The Temporal Structure of Memory Retrieval

John R. Anderson
Carnegie Mellon University

Jelmer P. Borst
University of Groningen

Jon M Fincham
Carnegie Mellon University

Avniel Ghuman
University of Pittsburgh

Caitlin Tenison
Carnegie Mellon University

Qiong Zhang
Carnegie Mellon University

1 This work was supported by the National Science Foundation grant DRL-1420008, a James S. McDonnell Scholar Award to JRA. The analyses and models in this paper can be obtained at http://actr.psy.cmu.edu/?post_type=publications&p=21105
Abstract

Neural imaging can be used to identify the stages of mental processes. We use a combination of hidden semi-Markov models and multivariate pattern matching (HSMM-MVPA) to analyze magnetoencephalography (MEG) data from memory experiments. The HSMM-MVPA methods are able to reveal brief “bumps” in the sensor data that mark the onset of different stages of processing. These bumps mark encoding, retrieval, decision, and response stages within single trials. The temporal distributions of these stages as well as the brain regions engaged indicate that the same stages occur in two different experiments (associative recognition of word pairs versus recall of arithmetic facts). The two experiments involved manipulations of the retrieval stage. Because the HSMM-MVPA technique allows the retrieval stage to be isolated in single trials, we can directly test theoretical claims concerning the distribution of retrieval times and activity in associated brain regions. The identified retrieval times have a Log-logistic distribution as would be predicted by the ACT-R (Anderson, 2007) theory of memory retrieval. Also as predicted, this stage is associated with an activity increase in the lateral inferior prefrontal cortex (LIPFC), which the theory claims is responsible for control of retrieval attempts.
The goal of this paper is to understand the cognitive processing that occurs during the brief period when someone retrieves a well-known fact. To take a current (2016) example, most people would have no problem answering the question “Who won the 2016 US election?” They are unlikely to introspect on how they answer this question nor would they come up with much insight if they did (much more likely to report on how they felt about the answer they retrieved). Answers to such questions seem to come effortlessly, but they do not come instantaneously. The goal of this paper is to identify the stages in generating such answers, what brain regions they involve, and how the stage durations vary on a trial-by-trial basis.

The first stage in answering such a question must involve encoding the terms in the question and the last stage must involve programming and executing the answer. While the identities of these two stages are fairly certain, there is still uncertainty as to their durations. There is much greater uncertainty about what happens in between. One class of models applied to recognition memory (e.g., Starns & Ratcliff, 2014; Wixted & Stretch, 2004; Shiffrin & Stevers, 1997) implies a continuous growth of evidence for an answer. When a threshold is met that answer is converted into a response. A second class of models, often applied to neuroimaging data, (e.g. Badre & Wagner, 2007; Thomson-Shill et al., 1999), proposes an initial access to some relevant knowledge in memory and then a decision process that determines an answer from what is retrieved. An instance of the second class of models is the ACT-R model of recognition memory and recall (e.g., Anderson & Schneider, 2012; Sohn et al., 2005). Appendix A provides a review of ACT-R’s mathematical theory of memory retrieval.
Classically, efforts to identify the number of mental stages in a task and their character have looked at the effects of various experimental manipulations on total processing time (e.g., Donders, 1868; Sternberg, 1969). Similarly, most fMRI studies trying to separate retrieval from post-retrieval (e.g. Barredo et al., 2016) have looked at total hemodynamic response on trials from different conditions. However, such inferences from total task measures are perilous. Inferences about temporal processing depend on assumptions about different stages such as whether they are independent (a controversial assumption, e.g., McClelland, 1979). Also, assumptions about latency distributions of stages not of interest, such as encoding and response, can bias inferences about stages of interest, such as memory and decision (Ratcliff, 2013; Verdonck & Tuerlinckx, 2016). Given the poor temporal resolution of the hemodynamic response, fMRI can tell about differences among conditions that are thought to affect different stages, but cannot actually identify which stage in a brief task produced the differences.

In recent work (Anderson et al., 2016; Borst & Anderson, 2015; Zhang et al., in press) we have tried to take advantage of the temporal resolution of EEG to identify the stages and when they were occurring in single trials. To cope with the trial-by-trial variability in the onset latencies of ERP components, we developed a novel method that involves applying hidden semi-Markov models and multivariate pattern analysis (HSMM-MVPA) to EEG data. Transitions from one cognitive stage to the next are signified by the onsets of bumps that sum with the task-irrelevant oscillatory activity in the EEG signal. The bumps have finite durations, amplitudes, and topographical distributions. The postulation that processing stages are signaled by such bumps is consistent with theories of ERP generation in EEG data (Makeig et al., 2002; Yeung,
Bogacz, Holroyd, & Cohen, 2004). The HSMM-MVPA attempts to recover the number, timing, and topographical distributions of these bumps.

In Anderson et al. (2016) we applied HSMM-MVPA to an EEG study of associative recognition where subjects learned pairs of words and then had to recognize word combinations they had studied and reject recombinations of the same words. Figure 1 shows the swimlane representation of the ACT-R retrieve-to-reject model (Anderson & Reder, 1999) for targets and foils in that experiment. Within the ACT-R model, a production (represented by red boxes in Figure 1a) evokes a change in processing, which we hypothesize produces a phasic response characterized by a bump in the EEG signal. The HSMM-MVPA method identifies both the distinctive scalp profiles (indicated in Figure 1a) and the variable latencies of such bumps in EEG trials. Two early bumps marked the encoding of the word pair, and a third bump marked the transition from encoding to retrieval. Following a variable period, a fourth bump marked the completion of retrieval and the start of comparison. A final, fifth bump marked completion of comparing the retrieved pair with the test pair, and was quickly followed by a behavioral response. Thus, the time to the third bump marked the Encoding stage, the time to the fourth bump the Retrieval stage, the time to the fifth bump the Decision stage, and the time to the trial end the Response stage.

In Figure 1b we show the predictions from the ACT-R model for the task. These predicted durations show a strong similarity to the average stage durations (Figure 1c) independently identified by applying the HSMM-MVPA technique to the ERP data. Critically, these analyses found evidence for separate Retrieval and Decision phases. This study found that the retrieval stage varied with interference (fan) and target versus
foil. Another study (Zhang et al., in press) found that the decision phase varies with the similarity of the foils to studied items. Thus, this method can identify which stage is being affected by which factors.

This paper extends the HSMM-MVPA methodology to two MEG (magnetoencephalography) experiments studying retrieval from long-term memory. The greater precision in localization that can be obtained with MEG enables testing of the activity during these stages in different brain regions. The two experiments are quite different, one involving recognition of word pairs and the other involving recall of newly learned arithmetic facts. Nonetheless, both involve comparable encoding and response structures, and both involve a manipulation of the retrieval stage and so serve as a strong test of whether that stage has a consistent character across tasks. In particular, going beyond past research, we will analyze whether the retrieval stage engages similar brain regions and has similar latency distributions in the two tasks.

**The MEG Experiments and their Data Representation**

The associative recognition experiment, similar to the experiment in Figure 1, was reported in Borst et al (2016), who used the imaging data to classify conditions. It consisted of two phases: a training phase in which subjects learned word pairs and a test phase in which subjects completed an associative recognition task. The test phase was scheduled the day after the training phase and took place in the MEG scanner. During the test phase subjects had to distinguish between targets (trained word pairs) and re-paired foils (re-arranged pairs). A critical manipulation was whether the words in the probe appeared uniquely in one pair or appeared in two pairs. This is referred to as a fan manipulation and for this reason we will refer to this experiment as the Fan experiment.
Subjects completed a total of 14 blocks (7 left-handed, 7 right-handed) with 64 trials per block. Detailed experimental procedure and materials can be found in the original paper (Borst et al., 2016). To correspond to the second experiment, we will only use the blocks with right hand responses. Hand did not have a significant effect on response time but did have a major effect on activity patterns. Figure 2 shows the mean times in this experiment and compares them with the earlier EEG results of Borst et al. (2013)

Tenison (2016) performed a MEG study of the learning of new arithmetic facts involving the pyramid relation. Pyramid problems (presented with a dollar symbol as the operator, e.g., 8$4=X) involve a base (8) that is the first term in an additive sequence and a height (4) that determines the number of terms to add. Each term in the sequence is one less than the previous (e.g., 8$4 = 8+7+6+5 = 26). We will refer to this experiment as the **Pyramid experiment**. Subjects solved a subset of these problems 54 times over the course of the experiment. Figure 3 shows how they sped up with repeated practice and compares this to other non-repeated problems they were solving at the same time. Initially, the height has a strong effect on response times and it continues to have a strong effect on the non-repeated problems. This is because it determines the number of additions. In contrast, subjects get much faster on repeated problems and the effect of height disappears (much like the results found with alpha-arithmetic, e.g., Logan, 1992; Barrouillet & Fayol, 1998; Zbrodoff, 1995). Previous studies (Tenison & Anderson, 2015; Tenison et al., 2016), using both behavioral and fMRI data, showed that subjects had stopped computing answers for repeated problems and were retrieving answers much like they would retrieve the answer to 8+3. We will be looking at repeated problems that take no longer than 3 seconds to solve. This is the same upper latency bound used for
the Fan experiment and the mean times were similar in the two studies (1.26 sec in Fan and 1.13 sec. in Pyramid).

