substantial contributions from these
248 variables (Figures 2 and 3; see also *Figure 4*).

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

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

269 The fact that substantial SMEs remained after accounting for a comprehensive set of external
270 variables may appear in conflict with findings that task context can affect the specific form of
271 SMEs, at least for recognition memory (*Kamp et al., 2017; Summerfield and Mangels, 2006; Otten
272 and Rugg, 2001; Staudigl and Hanslmayr, 2013; Fellner et al., 2013*). Task context manipulations
273 in these studies were designed to directly affect encoding processes (e.g., by asking participants
274 to perform different tasks on the study items) and their effects on SMEs suggest that they were
275 successful. Here we show that in the absence of direct manipulations of how study items are
276 presented or processed, external variables do not substantially contribute to the SME even when
277 they predict subsequent recall. These findings indicate that SMEs are not only effective measures
278 of memory formation, but that they reflect endogenous states that drive variability in cognitive
279 function.

280 Our findings align well with reports of sequential dependencies in human performance (*Kahana
281 et al., 2018; Gilden et al., 1995; Mueller and Weidemann, 2008; Verplanck et al., 1952*) as well as
282 with those of slow endogenous neural fluctuations that drive variability in evoked brain activity and
283 overt behavior (*Monte et al., 2008; Schroeder and Lakatos, 2009; Arieli et al., 1996; Fox et al., 2005,
284 2007; Fox and Raichle, 2007; Raichle, 2015*). Previous investigations of endogenous variability in
285 neural activity and performance have relied on exact repetitions of stimuli across many experimental
286 trials to limit variability in external factors. In order to study the effects of endogenous variability
287 on recall performance, we took a complementary approach by statistically removing the effects of
288 a comprehensive set of external factors. Despite the differences in methodologies and tasks, the
289 conclusions are remarkably consistent in establishing an important role for slowly varying neural
290 fluctuations in human cognition.

291 **Methods and Materials**

292 **Participants**

293 We analyzed data from 97 young adults (18–35) who completed at least 20 sessions in Experiment 4
294 of the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) in exchange for monetary
295 compensation. Recall performance for a large subset of the current data set was previously reported
296 (*Kahana et al., 2018*), but this is the first report of electrophysiological data from this experiment.
297 Data from PEERS experiments are freely available at <http://memory.psych.upenn.edu> and have
298 been reported in several previous publications (*Healey et al., 2014; Healey and Kahana, 2014, 2018;*
299 *Lohnas and Kahana, 2013; Siegel and Kahana, 2014; Lohnas et al., 2015; Weidemann and Kahana,*
300 *2016, 2019*). Our analyses included data from all participants with at least 20 sessions.

301 **Experimental task**

302 Each of up to 23 experimental sessions consisted of 24 study lists that each were followed by a
303 delayed free recall test. Specifically, each study list presented 24 session-unique English words
304 sequentially for 1,600 ms each with a blank inter-stimulus interval that was randomly jittered
305 (following a uniform distribution) between 800 and 1,200 ms. After the last word in each list,
306 participants were asked to solve a series of arithmetic problems of the form $A + B + C = ?$ where,
307 A , B , and C were integers in $[1, 9]$. Participants responded to each problem by typing the result
308 and were rewarded with a monetary bonus for each correctly solved equation. These arithmetic
309 problems were displayed until 24 s had elapsed and were then followed by a blank screen randomly
310 jittered (following a uniform distribution) to last between 1,200 and 1,400 ms. Following this delay,
311 a row of asterisks and a tone signaled the beginning of a 75 s free recall period. A random half of
312 the study lists (except for the first list in each session) were also preceded by the same arithmetic
313 distractor task which was separated from the first study-item presentation by a random delay
314 jittered (following a uniform distribution) to last between 800 and 1,200 ms. Each session was
315 partitioned into 3 blocks of 8 lists each and blocks were separated by short (approximately 5 min)
316 breaks. At each session participants were asked to rate their alertness and indicate the number of
317 hours they had slept in the previous night.

318 **Stimuli**

319 Across all lists in each session the same 576 common English words (24 words in each of 24 lists)
320 were presented for study, but their arrangement into lists differed from session to session (subject
321 to constraints on semantic similarity (*Healey et al., 2014*)). These 576 words were selected from a
322 larger word pool (comprising 1,638 words) used in other PEERS experiments. The 576-word subset
323 of this pool used in the current experiment were selected to maximize homogeneity, by removing
324 words that were atypical in frequency, concreteness, or emotional valence. Many participants
325 also returned for a 24th session that used words from the entire 1,638-word pool, but we are not
326 reporting data from that session here. We estimated the mean recallability of items in a list from
327 the proportion of times each word within the list was recalled by other participants in this study.

