Psychological Review
2016, Vol. 123, No. 5, 481–509

0033-295X/16/$12.00 http://dx.doi.org/10.1037/rev0000030

The Discovery of Processing Stages: Extension of Sternberg's Method

John R. Anderson and Qiong Zhang
Carnegie Mellon University

Jelmer P. Borst
University of Groningen

Matthew M. Walsh
Carnegie Mellon University

We introduce a method for measuring the number and durations of processing stages from the electroencephalographic signal and apply it to the study of associative recognition. Using an extension of past research that combines multivariate pattern analysis with hidden semi-Markov models, the approach identifies on a trial-by-trial basis where brief sinusoidal peaks (called *bumps*) are added to the ongoing electroencephalographic signal. We propose that these bumps mark the onset of critical cognitive stages in processing. The results of the analysis can be used to guide the development of detailed process models. Applied to the associative recognition task, the hidden semi-Markov models multivariate pattern analysis method indicates that the effects of associative strength and probe type are localized to a memory retrieval stage and a decision stage. This is in line with a previously developed the adaptive control of thought–rational process model, called ACT-R, of the task. As a test of the generalization of our method we also apply it to a data set on the Sternberg working memory task collected by Jacobs, Hwang, Curran, and Kahana (2006). The analysis generalizes robustly, and localizes the typical set size effect in a late comparison/decision stage. In addition to providing information about the number and durations of stages in associative recognition, our analysis sheds light on the event-related potential components implicated in the study of recognition memory.

Keywords: associative recognition, computational modeling, reaction time, hidden semi-Markov models, EEG

Even before the emergence of experimental psychology as a scientific field, it was apparent that a number of "stages" underlie behavioral responses (Boring, 1929). The stimulus must travel to the brain, the brain must process information, and the response must travel through the motor system. Further, for all but the simplest reaction time tasks, it is reasonable to suppose that multiple meaningful stages intervene between the arrival of the stimulus in the brain and the execution of the motor response. A challenge in psychology dating back to Donders (1969, translation) is how to identify the meaningful stages, how long they take, and what factors affect them. Sternberg's (1969) additive factors method led to a surge in the use of latency measures to identify stages. Since then, researchers have argued for different combinations of sequential, parallel, or overlapping stages (McClelland, 1979; Roberts & Sternberg, 1993; Schweickert, Fisher, & Goldstein, 2010). The typical approach to evaluating these accounts is to manipulate factors that affect different purported stages and test if they produce the theoretically expected results in the total response times. While researchers disagree about how to interpret these results, there is little controversy about how to understand the latency measure. Each stage progresses in time and the resulting reaction times reflect the cumulative processing of all stages. While latency has the advantage of an obvious relationship to the durations of underlying stages, it does not provide a direct measure of the individual stages but only of their cumulative time.

Recently, neural imaging techniques have been used to directly track ongoing processes (e.g., Blankertz, Lemm, Treder, Haufe, & Müller, 2011; King & Dehaene, 2014; Ratcliff, Philiastides, & Sajda, 2009; Sternberg, 2011; Sudre et al., 2012). To identify the number and duration of cognitive processing stages in functional magnetic resonance imaging (fMRI) data, we have developed a combination of hidden semi-Markov models (HSMMs) and multivariate pattern analysis (MVPA; e.g., Anderson, Fincham, Schneider, & Yang, 2012; Anderson & Fincham, 2014a,b). However, because the temporal resolution of fMRI is poor, this HSMM-MVPA methodology can only identify multisecond stages in tasks lasting on the order of 10+ s. Borst and Anderson (2015) successfully extended the methodology to electroencephalography (EEG) and parsed out much briefer stages in a memory task (data originally reported in Borst, Schneider, Walsh, & Anderson, 2013).

While neural signals from fMRI and EEG provide direct measures of underlying cognitive processes, they lack the obvious

This article was published Online First April 28, 2016.

John R. Anderson and Qiong Zhang, Department of Psychology, Carnegie Mellon University; Jelmer P. Borst, Department of Artificial Intelligence, University of Groningen; Matthew M. Walsh, Department of Psychology, Carnegie Mellon University.

This research was supported by the National Institute of Mental Health Grant MH068243 and a James S. McDonnell Foundation (220020162) Scholar Award. We thank Rob Kass for his comments on the research. The Matlab code and data for the analyses in this article is available at http://act-r.psy.cmu.edu/?post_type=publications&p=17655.

Correspondence concerning this article should be addressed to John R. Anderson, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: ja@cmu.edu

relationship of latency to stages. That is, neural signals do not directly indicate when one stage ends and another begins. To really understand what neural signals indicate about stages, one needs to postulate *linking assumptions* about how stages map onto the signals. By specifying linking assumptions one can both improve the statistical power of the machine-learning methods applied to the data, and produce theoretically more meaningful results. In their application of the HSMM-MVPA methodology to EEG data, Borst and Anderson (2015) used the simple linking assumption that periods with a constant EEG pattern corresponded to a processing stage. The current article proposes a more theoretically grounded linking assumption between the EEG signal and processing stages, which additionally provides a direct connection between the results of the HSMM-MVPA method and computational process models. The benefits of this approach are twofold: The direct connection enables strong tests of process models using neural imaging data, and process models provide detailed explanations of the stages identified in the neural imaging data.