In both experiments MEG data were recorded with a 306-channel Elekta Neuromag (Elekta Oy) whole-head scanner. The 306 channels are distributed into 102 sensor triplets, each containing one magnetometer and two orthogonal planar gradiometers. Data were digitized at 1 kHz. As part of standard MEG testing, four head position indicator (HPI) coils were attached to the subject’s scalp to track the position of the head in the MEG helmet. In addition, we obtained anatomical MRIs for purposes of source localization. In the Fan experiment subjects responded with a response glove and in the Pyramid experiment they responded with a numerical keyboard. The stimuli were projected onto a screen about 1 m in front of the subject. Stimulus onset was measured by a photodiode that was directly connected to the MEG recording system.

In each experiment our analyses focused on a binary factor that should affect the duration of the retrieval phase. In associative retrieval (18 subjects) the factor is fan, a contrast of probes whose words only occurred in one pair and probes whose words occurred in two pairs. In arithmetic retrieval (28 subjects), the factor is practice, contrasting items less than 30 or fewer practice opportunities (but with RTs faster than 3 seconds) with items that had more than 30 practice opportunities. In the ACT-R theory both Fan and practice are conceptualized as affecting retrieval time (see Appendix A). The analyzed data only involved correct trials. Subjects were quite accurate in both experiments (6% errors in Fan and 7% in Pyramid).

Although these are two rather different tasks, they were quite similar in terms of what the subject saw and the physical response. Figure 4 compares the sequence of
events for the two experiments. Each trial begins with a variable fixation period of 400 to 600 ms. Then the probe appears. While the probes are different (words versus numeric expression, in both cases two things need to be encoded (the two words or two numbers). The MEG data to be analyzed for each trial goes from the presentation to 50 ms. after the first key press. The key press in the Fan task is a binary choice providing the answer (target or foil) while in the Pyramid task it is a signal indicating the subject is ready to enter the answer. The sequence of events diverges after this key press. The Fan experiment simply ends with feedback as to correctness and an ITI (inter-trial interval). In the Pyramid task the subject must then enter the answer and is shown a more elaborate feedback explaining the answer, followed by an ITI.

Appendix B describes the details of the preprocessing that went into producing the data that will be used in our analyses. Only correct trials were used and only trials lasting between 400 and 3000 ms. (40 to 300 samples after down-sampling to 100 Hz). For purposes of data analysis, we also included 50-ms. after the response because this helps identify the response bump. The two resulting normed data sets involved 6736 trials (881,814 10-ms. samples) for Fan and 3540 trials (416,646 samples) for Pyramid.

As a step of dimensionality reduction and to deal with the fact that individual sensors do not necessarily correspond across subjects, we performed a multi-set canonical correlation analysis (MCCA) on our data (Zhang et al., submitted). This serves to compress the data from the 306 sensors into a reduced number (10 in our case) of uncorrelated dimensions that are common across subjects. The details of the MCCA analysis are explained in Appendix B. The outputs of these MCCA analyses are an 881,814 (samples) x 10 (components) matrix for Fan data (6736 trials, 18 subjects) and a
similar 416,646 x 10 matrix for the Pyramid data (3540 trials, 28 subjects). We also performed a third MCCA analysis where we put the data from both experiments together into a 1,298,460 x 10 matrix representing 10,276 trials and 46 subjects. The HSMM-MVPA analyses will estimate parameters that maximize the probability of the data as represented in these matrices.

**Bump Analysis of the MEG data**

We want to identify on a trial-by-trial basis the location of bumps that mark various stages in the generation of an answer to a memory probe. We assume that a bump is a 50-ms. half sin bump added to task-irrelevant oscillatory activity just as we had for our EEG data set (Figure 1). In between these 50 ms. periods are flats where we assume the activity pattern can be modeled by sinusoidal noise around a mean of 0. Bounded by trial onset and trial end, the placement of N bumps results in N+1 intervening flats.

Figure 5 provides a summary of assumptions, mathematics, and computations underlying the HSMM-MVPA analysis that estimates the bump parameters (see Anderson et al, 2016, for detailed derivations):

1. The estimation process finds a set of parameters that will maximize the likelihood of the data from all the trials, given some description of the structure of those trials.

2. The description of a trial is anchored by the placement of N bumps during the trials. The number of ways of placing N bumps in a trial of S samples is (S+N)!/(S! x N!) (a very large number, given trials averaging more than 100 samples). Every possible placement has a probability and the probability of the trial is the sum of all these probabilities. The essential power of the HSMM
algorithms is that they can efficiently compute these sums.

3. Any placement of the N bumps results in a partition of the trial into N bump regions and N+1 flat regions marked off by these bumps. The probability of such a partition is the product of the probabilities of the N bumps and the probabilities of the N+1 flats.

4. The probability of a flat varies with its duration. We assume that a simple latency-like distribution for these flats – gamma distributions with shape 2 and scale parameters $b_i$ to be estimated. In this paper we will explore how constraining this assumption is for the distributions of stage durations that are recovered.

5. The probability of a bump is determined by how well it matches the profile for that bump on each of the 10 MCCA dimensions. Because MCCA, like PCA, produces orthogonal dimensions the probability of a bump is the product of the probabilities on each dimension. Thus, each bump can be characterized by a set of 10 magnitudes $M_{ik}$ that represent the mean values on that dimension.

6. The 50 ms. (5 samples) that constitute the bump define a half-sin pattern that starts at 0 at the onset of the bump, reaches a maximum at 25 ms. and comes back down to 0 at 50 ms. Assuming the value of each sample represents the 5 ms. midpoint, the 5 sample weights are 0.309, 0.809, 1.000, 0.809, and 0.309. The probability of the bump is determined by how much variability in the sample can be explained by the assumption of such a bump

\[ M_{ik} \]

\[ \text{The value of the parameter } M_{ik} \text{ that will minimize } (S_{kl} - \text{halfsin}(M_{ik}))^2 \text{ at a fixed location is the same as what maximizes the more elaborate expression in Figure 5 at that location. However, the location that results in the maximal value of the larger expression can be determined.} \]
Thus, an N-bump model requires estimating N+1 scale parameters $b_k$ to describe the durations of the flats and 10N magnitude parameters $M_{ik}$ to describe the magnitudes of bumps $k$ on dimensions $i$. The scale parameters reflect the temporal pattern of the bump locations and the magnitude parameters reflect the multivariate pattern of the bumps. These parameters are estimated iteratively using the expectation maximization algorithm associated with an HSMM (Yu, 2010).

Results

Number of Cognitive Stages

We estimated models of 1 to 8 bumps for the two experiments. The estimation for each experiment was entirely independent of the estimation for the other experiment, including using separately obtained MCCA dimensions. The estimation process produces parameters that describe the topographies of the bumps and the durations of the flats. We then used these parameters estimated from one experiment to calculate the likelihood of the data from the other experiment. In essence we are using one experiment to predict the other. Figure 6 shows how the parameters from one experiment can predict the data from the other. In both directions the likelihood of data increases up to 5 bumps. The 5-bump model predicts the data better than fewer bumps for 16 out of the 18 Fan subjects and 20 of the 28 Pyramid subjects. Both directions are significant by a sign test. The fact that parameters estimated for one experiment can predict the data in the other is evidence that these bumps are tapping the same processes in the two experiments.