328 **EEG data collection and processing**

329 Electroencephalogram (EEG) data were recorded with either a 129 channel Geodesic Sensor net
330 using the Netstation acquisition environment (Electrical Geodesics, Inc.; EGI) or with a 128 channel
331 Biosemi Active Two system. EEG recordings were re-referenced offline to the average reference.
332 Because our regression models weighted neural features with respect to their ability to predict
333 (residuals of) recall performance in held out sessions, we did not try to separately eliminate artifacts
334 in our EEG data. Data from each participant were recorded with the same EEG system throughout all
335 sessions and for those sessions recorded with the Geodesic Sensor net, we excluded 26 electrodes
336 that were placed on the face and neck, rather than the scalp, from further analyses. The EGI system
337 recorded data with a 0.1 Hz high-pass filter and we applied a corresponding high-pass filter to the

338 data collected with the Biosemi system. We used MNE (*Gramfort et al., 2013, 2014*), the Python
339 Time-Series Analysis (PTSA) library (https://github.com/pennmem/ptsa_new), Sklearn (*Pedregosa*
340 *et al., 2011*) and custom code for all analyses.

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

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

355 **Item-based classifier**

356 For the item-based classifier we used a nested cross-validation procedure to simultaneously deter-
357 mine the regularization parameter and performance of L2-regularized logistic regression models
358 predicting each item's subsequent recall. At the top level of the nested cross-validation procedure
359 we held out each session once—these held out sessions were used to assess the performance
360 of the models. Within the remaining sessions, we again held out each session once—these held-
361 out sessions from within each top-level cross-validation fold were used to determine the optimal
362 regularization parameter, C , for Sklearn's LogisticRegression class. We fit models with 9 different
363 C values between 0.00002 and 1 to the remaining sessions within each cross-validation fold and
364 evaluated their performance as a function of C on the basis of the held out sessions within this
365 fold. We then fit another logistic regression model using the best-performing C value to all sessions
366 within each cross-validation fold and determined the model predictions on the sessions that were
367 held-out at the top level. We calculated the area under the ROC function on the basis of the
368 predictions from these held-out sessions.

369 **List-based regression models**

370 For the list-based regression models we followed the same procedure as for the item-based classifier
371 to determine the optimal level of regularization for ridge regression models predicting (residuals
372 of) list-level recall performance. Specifically, we used the same nested cross-validation procedure
373 described above to determine optimal values for α (corresponding to $1/C$), the regularization
374 parameter in Sklearn's Ridge class, testing 9 values between 1 and 65536. We applied these models
375 to the (logit-transformed) proportion of items recalled for each list, $p(\text{rec})$, as well as to the residuals
376 from the interlist and intersession models as described in the results section (*Kahana et al., 2018*).

377 **Data availability**

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

379 **Acknowledgements**

380 This work was supported by Grant MH55687 to MJK. We thank Ada Aka, Effie Li, Nicole Kratz, Adam
381 Broitman, Isaac Pedesich, Karl Healey, Patrick Crutchley and Elizabeth Crutchley and other members
382 of the Computational Memory Laboratory at the University of Pennsylvania for their assistance with
383 data collection and preprocessing and Nora Herweg and Ethan Solomon for helpful comments on a
384 draft of this manuscript.