This article contains five major sections. The first section provides background on the application of the HSMM-MVPA method to fMRI, and a description of an EEG experiment on associative recognition used in the study. The second section develops an HSMM-MVPA approach for EEG and applies it to data from the experiment. The third section introduces a theoretical model of associative recognition and shows how the model both provides an understanding of the HSMM-MVPA results and further extends them. The fourth section investigates how well the method and theoretical analysis extends to another data set collected in a different laboratory. The final section discusses the implications of this work for understanding associative recognition and for applying the HSMM-MVPA methodology more broadly.

Background: The Method and a Memory Experiment

HSMM-MVPA Applied to fMRI

Anderson and Fincham (2014a,b) applied the HSMM-MVPA analysis to fMRI data from a problem-solving task (see Anderson, Lee, & Fincham, 2014, for an application of the approach to a different task). In the task, participants solved a class of mathematical problems called pyramid problems. The HSMM-MVPA method addresses two issues. At a general level, it identifies the number of cognitive stages participants go through from perception of the problem to response generation. At a specific level, it segments each individual trial into the identified processing stages and thereby indicates the average duration of each stage over all trials. Because this is done for the different conditions in the experiment, the analysis indicates which processing stages vary in duration with experimental condition.

To identify stages, Anderson and Fincham (2014a, 2014b) combined MVPAs with HSMMs. Most neuroimaging analyses use a mass-univariate approach, which assumes that single voxels (fMRI) or sensors (EEG, magnetoencephalography) carry the relevant information. In contrast, MVPA looks at distributed patterns of brain activity across voxels or sensors (Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009). The power of this approach is seen in its successful application to the task of decoding noun representations from fMRI data (Mitchell et al., 2008) and magnetoencephalography data (Sudre et al., 2012). In the context of stage identification in fMRI, the assumption is that a constant pattern of activity across the brain signifies a certain combination of cognitive processes. Such a constant activity pattern can extend for a period of time—a processing stage. Qualitatively different patterns are interpreted as reflecting different processing stages.

Anderson and Fincham (2014a, 2014b) combined MVPA with hidden Markov models, which simulate a system that is in one of a distinct set of states at any time, and transitions to a new state at certain times (Rabiner, 1989). In the HSMM-MVPA analysis, each state corresponds to a processing stage. Because processing stages' durations are variable, Anderson and Fincham (2014a, 2014b) used a variable-duration HMM (Rabiner, 1989), which is also known as an HSMM (Yu, 2010).

The HSMM-MVPA method identified four distinct processing stages in Anderson and Fincham's (2014a, 2014b) task: (a) an encoding stage during which participants perceived the problem, (b) a planning stage during which they planned a solution path, (c) a solving stage during which the solution was calculated, and (d) a responding stage during which the response was entered. The encoding stage did not vary in duration with problem type, as all problems had similar perceptual features. The planning stage was longer for problems that demanded a novel solution path, the solving stage was longer for problems with more calculations, and the responding stage was longer when answers required more keystrokes.

Leaving out details that are specific to analysis of fMRI data, we will discuss seven features of the approach that are relevant to the EEG analysis in more detail (see Figure 1 for illustration).

Brain signatures. The brain pattern across voxels (or sensors for EEG data) that defines a stage is called its brain signature. To deal with the highly correlated nature of brain signals in different regions, Anderson and Fincham (2014a, 2014b) first performed a spatial principal component analysis (PCA) on the voxel data and

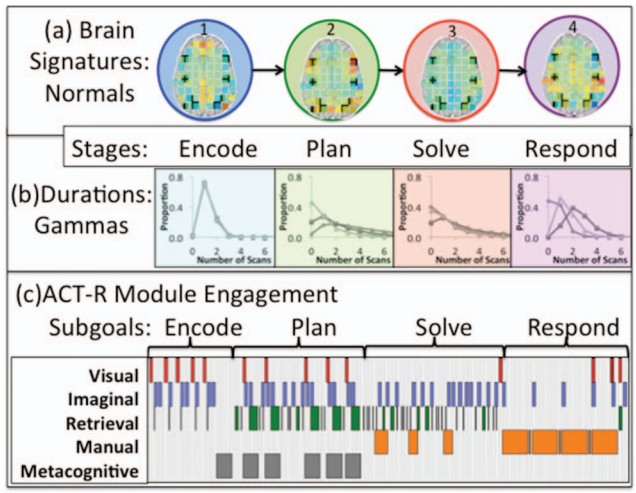

Figure 1. Illustration of the components in the approach developed for functional MRI: (a) brain signatures that define stages, (b) distribution of stage durations in different conditions, and (c) a swimlane representation of an ACT-R model with subgoals corresponding to the stages. The boxes in a row represent when a module is engaged. ACT-R = adaptive control of thought–rational. See the online article for the color version of this figure.

retained the first 20 PCA components (which accounted for about two thirds of the variance in the data). Thus, a brain signature is a constant activity pattern across a number of PCA components, rather than across raw voxel data. However, whole brain activation patterns for a stage can be reconstructed from the PCA weights that define the brain signature (Figure 1a). On each iteration of the HSMM's expectation maximization algorithm (Dempster, Laird, & Rubin, 1977), the MVPA estimates patterns in this 20 dimensional space for each stage that maximize the explained variance in the imaging data. The brain signature of a stage is by definition constant across experimental conditions,[1] and is assumed to represent a constant mixture of cognitive processes.