Using parameters estimated from the Pyramid experiment to predict the Fan data there is continued improvement up to 8 bumps (the 8-bump model fits at least 15 of the different than the location that minimizes the embedded expression.
18 Fan subjects better than any model with fewer bumps). However, there is not a similar improvement with more than 5 bumps when using fan-estimated parameters to predict the Pyramid data. Given that both directions show improved cross-experiment prediction for 5 bumps and given that 5 bumps correspond to the Anderson et al. (2016) EEG experiment, we will use a 5-bump model in the rest of the paper. We will use the results of a single 5-bump model fit to the results of a single MCCA performed on all 46 subjects (18 Fan and 28 Pyramid).

**Results of a Single 5-Bump Model for Both Experiments**

The HSMM-MVPA estimates on a trial-by-trial basis the time between bumps and hence the duration of stages. Figure 7 shows the mean estimated times for the two conditions of each experiment and compares them with the ACT-R predictions for these tasks. The ACT-R models predict that all stage durations except for the retrieval stage should be the same for both conditions of both experiments and the same as in the prior EEG experiment (Figure 1b). Thus, the correspondence of the predicted and estimated times for these non-retrieval stages was obtained without any parameter estimation. The differences among conditions and experiments are largely localized in the retrieval stage.

As explained in Appendix A, predictions of retrieval times required estimation of a single latency scale parameter, F, per experiment. This latency scale parameter scales a parameter-free prediction of the ratio of the retrieval times for the two conditions in that experiment. This ratio of times is determined by the activation levels of the retrieved memories in the two conditions. The activation level of a memory is a sum of activation spread from the terms in the probe plus base-level activation, which reflects things like practice. The Fan experiment manipulated activation spread and the Pyramid
experiment manipulated practice. The stage durations in Figure 7 are remarkably similar between these two experiments, the old EEG experiment, and the ACT-R model predictions. As predicted, in both experiments the major difference between the conditions is located in the retrieval stage.

We averaged sensor activity at all maximum likelihood bump positions for each subject for each trial. Figure 8 illustrates the t-values (the difference from zero divided by the standard deviation over the 46 subjects) for the 306 sensors at the 5 bumps. By chance 1 out of the 306 sensors should have an absolute t-value greater than 3.1. The numbers of sensors exceeding this threshold for the 5 bumps are 77, 15, 139, 28, and 74. Thus, the bump analysis is picking up reliable trends in the underlying sensor data. Moreover the patterns in Figure 8 are what we would expect from underlying sources because they display the opposing positive and negative bands across a local minimum that suggest underlying neural generators.

**Activity in Relevant Brain Regions**

As described in Appendix C, we used sample-by-sample sensor activity to infer patterns of activity in the 5124 sources. We then warped these data for each trial so that bumps would have the same position and stretched or shrunk activity between each bump. Since the average trial was 1.26 second, this resulted in 126 10-ms. samples per trial, which we averaged for each condition for each subject. Figure 9 shows the average activity of all sources over the course of the trials. For purposes of analysis we broke the time course of source activation into 3 periods: encoding (first three stages), retrieval (4th stage), and finishing (last 2 stages). We divided stages at the point of the peak of the bumps. We also treated Fan 1 and Late as one condition of faster retrieval and Fan 2 and
Early as one condition of slower retrieval. Past fMRI research (e.g., Anderson et al., 2008, Borst & Anderson, 2013) has associated ACT-R modules with regions of the brain and we will focus on three regions relevant to these experiments: the left fusiform as an index of early encoding activity, the left motor region as an index of response activity, and the left lateral inferior prefrontal cortex (LIPFC) as an index of retrieval. As described in the Appendix C, we obtained measures of how activity in these regions differed from global activity.

**(a) Fusiform (Figure 9a):** There is a highly significant difference between periods (F(2,88)=32.04, p < .0001) and an interaction between experiment and period (F(2,88)=9.09, p < .0005). Both experiments show significantly greater fusiform activation during the Encoding phase than later phases (Fan: t(17)=6.76, p < .0001; Pyramid: t(27)=2.53, p < .05) but the effect is larger in the Fan experiment. No other main effects or interactions are significant.

**(b) LIPFC (Figure 9b):** The only significant effect for the LIPFC is that of period (F(2,88)= 19.19, p< .0001). The two experiments seem to show remarkable similar changes over the course of the trial and the interaction with experiment does not approach significance (F(2,88)= 0.67). The activity shows a large rise from the Encoding Period to the Retrieval Period (t(45)=4.72, p < .0001) and maintains that level though the Finishing Period (t(45)= 1.45, p > .1).

**(c) Motor (Figure 9c):** Again the only strong effect is that of period (F(2,88)= 9.01, p< .0005) but there is a marginal interaction with experiment (F(2,88)= 2.72, p< .1). The Fan experiments shows significant drop in activation from Encoding to Retrieval (t(17)=3.57, p < .005) and significant rise during the Finishing Period (t(17)=4.05, p <
In the Pyramid experiments these effects are marginal (t(27) = 2.03, 1.85, both p < .10, 2-tailed). Both experiments seem to show a rise over the last 250 msec. and tests for a linear increase are significant for both experiments (fan: t(17)=2.71, pyramid: t(27)= 2.07, both p < .05).

The patterns of activity in Figure 9 provide evidence that these periods do involve the functions attributed to them. In particular it offers evidence for our identification of a Retrieval Period: Fusiform activity drops off with the onset of retrieval; LIPFC activity rises with the onset of retrieval, and motor activity rises with the offset of retrieval.

**Distributions of Durations**

As described in Appendix D, The HSMM-MVPA analysis provides an estimate of the maximum likelihood (ML) durations of each stage on each trial. We combined the early stages to get one estimate of Encoding times and the late stages to get one estimate of Finishing times. We used the time to end of the third bump as our Encoding Period, the time from the end of the third bump to the beginning of the fourth bump as our Retrieving Period, and the time from the beginning of the fourth bump to the response as our Finishing Period. While the stages embedded within the Encoding and Finishing periods may represent distinct processes, the brevity of some of the stages relative to the 10-ms. temporal grain size of the analysis limits any interpretation of their distributions. The retrieval stage, which is the focus of this paper, is sufficiently long to permit a detailed analysis.

Figure 10 shows distributions of the ML durations for these three periods separated into the two experiments and their two conditions:

(a) **Encoding.** Subjects take about 33 ms. less encoding the stimulus in the Pyramid
experiment. This could reflect the simpler nature of the Pyramid stimuli (“6$4” on one line versus “METAL SPARK” arranged on two lines). There is virtually no difference between the two conditions in either experiment – not only mean times but entire distributions. This is despite the fact that the two conditions have quite different total times.

**(b) Retrieving.** The large differences among conditions appear in the retrieving interval. This is the period of real interest that is buried between encoding and finishing. We explore below how to account for these distributions and their differences.

**(c) Finishing.** The ending times include both the decision stage and the response stage\(^3\). The late Pyramid responses are about 40 ms. faster than the rest, perhaps reflecting a motor speed up in operating the numeric keypad as the subject progresses through the experiment. The time to complete the response after the first key also speeds up in the experiment (from 1.15 sec to .96 seconds).

The Encoding and Finishing distributions contrast with the Retrieving distribution in that they do not begin to rise from zero until after some amount of time. The “irreducible intercepts” for Encoding reflects factors such as time for the signal to reach the brain and time to travel from early visual areas to regions responsible for the retrieval request. The irreducible Finishing time reflects similar physical processes such as time for the signal from the brain to result in a finger press.