References

- 385
386 **Arieli A**, Sterkin A, Grinvald A, Aertsen A. Dynamics of Ongoing Activity: Explanation of the Large Variability in
387 Evoked Cortical Responses. *Science*. 1996; 273:1868–1871. doi: [10.1126/science.273.5283.1868](https://doi.org/10.1126/science.273.5283.1868).
- 388 **Burke JF**, Long NM, Zaghoul KA, Sharan AD, Sperling MR, Kahana MJ. Human intracranial high-frequency
389 activity maps episodic memory formation in space and time. *NeuroImage*. 2014; 85 Pt. 2:834–843. doi:
390 [10.1016/j.neuroimage.2013.06.067](https://doi.org/10.1016/j.neuroimage.2013.06.067).
- 391 **DeLosh EL**, McDaniel MA. The role of order information in free recall: Application to the word-frequency
392 effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1996; 22:1136–1146. doi:
393 [10.1037/0278-7393.22.5.1136](https://doi.org/10.1037/0278-7393.22.5.1136).
- 394 **Fawcett T**. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27:861–874. doi:
395 [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010).
- 396 **Fellner MC**, Bäuml KHT, Hanslmayr S. Brain oscillatory subsequent memory effects differ in power and long-
397 range synchronization between semantic and survival processing. *NeuroImage*. 2013; 79:361–370. doi:
398 [10.1016/j.neuroimage.2013.04.121](https://doi.org/10.1016/j.neuroimage.2013.04.121).
- 399 **Fox MD**, Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance
400 imaging. *Nature Reviews Neuroscience*. 2007; 8:700–711. doi: [10.1038/nrn2201](https://doi.org/10.1038/nrn2201).
- 401 **Fox MD**, Snyder AZ, Vincent JL, Raichle ME. Intrinsic Fluctuations within Cortical Systems Account for Intertrial
402 Variability in Human Behavior. *Neuron*. 2007; 56:171–184. doi: [10.1016/j.neuron.2007.08.023](https://doi.org/10.1016/j.neuron.2007.08.023).
- 403 **Fox MD**, Snyder AZ, Zacks JM, Raichle ME. Coherent spontaneous activity accounts for trial-to-trial variability in
404 human evoked brain responses. *Nature Neuroscience*. 2005; 9:23–25. doi: [10.1038/nn1616](https://doi.org/10.1038/nn1616).
- 405 **Gilden D**, Thornton T, Mallon M. 1/f noise in human cognition. *Science*. 1995; 267:1837–1839. doi: [10.1126/sci-
406 ence.7892611](https://doi.org/10.1126/science.7892611).
- 407 **Gramfort A**, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L,
408 Hämäläinen M. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013; 7:267. doi:
409 [10.3389/fnins.2013.00267](https://doi.org/10.3389/fnins.2013.00267).
- 410 **Gramfort A**, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläi-
411 nen MS. MNE software for processing MEG and EEG data. *NeuroImage*. 2014; 86:446–460. doi:
412 [10.1016/j.neuroimage.2013.10.027](https://doi.org/10.1016/j.neuroimage.2013.10.027).
- 413 **Guderian S**, Schott BH, Richardson-Klavehn A, Duzel E. Medial temporal theta state before an event predicts
414 episodic encoding success in humans. *Proceedings of the National Academy of Sciences*. 2009; 106(13):5365–
415 5370. doi: [10.1073/pnas.0900289106](https://doi.org/10.1073/pnas.0900289106).
- 416 **Hanslmayr S**, Staudigl T. How brain oscillations form memories — A processing based perspective on oscillatory
417 subsequent memory effects. *NeuroImage*. 2014; 85:648–655. doi: [10.1016/j.neuroimage.2013.05.121](https://doi.org/10.1016/j.neuroimage.2013.05.121).
- 418 **Hanslmayr S**, Staudigl T, Fellner MC. Oscillatory power decreases and long-term memory: the information via
419 desynchronization hypothesis. *Frontiers in Human Neuroscience*. 2012; 6. doi: [10.3389/fnhum.2012.00074](https://doi.org/10.3389/fnhum.2012.00074).
- 420 **Healey MK**, Kahana MJ. Is memory Search Governed by Universal Principles or Idiosyncratic Strategies? *Journal*
421 *of Experimental Psychology: General*. 2014; 143:575–596. doi: [10.1037/a0033715](https://doi.org/10.1037/a0033715).
- 422 **Healey MK**, Kahana MJ. Age-Related Changes in the Dynamics of Memory Encoding Processes Provide a
423 Biomarker of Successful Aging. *Manuscript Submitted for publication*. 2018; .
- 424 **Healey MK**, Crutchley P, Kahana MJ. Individual differences in memory search and their relation to intelligence.
425 *Journal of Experimental Psychology: General*. 2014; 143:1553–1569. doi: [10.1037/a0036306](https://doi.org/10.1037/a0036306).
- 426 **Kahana MJ**, Aggarwal EV, Phan TD. The variability puzzle in human memory. *Journal of Experimental Psychology:*
427 *Learning, Memory, and Cognition*. 2018; doi: [10.1037/xlm0000553](https://doi.org/10.1037/xlm0000553).
- 428 **Kamp SM**, Bader R, Mecklinger A. ERP Subsequent Memory Effects Differ between Inter-Item and Unitization
429 Encoding Tasks. *Frontiers in Human Neuroscience*. 2017; 11. doi: [10.3389/fnhum.2017.00030](https://doi.org/10.3389/fnhum.2017.00030).
- 430 **Kim H**. Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies.
431 *NeuroImage*. 2011; 54(3):2446–2461. doi: [10.1016/j.neuroimage.2010.09.045](https://doi.org/10.1016/j.neuroimage.2010.09.045).