Distributions of stage durations. On different trials subjects can spend different amounts of time in a stage, making the model semi-Markov because stages are of variable duration. Variability in each stage's duration is represented by a discrete approximation to a gamma distribution (Figure 1b, discrete because the data describes activity in discrete intervals). If an experimental factor affects a stage, the stage will have different gamma distributions of durations for different values of that experimental factor (analogous to the logic of Sternberg's additive factors). This contrasts with brain signatures, which are constant for a stage across experimental conditions. As described above, Anderson and Fincham (2014a, 2014b) identified stages corresponding to encoding, planning, solving, and responding. The durations (and thus the gamma estimates) of the planning, solving, and responding stages varied by mathematical problem type, but the duration of the encoding stage did not.

Parameter estimation. To identify the optimal number of stages to describe the data, HSMMs with different numbers of states are fit to the data. For each HSMM, the standard expectation-maximization algorithm (Dempster et al., 1977; Rabiner, 1989) is used to estimate the optimal brain signatures (20 parameters per state) and gamma distributions (two parameters per state). The dynamic programming techniques of HSMMs make this a tractable enterprise (Yu, 2010).

Bottom-up stage identification. Using HSMMs with more stages and with more condition-specific durations will always improve the likelihood of the data, but at some point the improvement is merely due to overfitting. To prevent overfitting, Anderson and Fincham (2014a, 2014b) used leave-one-out cross-validation (LOOCV). They fit a HSMM to all but one subject, and then used the group's parameter estimates to calculate the likelihood of the remaining subject's data. Additional stages and gamma distributions were only justified if they increased the likelihood of the data for a significant number of subjects (as determined by a sign test). As discussed above, Anderson and Fincham (2014a, 2014b) identified four stages in their mathematical problem solving task, notionally labeled encoding, planning, solving, and responding. While they applied these labels to the stages, they had really only identified an HSMM that characterized the imaging data. It remained to be determined what processes actually took place during the stages.

A process model. Guided in part by the HSMM, Anderson and Fincham (2014a, 2014b) developed an adaptive control of thought–rational (ACT-R) model that was capable of solving the math problems (Figure 1c) and that explained what was happening in the stages. The model involved the interaction of five modules in ACT-R: a visual module for encoding the problem and tracking the response output, an imaginal module for holding and transforming problem representations during problem solving, a retrieval module for retrieving information about problem type as well as general arithmetic facts, a manual module for controlling hand movements, and a metacognitive module for holding and transforming information about solution methods. The swimlane representation in Figure 1c shows when the modules are active during the 13 s that it takes for the model to solve a particular problem. The activity of each module is represented along a row and the widths of the boxes reflect how long each module was engaged.

The fMRI linking assumption. Earlier work (Anderson, 2007; Anderson et al., 2008) explicitly linked the activity of ACT-R modules to fMRI blood-oxygen-level-dependent (BOLD) responses. When a module is active, it makes demands on corresponding brain regions. Demand is convolved with a hemodynamic response function to produce the fMRI BOLD response in those regions. This is basically an extension of the typical assumption in fMRI brain analysis, using swimlanes as design matrices (Borst, Nijboer, Taatgen, Van Rijn, & Anderson, 2015; Friston, Ashburner, Kiebel, Nichols, & Penny, 2007). Given the slow and sloppy hemodynamic response function, the BOLD response is insensitive to the exact time of these demands, which are typically on the order of 50 to 500 ms (Figure 1c). Instead, the BOLD response reflects the average activity in a module over several seconds. The ACT-R model sets subgoals that result in approximately constant module activity over several seconds, as indicated in Figure 1c. The stages identified in the fMRI analysis correspond to these subgoals. The relatively constant mix of module activity during a subgoal results in a constant demand on the brain regions that support those modules, resulting in the stable pattern that is the stage's brain signature.

Increased insight. Informed by this linking assumption, Anderson and Fincham (2014a, 2014b) were able to run the ACT-R model through 128 different problems and identify the durations of stages for each. They used these model-based stage durations per problem to construct a HSMM based on the ACT-R model. The model-based HSMM contained far fewer parameters and was better able to account for the data than the HSMM that was discovered in a purely bottom up manner. This highlights the fact that there are an infinite variety of HSMMs, and that without some top-down theoretical constraints one will find a good approximation at best. In addition to finding a better statistical characterization of the data, Anderson and Fincham (2014a, 2014b) produced a process model that explained what was happening during each stage. The model provided an overall computational understanding of how people solve a class of mathematical problems in which they must reflect on and extend what they have previously learned.