We tried fitting three types of distributions to these densities:

1. **Gamma.** This was chosen because this is assumed in the HSMM-MVPA

\(^3\) But excludes the 50 ms. beyond the response that was used in fitting stages.
analysis. It is used in the HSMM-MVPA analysis only as a way to constrain the estimated stage distributions to have a sensible shape. Left unconstrained as to distribution shape the HSMM will drastically overfit the flat durations (it can estimate a separate probability for each time bin). The parameters of this distribution are $\alpha$ (shape) and $\beta$ (scale).

2. **ExGaussian.** This was chosen because it often gives a good fit to full latency distributions (e.g., Matze & Wagenmakers, 2009). It is conceived of as a sum of a normal and an exponential distribution. Its parameters are $\mu$ and $\sigma$ (mean and standard deviation of the normal), plus $\tau$ (mean duration of the exponential).\(^4\)

3. **Log-logistic.** This was chosen because the ACT-R theory assumption that the trial-to-trial variability in memory activation conforms to a logistic distribution. Since retrieval time is related to activation by an exponential function, retrieval times should have a Log-logistic distribution. The parameters of this distribution are $\mu$ (the mean of the activations) and $s$ which determines the trials to trial variation (the standard deviation of the activations is $\pi s/3^{1/2}$.)

These three distributions were fit to the ML durations of the periods for each condition for each subject. Table 1 summarizes the range of parameters estimated across subjects. Table 2 gives the mean BIC measures of fit that attempt to correct for the extra parameter of the ExGaussian. While Table 2 breaks the results out by condition, to determine which distribution was best we summed the BIC scores for the two conditions for each subject. The results for the three periods were different:

1. **Encoding Period.** The ExGaussian performs best, having the best BIC scores for

---

\(^4\) Similar quality of fits would be obtained with a shifted Wald distribution (Matze & Wagenmakers, 2009).
32 of the subjects. It superiority arises from the fact that it is best able to capture both the
delay to rise (with its normal) plus the long tail (with its exponential).

2. **Retrieval period.** The Log-logistic performs best, with top BIC scores for 33 of
the 46 subjects. While the ExGaussian usually fit a subject better in terms of log
likelihood, the advantage is not enough to compensate for the cost of the extra parameter
in the BIC calculation. The Gamma distribution, which has the same number of
parameters as the Log-logistic, has difficulty producing the combination of an early peak
and long tail.

3. **Finishing Period.** The Gamma performs best, with a top BIC score for 32 of the
46 subjects. While the ExGaussian again tends to have better log-likelihoods, it is not
enough to compensate for the extra parameter. The Log-logistic has difficulty
simultaneous fitting the delayed onset and long tail.

Many theories, including ACT-R, treat the Encoding and Finishing distributions
as normal or close to normal\(^5\), which they definitely are not. Normal distributions are
often just assumed for convenience with the theoretical interest on the distribution of an
embedded stage like retrieval. The ACT-R prediction of a Log-logistic distribution for
retrieval times is confirmed. A good fit for the Log-logistic to the retrieval distribution
is not trivial as indicated by the poor fit of the Log-logistic to the other intervals.

The conclusions from Table 2 depend on a number of assumptions about the
HSMM-MVPA estimation process and the appropriateness of the BIC statistic for
deciding among distributions. These issues are discussed in Appendix E where we take

\(^5\) In ACT-R the distribution of times for a process, except the retrieval is assumed to
be uniform around the mean. A convolution of a number of uniforms comes to
closely approximate a normal. Both Encoding and Finishing involve a number of
processes (see Figure 1).
an alternative approach to testing conclusions about the retrieval distributions of interest. There we generate synthetic data with different underlying retrieval distributions (Gamma, ExGaussian, Log-Logistic) and apply the HSMM-MVPA estimation process the synthetic data. We found that the retrieval distribution would not be fit by Log-logistic as well as it is if the true distribution were Gamma or ExGaussian, but only if it were indeed Log-Logistic.

Encouraged by the success of the Log-logistic in fitting the retrieval distribution, we went on to examine an ACT-R-specific prediction about the retrieval distribution. According the ACT-R theory, both conditions in an experiment should share a common variability in activation values. Therefore, a 3-parameter model with a shared noise parameter should fit the two conditions in an experiment better than a 4-parameter model with separate noise parameters. In fact, a 3-parameter model has a better BIC measure for 41 of the 46 subjects better. Figure 11 shows the fit of the 3-parameter models to the distributions of the two experiments.

The Encoding, Retrieval, and Finishing times sum to produce the total times that subjects took. This distribution of total time, shown in Figure 12, is unequivocally best fit by an ExGaussian distribution, resulting in lower BIC values for 45 of the 46 subjects than Log-logistic or Gamma. The fits to the three periods suggest that the total time is a combination of three different distributions – ExGaussian for encoding, Log-logistic for retrieval, and Gamma for finishing. For each subject and each condition we calculated the convolution of these estimated distributions and compared this to the actual distribution of times and to ExGaussian fit to the total time distribution. Figure 12 compares the average convolutions over all subjects and conditions with the frequency of
different total times (binned into 10-ms. samples)\(^6\). The convolutions and the ExGaussians were highly correlated (.961) and much more correlated than either were to the data (.577 for the convolution and .595 for the ExGaussian).

**Conclusions**

We have combined MEG studies of two different memory retrieval tasks – associative recognition and retrieval of arithmetic facts. The experiments involved two different manipulations of memory access – associative interference versus practice. Still the evidence is that the two experiments involve the same retrieval processes and that these processes match the expectation of the ACT-R theory. Parameters estimated from one experiment could be used to predict the data in the other experiment (Figure 6). The effect of the manipulation in both experiments was principally located in the retrieval period and varied as predicted by the ACT-R theory (Figure 7). In both experiments activity in the LIPFC rose with the onset of retrieval and the two experiments did not differ in the magnitude of the rise (Figure 9b). The retrieval distributions for the two conditions in each experiment were best fit by Log-logistic distributions that shared a noise parameter, as predicted by the ACT-R theory (Figure 11).

As described in Appendix A the ACT-R prediction of a log-logistic distribution was based on two fundamental assumptions – that the speed of retrieval of a memory controlled by the activation of the memory and that there was trial-to-trial variability in that activation, which we thought would be approximately normal. The logistic distribution was chosen to model that variability for computational reasons that are not particularly relevant two decades later. A normal distribution could have been used in

\(^6\) A good fit to the individual periods implies a good fit to total time only if the distributions for the individual periods are independent.
which case the predicted distributions would be Lognormal. In fact, the Lognormal fits the retrieval distributions basically as well as the Log-logistic (better BIC score for 20 and of the 46 subjects, and an average BIC score 1.1 worse). While the distributional results confirm the ACT-R predictions, they would be consistent with any model that assumes the rate of retrieval has an exponential relationship to a quantity that varies according to an approximately normal distribution.

The LIPFC has been observed to be active during controlled retrieval and its activation has been related to both the retrieval process itself and post-retrieval decision. There are two noteworthy aspects of the source estimates observed in the LIPFC for both experiments (Figure 9b). First, we observe a rise in its activity with the onset of the retrieval stage but, second, that activity continues beyond the retrieval stage. This would imply that the prefrontal region is engaged in both retrieval and post-retrieval processes. The ACT-R model only predicts greater LIPFC engagement during the retrieval stage. There are suggestions (e.g., Badre et al., 2005), based on the greater spatial precision of fMRI, that different regions of the LIPFC are engaged in these two processes.

Second, we found no significant difference in the LIPFC between conditions of easy retrieval and hard retrieval (F(1,44)= 0.03). This might seem to contradict the typical fMRI result of greater activity in harder retrieval condition (e.g. Danker et al., 2008). In an earlier analysis of fan data, Borst et al (2016) found significantly greater prefrontal differences in the fan-1 condition at the end of the trial (150 – 500 msec before the response). This is the opposite of the fMRI effect. The key to understanding the difference between the fMRI experiments and the MEG experiments is to note that the hemodynamic response used in fMRI integrates activity over the short periods as in these
experiments. Therefore, greater activity in fMRI in high fan conditions is expected as a consequence of the longer retrieval times.