- 432 **Lohnas LJ**, Kahana MJ. Parametric Effects of Word Frequency in Memory for Mixed Frequency Lists. *Journal of*
433 *Experimental Psychology: Learning, Memory, and Cognition*. 2013; 39:1943–1946. doi: 10.1037/a0033669.
- 434 **Lohnas LJ**, Polyn SM, Kahana MJ. Expanding the scope of memory search: Modeling intralist and interlist effects
435 in free recall. *Psychological Review*. 2015; 122:337–363. doi: 10.1037/a0039036.
- 436 **Long NM**, Kahana MJ. Successful memory formation is driven by contextual encoding in the core memory
437 network. *NeuroImage*. 2015; 119:332–337. doi: 10.1016/j.neuroimage.2015.06.073.
- 438 **Merritt PS**, DeLosh EL, McDaniel MA. Effects of word frequency on individual-item and serial order retention:
439 Tests of the order-encoding view. *Memory & Cognition*. 2006; 34:1615–1627. doi: 10.3758/bf03195924.
- 440 **Monto S**, Palva S, Voipio J, Palva JM. Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and
441 Oscillation Amplitudes in Humans. *Journal of Neuroscience*. 2008; 28:8268–8272. doi: 10.1523/jneurosci.1910-
442 08.2008.
- 443 **Mueller ST**, Weidemann CT. Decision noise: An explanation for observed violations of signal detection theory.
444 *Psychonomic Bulletin & Review*. 2008; 15:465–494. doi: 10.3758/pbr.15.3.465.
- 445 **Murdock BB Jr.** The serial position effect of free recall. *Journal of Experimental Psychology*. 1962; 64:482–488.
446 doi: 10.1037/h0045106.
- 447 **Otten LJ**, Quayle AH, Akram S, Ditlewicz TA, Rugg MD. Brain activity before an event predicts later recollection.
448 *Nature Neuroscience*. 2006; 9(4):489–491. doi: 10.1038/nn1663.
- 449 **Otten LJ**, Rugg MD. Electrophysiological correlates of memory encoding are task-dependent. *Cognitive Brain*
450 *Research*. 2001; 12:11–18. doi: 10.1016/s0926-6410(01)00015-5.
- 451 **Paller KA**, Wagner AD. Observing the transformation of experience into memory. *Trends in Cognitive Sciences*.
452 2002; 6:93–102. doi: 10.1016/s1364-6613(00)01845-3.
- 453 **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg
454 V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in
455 Python. *Journal of Machine Learning Research*. 2011; 12:2825–2830.
- 456 **Raichle ME**. The restless brain: how intrinsic activity organizes brain function. *Philosophical Transactions of the*
457 *Royal Society B: Biological Sciences*. 2015; 370:20140172. doi: 10.1098/rstb.2014.0172.
- 458 **Schroeder CE**, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in*
459 *Neurosciences*. 2009; 32:9–18. doi: 10.1016/j.tins.2008.09.012.
- 460 **Siegel LL**, Kahana MJ. A retrieved context account of spacing and repetition effects in free recall. *Journal of*
461 *Experimental Psychology: Learning, Memory, and Cognition*. 2014; 40:755–764. doi: 10.1037/a0035585.
- 462 **Staudigl T**, Hanslmayr S. Theta Oscillations at Encoding Mediate the Context-Dependent Nature of Human
463 Episodic Memory. *Current Biology*. 2013; 23(12):1101–1106. doi: 10.1016/j.cub.2013.04.074.
- 464 **Summerfield C**, Mangels JA. Dissociable Neural Mechanisms for Encoding Predictable and Unpredictable
465 Events. *Journal of Cognitive Neuroscience*. 2006; 18:1120–1132. doi: 10.1162/jocn.2006.18.7.1120.
- 466 **Sweeney-Reed CM**, Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Richardson-Klavehn A, Hinrichs H,
467 Heinze HJ, Knight RT, Rugg MD. Pre-stimulus thalamic theta power predicts human memory formation.
468 *NeuroImage*. 2016; 138:100–108. doi: 10.1016/j.neuroimage.2016.05.042.
- 469 **Verplanck WS**, Collier GH, Cotton JW. Nonindependence of successive responses in measurements of the visual
470 threshold. *Journal of Experimental Psychology*. 1952; 44:273–282. doi: 10.1037/h0054948.
- 471 **Weidemann CT**, Kahana MJ. Assessing recognition memory using confidence ratings and response times. *Royal*
472 *Society Open Science*. 2016; 3:150670. doi: 10.1098/rsos.150670.
- 473 **Weidemann CT**, Kahana MJ. Dynamics of brain activity reveal a unitary recognition signal. *Journal of Experi-*
474 *mental Psychology: Learning, Memory, and Cognition*. 2019; 45:440–451. doi: 10.1037/xlm0000593.
- 475 **Weidemann CT**, Kragel JE, Lega BC, Worrell GA, Sperling MR, Sharan AD, Jobst BC, Khadjevand F, Davis KA,
476 Wanda PA, Kadel A, Rizzuto DS, Kahana MJ. Neural activity reveals interactions between episodic and semantic
477 memory systems during retrieval. *Journal of Experimental Psychology: General*. 2019; 148:1–12. doi:
478 10.1037/xge0000480.
- 479 **Weidemann CT**, Mollison MV, Kahana MJ. Electrophysiological correlates of high-level perception during spatial
480 navigation. *Psychonomic Bulletin & Review*. 2009; 16:313–319. doi: 10.3758/pbr.16.2.313.