The HSMM-MVPA method is theory-agnostic. As long as one has a task model that specifies when various cognitive resources are required to perform the task, one can link the model to fMRI data in the same way. This can be done even without a priori predictions about what neural regions implement the cognitive

[1] Alternatively, it is possible to use different brain signatures for different experimental conditions, if it assumed that subjects go through qualitatively different processing stages in different conditions.

processes evoked by the task. If one can define periods of time when the mixture of cognitive activity is relatively constant, the HSMM-MVPA methodology can assess how well these model-defined periods account for the neural data.

Although the HSMM-MVPA method successfully extracted multisecond stages from fMRI data, models in cognitive science typically assume basic processes that take well under 1 s. fMRI is not sensitive to the exact sequence of these processes or their durations, but rather their relative frequencies over multisecond periods. This article extends the HSMM-MVPA methodology to EEG data in order to capitalize on the greater temporal precision that EEG offers.

A Study of Associative Recognition

We will apply the HSMM-MVPA methodology to data from an experiment by Borst et al. (2013) that manipulated associative fan—the number of episodic associations that a word has with other words in memory. Fan has strong effects on response time and accuracy, with higher fan resulting in longer latencies and lower accuracy (Anderson, 2007; Anderson & Reder, 1999; Schneider & Anderson, 2012). The experiment examined the potential effects of associative fan on EEG correlates of recognition memory, namely, the FN400 and the parietal old–new effect. The FN400 is a frontocentral negativity that is larger for items judged as "new" versus "old" (Curran, 2000), whereas the parietal old–new effect is a late posterior positivity that is larger for items judged as "old" versus "new" (Curran, 2000; Düzel, Yonelinas, Mangun, Heinze, & Tulving, 1997). According to dual-process models of memory, these components reflect qualitatively distinct memory processes (Diana, Reder, Arndt, & Park, 2006; Rugg & Curran, 2007; Yonelinas, 2002). The FN400 corresponds to a familiarity process that provides information about whether an item has been seen, but does not involve the retrieval of specific details from when it appeared. In contrast, the parietal old–new effect corresponds to a recollection process that does involve the retrieval of associated information from when an item appeared.

The experiment consisted of a study phase and a test phase. In the study phase, subjects memorized 32 word pairs that varied in fan. Each word in a Fan 1 pair appeared only in that pair, whereas each word in a Fan 2 pair appeared in two pairs.[2] In the test phase, subjects viewed probe word pairs. Pairs included targets (i.e., two words previously studied together), repaired foils (i.e., two words previously studied separately), and new foils (i.e., two novel words not previously studied). Each target and repaired foil appeared 13 times over the course of the experiment whereas each new foil appeared only once. Subjects responded "yes" to targets, and "no" to repaired foils and new foils. The test phase contained five trials types (Fan 1 target, Fan 2 target, Fan 1 foil, Fan 2 foil, and new) that appeared with equal frequency (see Table 1). Twenty subjects completed 13 blocks of 80 trials, yielding a total of 20,800 potential trials.

During the test phase, the EEG was recorded from 32 Ag-AgCl sintered electrodes (10–20 system). Electrodes were also placed on the right and left mastoids. The right mastoid served as the reference electrode, and scalp recordings were algebraically rereferenced offline to the average of the right and left mastoids. The EEG signals were amplified by a Neuroscan bioamplification system (Neuroscan, Inc., Sterling, VA) with a bandpass of 0.1–70.0 Hz and were digitized at 250 Hz. Electrode impedances were kept below 5 kΩ.

The EEG recording was decomposed into independent components using the EEGLAB infomax algorithm (Delorme & Makeig, 2004). Components associated with eyeblinks were visually identified and projected out of the EEG recording. A 0.5–30 Hz bandpass filter was then applied to attenuate noise. Epochs beginning 200 ms before probe presentation and continuing 160 ms beyond the response were extracted from the continuous recording and corrected using a linear baseline (cf. Anderson & Fincham, 2014a). The baseline was defined as the slope between the average of −200 to 0 ms before stimulus onset and the average of 80 to 160 ms after the response, and was subtracted from the data in the trial. We applied this linear baseline on a trial-by-trial basis to remove random signal drift within trials. Averaging across trials reduces drift in standard event-related potential (ERP) analyses, but because the HSMM-MVPA analysis is applied to single-trial data, it is important to remove drift from within trials. For EEG analyses, we only included correct trials with latencies within three standard deviations of the mean per condition per subject, and shorter than 3,000 ms. In total, 12.2% of trials were excluded from analysis.

Table 1 shows behavioral results from the five test conditions. Participants responded more quickly to Fan 1 than to Fan 2 items, and to targets than to repaired foils. They responded most quickly to new foils. All differences among conditions in latency were significant, replicating the classic effects of fan and probe type.