In conclusion, these experiments and analyses have produced new detailed information about the retrieval process. Some aspects of what we found were predicted by the ACT-R theory but other aspects were not. This illustrates that this HSMM-MVPA method is theoretically neutral, judging theories according to the information in the data. It can extract information unseen by other methods because it combines information from the trials rather than averaging the trials, and because it considers all ways a trial might break out into stages. Thus, given enough trials and an imaging methodology with enough temporal and spatial resolution, it is possible to assess the fine-grained temporal details of the stages underlying performance of a task.
References


http://doi.org/10.1016/j.jneumeth.2003.10.009

Donders, F. C. (1868). De snelheid van psychische processen (On the speed of mental processes).


http://doi.org/10.1016/j.neuroimage.2012.01.021


http://doi.org/10.1016/j.neuroimage.2013.10.027


http://doi.org/10.1016/j.cogpsych.2008.02.004


Table 1
Mean parameter values and range across subjects

(a) Gamma

<table>
<thead>
<tr>
<th></th>
<th>Encoding Period</th>
<th>Retrieval Period</th>
<th>Finishing Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fan 1</td>
<td>Fan 2</td>
<td>Fan 1</td>
</tr>
<tr>
<td>Fan 2</td>
<td>38.86 (28-59.5)</td>
<td>38.03 (21.9-69.3)</td>
<td>3.27 (2.2-4.8)</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>39.98 (11.3-106.6)</td>
<td>3.52 (2.4-5)</td>
<td>11.21 (8.6-17.6)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>43.78 (11.2-147.2)</td>
<td>2.37 (1.3-4.4)</td>
<td>10.24 (5.9-15.4)</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fan 2</td>
<td>Pyr Early</td>
<td>Pyr Late</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>25.03 (20.7-32.4)</td>
<td>2.37 (1.3-4.4)</td>
<td>10.24 (5.9-15.4)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>21.65 (16.4-29.8)</td>
<td>2.44 (1.5-6.9)</td>
<td>11.12 (5.8-16.5)</td>
</tr>
</tbody>
</table>

(b) ExGaussian

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Encoding Period</td>
<td>Retrieval Period</td>
<td>Finishing Period</td>
</tr>
<tr>
<td></td>
<td>Fan 1</td>
<td>Fan 2</td>
<td>Fan 1</td>
</tr>
<tr>
<td>Fan 2</td>
<td>24.96 (20.4-31.5)</td>
<td>25.03 (20.7-32.4)</td>
<td>24.5 (21-29.2)</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>22.23 (16.9-30.2)</td>
<td>2.44 (1.5-6.9)</td>
<td>5.83 (-481.9-119.1)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>21.65 (16.4-29.8)</td>
<td>1.94 (0.4-5.4)</td>
<td>12.65 (-14.8-33.9)</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fan 2</td>
<td>Pyr Early</td>
<td>Pyr Late</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>25.03 (16.9-48.6)</td>
<td>5.83 (-481.9-119.1)</td>
<td>24.43 (17.9-31.4)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>21.65 (16.4-29.8)</td>
<td>6.59 (0.1-23.4)</td>
<td>21.52 (15.6-31)</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fan 2</td>
<td>Pyr Early</td>
<td>Pyr Late</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>33.36 (16.9-48.6)</td>
<td>10.57 (0.3-67.6)</td>
<td>24.43 (17.9-31.4)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>12.65 (-14.8-33.9)</td>
<td>6.59 (0.1-23.4)</td>
<td>21.52 (15.6-31)</td>
</tr>
<tr>
<td></td>
<td>Encoding Period</td>
<td>Retrieval Period</td>
<td>Finishing Period</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>Fan 1</strong></td>
<td>3.35 (3.18-3.55)</td>
<td>3.86 (3.3-4.39)</td>
<td>3.43 (3.29-3.57)</td>
</tr>
<tr>
<td><strong>Fan 2</strong></td>
<td>3.36 (3.2-3.57)</td>
<td>4.3 (3.76-4.7)</td>
<td>3.49 (3.36-3.57)</td>
</tr>
<tr>
<td><strong>Pyr Early</strong></td>
<td>3.25 (3-3.52)</td>
<td>4.26 (3.08-5.03)</td>
<td>3.45 (3.23-3.62)</td>
</tr>
<tr>
<td><strong>Pyr Late</strong></td>
<td>3.22 (2.97-3.51)</td>
<td>3.7 (2.57-4.72)</td>
<td>3.34 (3.04-3.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>µ</strong></td>
<td>0.09 (0.07-0.11)</td>
<td>0.34 (0.27-0.44)</td>
<td>0.17 (0.14-0.2)</td>
</tr>
<tr>
<td><strong>σ</strong></td>
<td>0.09 (0.07-0.12)</td>
<td>0.34 (0.28-0.44)</td>
<td>0.19 (0.16-0.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.18 (0.15-0.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.17 (0.14-0.23)</td>
</tr>
</tbody>
</table>
Table 2
Average BIC measures and number of lowest BIC measures over subjects

<table>
<thead>
<tr>
<th></th>
<th>Encoding Period</th>
<th>Retrieval Period</th>
<th>Finishing Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gamma</td>
<td>ExGaussian</td>
<td>Log-logistic</td>
</tr>
<tr>
<td>Fan 1</td>
<td>1183.7 (2)</td>
<td>1168.2 (13)</td>
<td>1174 (3)</td>
</tr>
<tr>
<td>Fan 2</td>
<td>1058.5 (2)</td>
<td>1044.2 (10)</td>
<td>1049.4 (6)</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>357.2 (4)</td>
<td>344.8 (14)</td>
<td>350.6 (10)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>380.2 (7)</td>
<td>371.2 (15)</td>
<td>375.6 (6)</td>
</tr>
<tr>
<td>Fan 1</td>
<td>1868.1 (4)</td>
<td>1864.1 (3)</td>
<td>1862 (11)</td>
</tr>
<tr>
<td>Fan 2</td>
<td>1786.2 (4)</td>
<td>1785.8 (2)</td>
<td>1783.5 (12)</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>614.4 (2)</td>
<td>611.6 (9)</td>
<td>609.9 (17)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>614.2 (2/28)</td>
<td>611.3 (3)</td>
<td>606.8 (23)</td>
</tr>
<tr>
<td>Fan 1</td>
<td>1460.2 (16)</td>
<td>1466 (1)</td>
<td>1463.9 (1/18)</td>
</tr>
<tr>
<td>Fan 2</td>
<td>1341.2 (14)</td>
<td>1347.8 (1)</td>
<td>1345.8 (3)</td>
</tr>
<tr>
<td>Pyr Early</td>
<td>456.6 (19)</td>
<td>459.6 (2)</td>
<td>457.1 (7)</td>
</tr>
<tr>
<td>Pyr Late</td>
<td>480.5 (12)</td>
<td>481.6 (0)</td>
<td>479.1 (16)</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. (a) Swimlane representation of the engagement of ACT-R modules during an associative recognition task. Boxes indicate when various modules are engaged in the task. (b) Predictions from the ACT-R model for the conditions in the experiment of Borst et al (2013). (c) Estimated stage durations using the 5-bump model of Anderson et al (2016) fit separately to the times of the different conditions.

Figure 2. (a) Correct latencies in Borst et al. (2013) for the two levels of fan and the two types of foils. (b) Correct latencies in Borst et al. (2016) for the same conditions.

Figure 3. Average correct latencies in Tenison (2016) as a function of amount of practice (number of problems solved) for repeated and non-repeated pyramid problems of various heights.

Figure 4. Comparison of the experimental procedures in the Fan task and the Pyramid task.

Figure 5. A summary of assumptions, mathematics, and computations underlying the HSMM-MVPA analysis.

Figure 6. The average likelihood of a trial from one experiment given a N-bump model estimated from the other experiment.

Figure 7. (a) The same ACT-R model for the two conditions of the two experiments with only duration of the 4th retrieval stage varying. (b) Comparison of predictions (solid lines) and estimated durations (dotted lines) for the Fan experiment. (c) Similar comparison for the Pyramid experiment.

Figure 8. t-values for sensor (blue-to-red scale ranges from -6 to 6).
Figure 9. Average residual source activity for Fan experiment and Pyramid experiment. (a) -- Talairach coordinates: -42, -60, -8; (b) LIPFC -- Talairach coordinates: -43, 23,24; motor Talairach coordinates: -42, 19,50.