Figure 2 shows activity at two electrodes over the frontal (left column) and posterior (right column) scalp. Stimulus-locked activity is shown in Figure 2a for the first 750 ms (top), and response-locked activity is shown in Figure 2b for the final 750 ms (bottom). Data from all trials of a condition are averaged together within participant, and grand averaged waveforms are created from the average of the individuals' waveforms. The stimulus-locked averages in Figure 2a show a number of classic ERP components including the N1 and P2. These components are typically observed in paired associate studies (e.g., Bader, Mecklinger, Hoppstädter, & Meyer, 2010; Rhodes & Donaldson, 2008).

Figure 2a also shows what may be interpreted as an FN400: New foils are accompanied by a negativity around 400 ms, though the differences between new foils and all other items extend beyond frontal sites and over the central scalp (Figure 2c). The topographical distribution of the effect is consistent with the results of two other studies in which participants were tested on studied versus new word pairs (Bader et al., 2010; Wiegand, Bader, & Mecklinger, 2010). The results of those studies were interpreted in terms of conceptual fluency and a decreased N400—the integration of word pairs during the initial study phase increased conceptual fluency and subsequently decreased the amplitude of the N400 during the test phase. The FN400 and N400 accounts of our data explain the early effect as a positivity related to familiarity superimposed on targets and repaired foils (FN400), or a negativity related to processing difficulty superimposed on new foils (N400). As we will see, the HSMM-MVPA points to yet another interpretation of this effect. Lastly, Figure 2d most clearly shows the parietal old–new effect: Waveforms are more positive

[2] When a word appeared in two pairs, both appearances were either in the first position or the second position.

Table 1
Associative Recognition Task: Example Material, Conditions, Observations, and Model Predictions

Study phase	Test phase	Condition	Error rate	Latency (ms)[a]	Useable trials	Model's latency
Flame–cape	Flame–cape	Fan 1 target	3.1%	994	3,767	948
Metal–motor	Metal–motor	Fan 2 target	6.7%	1,189	3,473	1,216
Metal–spark	Flame–deck	Fan 1 foil	3.1%	1,061	3,714	1,120
Jelly–motor	Metal–peach	Fan 2 foil	6.0%	1,342	3,444	1,303
Book–deck	Jail–giant	New foil	.1%	702	3,873	696
House–peach						
Flag–peach						
. . .						

[a] Mean latencies are from useable trials only.

before participants respond to targets than to repaired foils. The HSMM-MVPA will also inform the interpretation of the parietal old–new effect.

HSMM-MVPA for EEG

This section describes an HSMM-MVPA analysis of EEG that can analyze the signal from individual trials and identify stages in each. We begin by presenting our proposal for a linking assumption that connects processing stages to the EEG signal. We then define the approach and use it to identify the number of stages and their durations in the associative recognition task. This section describes a bottom-up analysis of the EEG signal, while the next section shows how a theoretical model can further guide the interpretation of the data.

Linking Assumption: Modeling the EEG Signal

Given its high temporal resolution, EEG is ideal for learning about the latencies and durations of cognitive processes. However, extracting this information from the EEG signal is challenging. A modulation in the EEG signal that is not time-locked with an observable event, such as the presentation of a stimulus or the commission of a response, may be distorted or lost in the average waveform (Luck, 2005). This problem becomes more pronounced as the variability of response latencies and endogenous ERP component onsets increases with task difficulty and/or complexity.

Researchers have proposed several solutions to this problem. These solutions include template matching (Woody, 1967), peak-picking (Assecondi et al., 2009; Gratton, Kramer, Coles, & Donchin, 1989), response-time binning (Poli, Cinel, Citi, & Sepulveda, 2010), independent component analysis (Jung et al., 2001), and maximum-likelihood estimation (D'Avanzo, Schiff, Amodio, & Sparacino, 2011; Tuan, Möcks, Kohler, & Gasser, 1987). A complete review of these methods is beyond the scope of this article. However, as of yet, we lack a method to detect the onsets of multiple signals that occur with variable latencies in a trial, and to do so while making minimal assumptions about the shape and spatiotemporal distribution of those signals. Here, we offer a powerful way of aligning trials that follows from either of two theories about the source of the EEG signal.[3]

Makeig et al. (2002) described these two theories of ERP generation. According to the **classical** theory, significant cognitive events produce phasic bursts of activity in discrete brain regions (Shah et al., 2004). This results in the addition of a sinusoidal peak to an ongoing signal of uncorrelated sinusoidal variation (see Figure 3). Averaging the signal across trials will identify the peak and the other variation will average to zero. To the extent that the timing of the peak is jittered due to trial-to-trial variability, the average signal will show a greater width and lower magnitude. With enough variability the peak will disappear completely from the averaged signal. According to the *synchronized oscillation* theory, significant cognitive events cause the phase of the oscillation in a certain frequency range to reset (Basar, 1980). By this view, it is temporal synchronization, rather than the collection of oscillations, that change during task processing. The frequency range that is reset in the synchronized oscillation theory can be mapped to the frequency of the sinusoidal peak in the classic theory: This will produce waveforms very similar to a classic bump with frequency *freq* if the maximum frequency of the reset range is $2 \times freq$.[4] In fact, Yeung and colleagues (Yeung, Bogacz, Holroyd, & Cohen, 2004; Yeung, Bogacz, Holroyd, Nieuwenhuis, & Cohen, 2007) showed that the classic and synchronized theories of ERP generation could produce indistinguishable waveforms. Our simulated results in the Appendix use Yeung et al.'s generator of the EEG signal.

According to either theory, if we could identify the locus of the component or the oscillation reset in each trial, we could identify when significant cognitive events occurred. In this article we use the classical theory as the hypothesized generator of the EEG signal mainly because of its conceptual simplicity. Yeung et al. were interested in detecting an ERP component called the error-related negativity, whereas we will explore whether the superposition of components on ongoing neural activity can be used more generally to detect the onsets of multiple significant cognitive events. Specifically, we look for *bumps*—multidimensional distributions of scalp activity (i.e., across electrodes) that begin to rise with the onset of a significant cognitive event. Bumps are conceptually related to brain signatures in the HSMM-MVPA analysis of

[3] As both theories provide a rationale for our linking assumption, we do not choose between the two.

[4] The mean value of sine functions with frequencies uniformly from 0 to F Hz at time t after a reset is $[(1 - \cos(2 \times F \times t \times \pi)]/(2 \times F \times t \times \pi)$ which rises to a maximum at $t = .371/F$ and returns to zero at $t = 1/F$. After returning to zero, the reset rapidly dampens because of the t in the denominator.

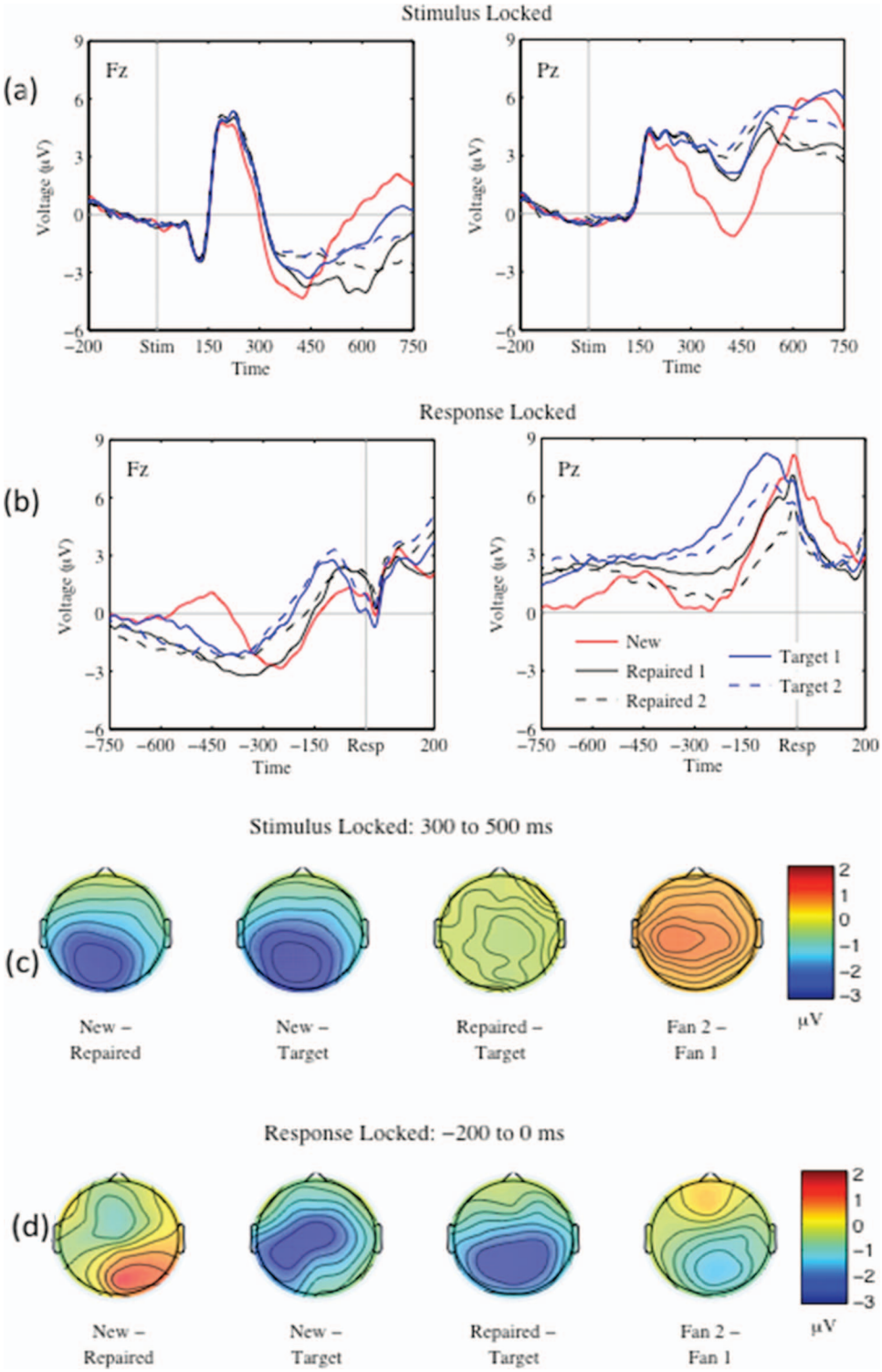

Figure 2. (a) Stimulus-locked averages of the Fz and Pz electrodes. (b) Response-locked averages of the Fz and Pz electrodes. (c) Topographic distribution of differences between conditions from 300 to 500 ms after the stimulus onset. (d) Topographic distribution of differences between conditions from −200 to 0 ms prior to response. Fz = frontal; Pz = parietal. See the online article for the color version of this figure.