Figure 10. Smoothed ML densities (a) encoding duration (time until the end of the third bump) ; (b) retrieval duration (4th flat); (c) finishing duration (time from the onset of the fourth bump until the end of the trial).

Figure 11. Fit of the Log-logistic distribution to the two conditions of each experiment where the conditions share a noise parameter and only have a difference in mean activation. Thick lines are the smoothed ML densities and the thin lines are weighted averages of individual subject fits.

Figure 12. Number of observation of total time in each 10 ms. time bin and predictions (averaged over subjects) from ExGaussian fits to total time and convolution of encoding ExGaussian, retrieval Log-logistic, and finishing Gamma.
Figure 1. (a) Swimlane representation of the engagement of ACT-R modules during an associative recognition task. Boxes indicate when various modules are engaged in the task. (b) Predictions from the ACT-R model for the conditions in the experiment of Borst et al (2013). (c) Estimated stage durations using the 5-bump model of Anderson et al (2016) fit separately to the times of the different conditions.

Figure 2. (a) Correct latencies in Borst et al. (2013) for the two levels of fan and the two types of foils. (b) Correct latencies in Borst et al. (2016) for the same conditions.
Figure 3. Average correct latencies in Tenison (2016) as a function of amount of practice (number of problems solved) for repeated and non-repeated pyramid problems of various heights.

Figure 4. Comparison of the experimental procedures in the Fan task and the Pyramid task.
Figure 5. A summary of assumptions, mathematics, and computations underlyng the HSMM-MVPA analysis.

1. Maximize the probability of the data from ∼10,000 trials.

2. Sum the probability of all (>10^8 in case of 5 bump model) placements of the bumps.

3. The probability of a resulting partition is the product of the probabilities of the flats and bumps.

4. The probability of a flat varies with the probability of its duration t_k, according to a gamma with shape 2.

5. The probability of a flat is the product of the probability of the signal values S_ki for each factor.

6. The probability of the 50 msec. sample is determined by how much signal a half sin with that magnitude will explain.

Figure 6. The average likelihood of a trial from one experiment given a N-bump model estimated from the other experiment.
Figure 7. (a) The same ACT-R model for the two conditions of the two experiments with only duration of the 4th retrieval stage varying. (b) Comparison of predictions (solid lines) and estimated durations (dotted lines) for the Fan experiment. (c) Similar comparison for the Pyramid experiment.

Figure 8. t-values for sensor (blue-to-red scale ranges from -6 to 6).
Figure 9. Average residual source activity for Fan experiment and Pyramid experiment. (a) Talairach coordinates: -42, -60, -8; (b) LIPFC -- Talairach coordinates: -43, 23, 24; motor Talairach coordinates: -42, 19, 50.

Figure 10. Smoothed ML densities (a) encoding duration (time until the end of the third bump); (b) retrieval duration (4th flat); (c) finishing duration (time from the onset of the fourth bump until the end of the trial).
Figure 11. Fit of the Log-logistic distribution to the two conditions of each experiment where the conditions share a noise parameter and only have a difference in mean activation. Thick lines are the smoothed ML densities and the thin lines are weighted averages of individual subject fits.

Figure 12. Number of observation of total time in each 10 ms. time bin and predictions (averaged over subjects) from ExGaussian fits to total time and convolution of encoding ExGaussian, retrieval Log-logistic, and finishing Gamma.
Appendix A: Review of the ACT-R Memory Theory and Predictions

This is a review of the relevant mathematics of the ACT-R memory theory as described in Anderson & Lebiere (1998). The availability of a memory $i$ is determined by its level of activation, $A_i$, which reflects both the past history of practice that the memory has received and the cues $j$ in the environment according to the equation:

$$A_i = B_i + \sum_j W_j S_{ji}$$

Activation Equation

where $B_i$ is the base-level activation, the summation is over attended cues $j$, $W_j$ are the weights given to cues $j$ (typically $= 1/n$ where $n$ is the number of relevant cues) and $S_{ji}$ are the strengths of association between cue $j$ and memory $i$. The Fan experiment manipulated the strengths of associations $S_{ji}$ while the Pyramid experiment manipulated the base level activations $B_i$.

**Base-Level Activation.** The base-level activation reflects an accumulation of individual strengthenings, each of which is decaying away with experience

$$B_i = \ln \left( \sum_{k=1}^{n} t_k^{-d} \right)$$

Base-Level Equation

where $n$ is the number of prior presentations, $t_k$ is the time that has past since that presentation, and $d$ is the decay rate. This equation has been shown to closely approximate the log odds of something appearing in the environment as well as match practice and retention effects in human memory (Anderson & Schooler,
In the case where the presentations are approximately equally spaced, this equation is closely approximated by
\[ B = \ln\left(\frac{n}{1 - d}\right) - d \cdot \ln(T) \]  
where \( T \) is the time since the memory was created. Thus, activation grows with the logarithm of practice and decreases with the logarithm of time.

**Strength of Association.** The strength of association reflects how much more likely a cue makes the memory over its base probability:

\[ S_{ji} = \ln\left(\frac{\text{prob}(i \mid j)}{\text{prob}(i)}\right) \]  
One way of manipulating strength of association is by \( \text{fan} \), which is the number of items associated to a cue. In that case \( \text{prob}(i \mid j) = 1/\text{fan} \) and the equation can be expressed as

\[ S_{ji} = S - \ln(\text{fan}_j) \]  
where \( S \) reflects \( \ln(1/\text{prob}(i)) \) and \( \text{fan}_j \) is the number of memories associated to cue \( j \).

Activation is related to time to retrieve a memory by the equation

\[ \text{Time} = Fe^{-A} \]  
where \( F \) is the latency scale constant. As developed in Anderson & Lebiere, this equation is based on the assumption that retrieval involves sorting through competing memories to identify the correct one and that the number of effective competitors increases as the the activation of the trace gets lower. Following the earlier analysis of Anderson & Schooler (1991) we assumed that the there was a Zipf's law relationship between activation and number of competitors. Anderson
(2007) sketched how this might be realized in accumulator models (e.g. Ratcliff et al. 2007, Brown & Heathcote, 2008).

With this latency equation we can turn the activation assumptions into predictions for the two experiments:

**Fan experiment.** Substituting the activation equation into the latency equation, the relevant equation for this experiment becomes

\[
Time = F' e^{-\frac{(S_1 + S_2)}{2}}
\]

Fan Experiment Equation

where \(F'\) is the latency scale that has absorbed the base-level component, which is constant across the two conditions. \(S_i\) are the strengths of association from the two cues in the probe. Their sum is divided by 2 to reflect the equal \(w_j\) weighting of the two cues. The strengths of association are \(S\) in the 1-fan condition and \(S-ln(2)\) in the 2-fan condition. In the case of targets, both cues are present and the predicted latencies are \(F'\exp(-S)\) and \(F'\exp(-[S-ln(2)])\) for the 1 and 2 fan conditions. Only one cue will be connected to the retrieved memory for foils and so the predicted times are \(F'\exp(-S/2)\) and \(F'\exp(-[S-ln(2)]/2)\). We used the value \(S=1\) from Anderson et al. (2017). Thus, the only parameter to be estimated is \(F'\), which was estimated as 1400 msec for Figure 7b.

**Pyramid experiment.** In the Pyramid experiment the associative factors are constant and the practice varies. Thus, the relevant equation for this experiment becomes

\[
Time = F'' e^{-\frac{\ln(n/(1-d))}{2}} = F'' n^{-(1-d)}
\]

Pyramid Experiment Equation

where \(F''\) has absorbed both the effects of association strengths and time that are constant. The standard value of the decay parameter \(d\) in ACT-R is .5 making our
predicted retrieval times proportional to the square root of the amount of practice, \(n\). Fitting the retrieval stage times only requires estimation of \(F''\), which was estimated as 2967 msec for Figure 7c.

In the main text we make predictions about latency distributions. Distributions of retrieval latencies in ACT-R follow from the assumption that there is trial to trial variability in activation values about the value specified by the Activation Equation. At the time of the development of the original ACT-R software, we chose a logistic distribution for the variable noise added to activation values. This was purely because it was simpler and easier to compute than a normal distribution. This assumption in conjunction with the Latency equation leads to the prediction of a Log-logistic distribution of retrieval times.

**Appendix B: Data Preparation and Multi-set Canonical Correlation Analysis (MCCA)**

The MEG data were visually inspected to reject flat or noisy channels. Next, the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011) for MATLAB was used to apply a band-pass filter (0.5-50 Hz) and subsequently down-sample the data to 250 Hz (Fan) or 200 Hz (Pyramids). Eye blinks and saccades were removed by decomposing the MEG data into independent components with EEGLAB’s (Delorme & Makeig, 2004) independent component analysis (ICA) algorithms as implemented in FieldTrip. Components were automatically marked for rejection if their correlations with either the horizontal or vertical EOG recordings were higher than 0.5 and these marked components were also manually inspected before rejection. The remaining components
were projected back to sensor space. Finally, sensor data were realigned to the average head position of all subjects.

Data from the 306 sensors was then downsampled to 10 ms. samples. Trials were removed that had extreme outliers for a sensor: First, trials were eliminated that had samples that had any observations for any sensor more than 50 standard deviations from the mean for that sensor. Then trials were eliminated that had more that .1% of their observations more than 5 standard deviations for that sensor. This resulted in the loss of 35 trials out of 6771 potentially relevant trials for the Fan data and 23 trials out of 3563 potentially relevant trials for the Pyramid.

To normalize the variance among the three types of sensors (magnetometer, planar gradiometer, axial gradiometer), the average activity of all sensors of a particular type were divided by the standard deviation for that type of sensors. For each trial and each sensor the activity was baselined by subtracting the mean activity in the 200 ms. prior to trial onset.

This sensor data as processed had two properties that made them inappropriate for a bump analysis and motivated the MCCA final step. To demonstrate these problems and prepare the data for this final step, we calculated average trial patterns for subjects by warping all trials between 600 to 2400 ms.\(^7\) to a common length of 1200 ms., by stretching or shrinking the trial by no more than a factor of 2. We averaged these warped trials to get average patterns for each sensor for each subject. Examining the resulting 120 samples-by-360 sensor arrays, two features are quite striking:

1. Sensors are poorly correlated across subjects. The average between-

\(^7\) All trials (400 to 3000 msec) were used in later analyses
subject correlation for corresponding sensors over the 120 samples was .095 for the Fan study and .192 for the Pyramid experiment. These correlations are much worse than for EEG data. This is not due to noisy data. Splitting each subject’s data into odd and even trials, the average odd-even correlation for the same sensor within a subject is .765 for the Fan study and .591 for the Pyramid experiment. This lack of between-subject sensor correspondence reflects the better spatial sensitivity of MEG, individual differences in both structural and functional anatomy of the brain, and differences in the location of the sensors relative to the brain.

2. Individual sensors are quite correlated within subjects. A principal component analysis on each subject’s data reveals that just 10 dimensions cover 97.7% of the variance for the Fan data and 97.3% of the variance for the Pyramid data. Thus, we cannot treat the dimensions independently as is assumed by the bump analysis.

We used the multi-set canonical correlation analysis (MCCA) as described in Zhang et al (submitted) to deal with both issues. Figure B1 illustrates how the method was applied to the set of 306 sensors by 120 sample matrices. These matrices were reduced by a PCA to a smaller (20 in this application) of dimensions. These 20 dimensions captured of 99% of the variance in the average subject’s data\(^8\). The MCCA finds for each subject a mapping of these 20 PCA dimensions (and hence of the original 306 sensors) onto a set of orthogonal canonical components. The components are ordered such that the first component has the highest correlation among subjects, the second component the second

---

\(^8\) The resulting PCAs are not much better correlated across subjects than the sensors: \(r = .219\) for Fan and .199 for Pyramid.
highest, etc. The bump modeling used the first 10 components.

Figure B2 displays various measures of correspondence of the 10 components. The line “Within Experiments” reflects the average correlations between subjects within an experiment. These correlations are much higher than the original sensor correlations. As a test of the reliability of the MCCA components we broke the data from each experiment into odd and even trials and independently performed separate MCCAs on the two halves. The line labeled “Within Halves” reflects the average correlation between subjects within either half of the data. It is slightly lower that the within-experiment correlation because of the halving of the data to compute the MCCA. The line labeled “Between Halves” shows the average between-subjects correlation of the odd and even halves and is a test of the reliability of the result. Except for the 10th component, these correlations are close to the within-halves correlation. As a final measure, the line “Between Experiments” gives the mutual correlations between the corresponding components of the 18 Fan subjects and the 28 Pyramid subjects. The average between-experiment correlation between corresponding components is .656. For comparison, the average correlation between non-corresponding components is close to zero (.004). These results indicate that MCCA is capturing reliable temporal trends in the data and temporal trends that appear to generalize across experiments.

Figure B3 displays the temporal pattern of the 2nd, 4th, 6th, and 8th MCCA components over the 120 samples for the two experiments. The correspondence between the two experiments is striking. All ten of the components display a pattern that can be observed in these four: There is a wavy fluctuation (but not simply sinusoidal) with more fluctuations for the later components. While the average patterns in Figure B3 are quite
regular, individual trials can be quite complex as Figure B4 illustrates. These MCCA components have complex power spectra just as the sensors they are derived from. Patterns like those in Figure B3 only emerge in average data.

The average 10 MCCA components for the experiment do a good job of capturing the variance in the average pattern of one of the subject’s 306 sensors -- average variance accounted for is .911 for the Fan experiment and .848 in the Pyramid experiment. Given that these factors are uncorrelated they provide simultaneously a reliable organization of the variance in the experiment and a compression of the dimensionality of the data. The fact that similar MCCA components appear for the two experiments is one piece of evidence that the two tasks involve the same processes. We took the subject weightings from the MCCA performed on their average data and applied it to the trial-by-trial data to represent each trial in 10 dimensions. The HSMM-MVPA bump analysis assumes that the bumps diverge from a zero baseline. To approximate a zero baseline throughout the trial, each point is calculated as the difference between it and the surrounding 31 points (-150 ms to +150 ms). We then z-scored the data from each trial to give each dimension and each trial equal weighting and to provide data with zero mean and a standard deviation of one. Figure B4 shows these normalized values for one trial. It illustrates that the trial-by-trial MCCA components display the same complex structure that the sensor signals do.

**Appendix C: Calculation of Source Activity**

Because the measured magnetic signal does not directly indicate the location and magnitude of cortical currents, we projected the sensor data onto the cortical surface using cortically constrained minimum norm estimates (MNE; Gramfort et al., 2014). The
MNE method attempts to find the distribution of currents on the cortical surface with the minimum overall power that can explain the MEG sensor data. To this end we first constructed 3D cortical surface models from the subjects’ structural MRIs using FreeSurfer (Dale, Fischl, & Sereno, 1999; Fischl, 2012). These cortical surface models were manually co-registered with the MEG data on the basis of the recorded fiducials and scalp surface points. Then, 2,562 source dipoles were placed on the gray-white matter boundary of each hemisphere, approximately 6 mm apart, and a forward operator from these sources was calculated using a single compartment boundary-element model. A noise-covariance matrix was computed for each subject from 400 ms to 50 ms before stimulus presentation on each trial. The noise-covariance matrix and the forward operator were combined into an inverse operator (loose orientation constraint of 0.1, no depth weighting, default MNE SNR of 7), which can be used to project sensor data onto the source dipoles on the cortical surface. The inverse operator was used to project the raw stimulus-locked epochs of the Pyramid sensor data and average stimulus-locked epochs per subject and condition of the Fan sensor data to the source space. These source estimates were then morphed onto the standard MNI brain using MNE’s surface-based normalization procedure (Gramfort et al., 2014).