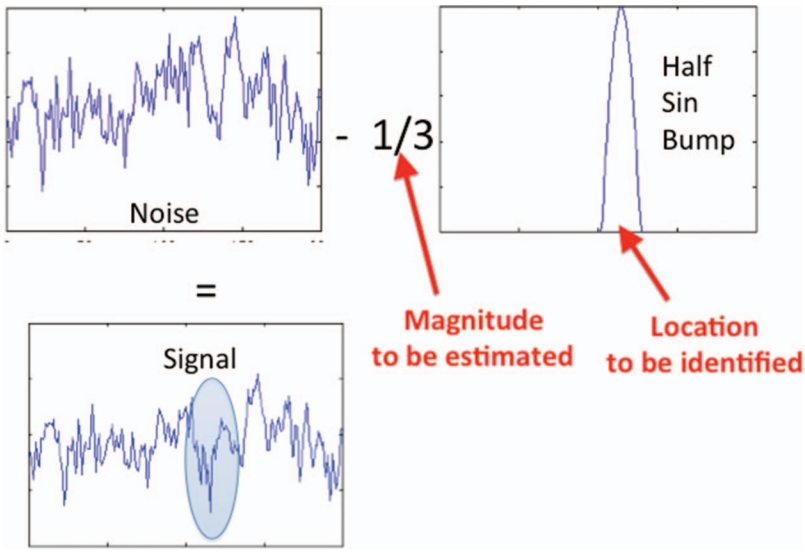

Figure 3. Illustration of Yeung and colleagues' (Yeung et al., 2004; Yeung et al., 2007) addition of a bump to an ongoing signal of sinusoidal noise. See the online article for the color version of this figure.

fMRI data. These bumps, which are latent in the EEG signal, summate with one another and with ongoing sinusoidal noise as in Figure 3, to produce the peaks and dips seen in ERPs.

Figure 4 shows the scalp topologies of five bumps identified in the EEG data of this experiment. Each bump is modeled as a 50-ms[5] half-sine multidimensional peak across the scalp added to the ongoing EEG signal, and each is interpreted as the beginning of a significant change in the information processing. As Figure 4 illustrates, variable-duration *flats* separate bumps. The new process signaled by the bump continues throughout the flat but the oscillatory EEG signal returns to ongoing sinusoidal noise with a mean of 0. Thus, an intermediate stage is comprised of a bump plus the subsequent flat. The first and last stages have somewhat different interpretations. The first stage is initiated by trial onset and does not include a bump. It includes the time for the signal to reach the brain, and so reflects both precortical processing and time until the signal initiates cognitive processing. The last stage terminates with the response, and so its end does not necessarily reflect the conclusion of an ongoing cognitive process.

The identification of bumps differs from the HSMM-MVPA approach used for fMRI data, and the approach used by Borst and Anderson (2015) for EEG data. Those approaches looked for periods of time with constant patterns of activity (i.e., brain signatures), which were defined as stages. Here, rather than looking for constant patterns of activity, we look for the bumps that mark the onsets of stages. If the assumptions underlying this analysis are correct, the EEG signal will return to sinusoidal noise around 0 after each bump. As we will see, this is true for all of the flat periods except for the one encompassing the parietal old–new effect. In addition to its somewhat better characterization of the data, this approach enables a precise connection between the neural data and a process model of the task. However, qualitatively the results are quite similar to the results of the Borst and Anderson (2015) analysis.

HSMM-MVPA Applied to EEG

We performed two steps of data reduction to simplify the analysis and make the computations more efficient and tractable. First, we down-sampled the data to 100 Hz (i.e., 10-ms samples). Second, to deal with the highly intercorrelated nature of the EEG signal we performed spatial PCA (i.e., across electrodes) and retained the first 10 PCA components. These accounted for 95% of the variance in the signal.[6] These PCA components were z scored for each trial so that each dimension on each trial has mean 0 and the same variability (i.e., 1) as all other dimensions on all other trials. Thus, the data stream no longer included 32 correlated electrodes sampled every 4 ms and with large intertrial variability, but rather 10 orthogonal PCA components sampled every 10 ms and with constant variability across trials. Still, EEG scalp patterns can be reconstructed from the normalized PCA representations (see Figure 4).