As we did for the sensors (Figure 8) we warped each trial of source data so that the locations of the bumps were at the average positions (samples 9, 17, 27, 95, and 113) by stretching of compressing the observations between the bumps. Figure C1 shows the average activity of all sources over the course of the trials. The estimated current appears to be much larger in the Pyramid experiment. However, the difference is only marginal, reflecting the large differences in absolute value of activations obtained for different
subjects (F(1,44) = 2.93, < .1). The other two within-subject factors were significant, however – retrieval speed (F(1,44)=4.71, p < .05) and period (F(2,88)=3.65, p < .05). There were no significant interactions. Global activity tended to be greater in the faster condition and lower during the retrieval period. It can be noted in Figure 9 that there is a detectible rise in global activity during bumps and the effect of period probably reflects that there are bumps during the Encoding and Retrieval periods.

While such global effects may be of some interest, our interest is in the activity in specific regions. To remove global trends, we regressed each region of interest against the average activity on that trial and then averaged the residual activity across trials for each subject, condition and region. Figure 9 plots the residual activity in a region not predicted by the global trend.

Appendix D: Distributions of Individual Stages

Each trial yields estimates of the probability that the flats demarked by the bumps were of various durations on each trial. Averaging these trial-by-trial distributions one can get distributions of flat durations. Alternatively, one can take the maximum likelihood (ML) duration of each flat on each trial and build up distributions from the frequencies of these ML durations over trials. Figure D1 presents compares the average distributions with the ML distributions. The ML distribution tends to be a bit narrower and somewhat bumpy. While one could use either the average of trial estimate or the ML estimates, in the main paper we only use ML estimates to simplify tests about different distributions. With the ML estimates we get a single time for each period for each trial for each subject and we can investigate how probable these times would be for a particular distribution.
The shape of the distributions estimated by either method will be partly
determined by the estimation assumption of a Gamma-2 distribution that is used in the
HSMM-MVPA estimation. Nonetheless, some of these densities deviate substantially
from any Gamma distribution. In particular, Flat 1 has a peculiarly shifted density.

**Appendix E: Monte Carlo Exploration of the Fits to the Retrieval Duration**

The HSMM-MVPA bump estimation process raises questions about the
conclusions suggested by Table 1. The HSMM-MVPA bump estimation is based on
fitting all the data because its accuracy depends on having a great deal of data. However
as a consequence of this, the distributions of stage durations for different subjects are not
independent. The estimation process also works with an assumption of a Gamma 2.
Even though this does not force the distributions distribution we are fitting to be Gamma
(particularly encoding and finishing distributions that are combinations), it complicates
assessment of the relative fits. These considerations raise questions both about whether
there is some hidden bias in the HSMM-MVPA and the appropriateness of the using BIC
to decide.

Because our major interest is in the retrieval distribution we decided to further test
conclusions about its shape by Monte Carlo simulation with various assumptions about
the retrieval distribution. For each subject we generated the same number trials as that
subject had in each condition. We generated each trial by starting with the random
sinusoidal noise of the same power spectrum as the data and adding the estimated bumps
to the signal. The positions for these bumps were randomly generated for each trial
according to a set of 6 flat distributions. The flat distributions for all but the 4th retrieval
flat were based on that subject’s ML flat durations. The flat distribution for the retrieval
stage was alternatively generated according to the three distributions. For each subject these were generated with the maximum likelihood parameters for that subject. This generated three possible synthetic data sets, with approximately the same number of 10 ms. samples as the actual data, each data set based on a different assumption about the retrieval distribution. We fit a 5-bump model to these synthetic data sets just as we had to the original data set, extracted estimates of the stages, and then fit subject-by-subject Log-logistic, ExGaussian, and Gamma distributions to the maximum likelihood estimates of Stage 4 retrieval durations. We generated 100 experiments worth of synthetic data for each distribution.

Figure E1 shows the log likelihood of an average fit to a single trial of each distribution to each synthetic data set (averaged) as well as to the data. Without a complexity correction for its extra parameter, the ExGaussian tends to fit all the data sets at least as well as its competitors. As one might expect, it has the greatest advantage when with distribution was generated with an ExGaussian. More generally, each distribution has its best performance relative to the other distributions when fit to data generated with that assumed distribution – i.e. the advantage of the ExGaussian is greatest for data generated by the ExGaussian, the disadvantage of Log-logistic to the Exgaussian is least for data generated by the Log-logistic, and the disadvantage of the Gamma to the Exgaussian is least for data generated by the Gamma.

To address how strong this evidence is for Log-logistic we investigated two questions:

1. How often would an ExGaussian fit no better than a Log-logistic (as observed in

\[^9\] To avoid the impact of the occasional odd simulated subject, we took the median over subjects of their average log trial likelihoods.
the actual experiments), if the retrieval distribution really conformed to an ExGaussian? Only 1 of the 100 synthetic ExGaussian data sets yielded this result. On the other hand, 23 of the Log-logistic data sets resulted in a tie or better with the ExGaussian.

2. How often would a Gamma fit at least .048 worse than a Log-logistic (as observed) if the retrieval distribution really conformed to a Gamma? This was not observed in any of the data sets generated with a Gamma distribution. However, this strong of evidence for the Log-logistic over the Gamma also never occurred in synthetic data generated from a Log-logistic assumption. Thus, the actual data must contain evidence against a Gamma distribution that is not produced by the synthetic data.

In conclusion, the pattern observed in the actual data (near tie of Log-logistic with ExGaussian and a strong advantage over the Gamma), indicates that the Log-logistic is a better characterization of the retrieval times than the other two distributions.
Appendix Figure Captions

Figure B1. An illustration of multi-set canonical correlation applied to the MEG data.

Figure B2. The between-subject correlations between MCCA dimensions independently obtained for the two experiments and for the odd and even trials within an experiment.

Figure B3. The average pattern of the 2\textsuperscript{nd}, 4\textsuperscript{th}, 6\textsuperscript{th} and 8\textsuperscript{th} MCCA components over the 120 samples. The blue line is the pattern for the Fan experiment and the red line is the pattern for the Pyramid experiment. The values displayed are the output of the MCCA.

Figure B4. Bottom: The 2\textsuperscript{nd}, 4\textsuperscript{th}, 6\textsuperscript{th} and 8\textsuperscript{th} MCCA component on a trial that took 1 sec (100 samples). The values have been normalized as described in the text. Above: The expected magnitudes for these 4 dimensions at the maximum likelihood positions for the 5 bumps.

Figure C1. Average overall source activity after warping every trial so that the maximum likelihood locations of the bumps correspond to the average locations.

Figure D1. (a) Average distributions (blue smooth lines): average of trial-by-trial probabilities of the durations of the 6 flats; (b) ML distributions (black sometimes jagged lines): proportion of maximum likelihood durations in each 10 ms. bin.

Figure E1. Mean log likelihood of a trial generated with synthetic data based on three retrieval distributions as well as the data from the experiment. The
separate bars reflect fits of the three distributions to the resulting estimates of retrieval durations.
Figure B1. An illustration of multi-set canonical correlation applied to the MEG data.

Figure B2. The between-subject correlations between MCCA dimensions independently obtained for the two experiments and for the odd and even trials within an experiment.
Figure B3. The average pattern of the 2\textsuperscript{nd}, 4\textsuperscript{th}, 6\textsuperscript{th} and 8\textsuperscript{th} MCCA components over the 120 samples. The blue line is the pattern for the Fan experiment and the red line is the pattern for the Pyramid experiment. The values displayed are the output of the MCCA.

Figure B4. Bottom: The 2\textsuperscript{nd}, 4\textsuperscript{th}, 6\textsuperscript{th} and 8\textsuperscript{th} MCCA component on a trial that took 1 sec (100 samples). The values have been normalized as described in the text. Above: The expected magnitudes for these 4 dimensions at the maximum likelihood positions for the 5 bumps.
Figure C1. Average overall source activity after warping every trial so that the maximum likelihood locations of the bumps correspond to the average locations.

Figure D1. (a) Average distributions (blue smooth lines): average of trial-by-trial probabilities of the durations of the 6 flats; (b) ML distributions (black sometimes jagged lines): proportion of maximum likelihood durations in each 10 ms. bin.
Figure E1. Mean log likelihood of a trial generated with synthetic data based on three retrieval distributions as well as the data from the experiment. The separate bars reflect fits of the three distributions to the resulting estimates of retrieval durations.