We made several assumptions to facilitate the analysis of the temporal structure of the signal. First, the bumps were given a 50-ms width (i.e., half sines with frequency 10 Hz). Thus, a bump occupies 5-cs samples. Using the value of a sine function halfway through the 10-ms intervals results in the following weights, P_i, for the five samples: 0.309, 0.809, 1.000, 0.809, and 0.309. The choice of 50 ms as the width of the bump is somewhat arbitrary. In the Appendix, we show that even if the true bumps are of different widths (ranging from 30 to 150 ms), an analysis using 50-ms bumps still correctly identifies the durations of the underlying stages. The choice of 50 ms is possibly briefer than the width of

[5] This produces peaks with the same width as the function *Makinen* at the website http://www.cs.bris.ac.uk/~rafal/phasereset/, which is based on Figure 1 Mäkinen, Tiitinen, and May (2005) generated by resetting frequencies between 4 HZ and 16 Hz. This produces bumps that peak at 22 ms and return to 0 at 50 ms, on average.

[6] The first PCA component accounts for 65% of the variance.

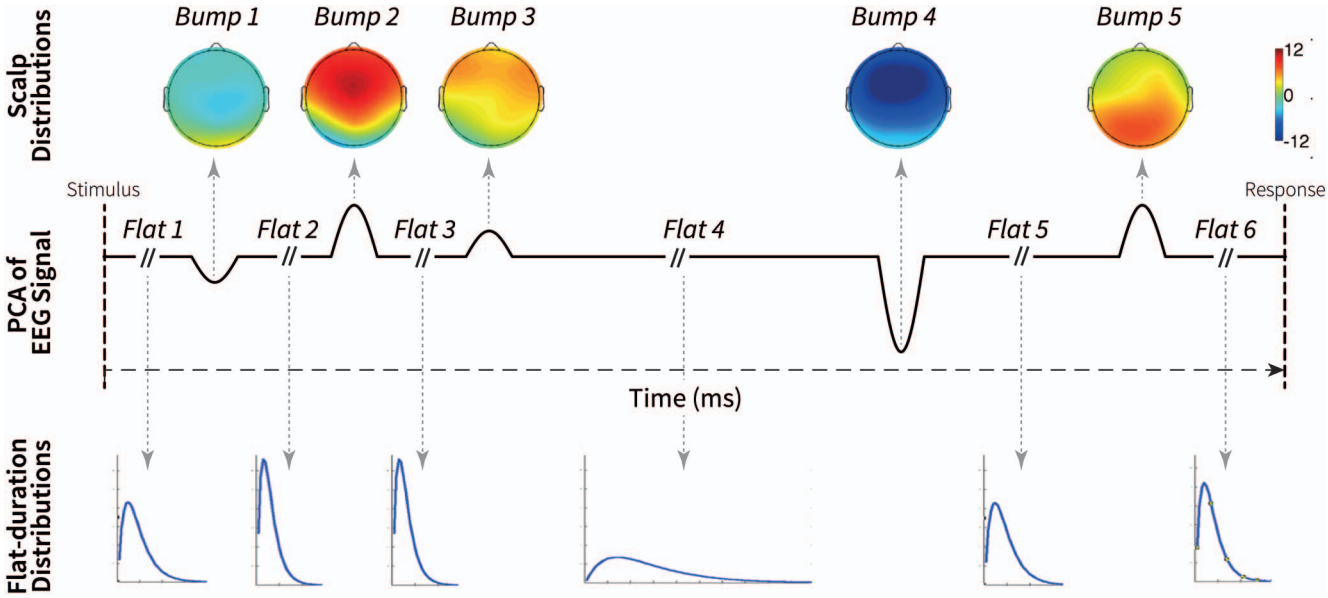

Figure 4. An illustration of the five-bump solution for the full data set: the scalp profiles of the five bumps and the distributions of the durations of the six flats for Fan 1 targets. The individual blue line graphs are the probability distributions for the durations of the flats. PCA = principal component analysis. See the online article for the color version of this figure.

true bumps, but its narrowness promotes precision in the identification of stage boundaries. The analysis also assumes the bumps do not overlap, but as the Appendix shows, the method accurately recovers bump locations even if they occasionally do.

Another assumption that we made was to model the distributions of the flat durations between the bumps as gamma distributions with a shape parameter of 2 and a free scale parameter. The results are not sensitive to the exact choice of the shape parameter, but setting it simplifies estimating the flat distributions between bumps. To denote that the shape parameter is fixed to 2, we refer to these as gamma-2 distributions. Gamma-2 distributions have the right sort of shape for a temporal distribution with a long right tail, but the small shape parameter still allows for a wide range of possible durations on individual trials. See the Appendix for further discussion of the shape parameter.

An n bump HSMM requires estimating $n + 1$ stage distributions to describe the durations of the flats plus the n multidimensional bumps that mark stage transitions (see Figure 4). A bump is defined as a two-dimensional matrix of values B_{ij}, where j indexes the five samples and i indexes the 10 PCA dimensions. The bump values are calculated as $B_{ij} = P_j \times M_i$, where P_j are the five sample weights and M_i are the 10 PCA magnitudes for Bump i. The half-sine shape sets the P_j and the M_i is estimated for each bump. Thus, estimating n bumps requires estimating $10 \times n$ magnitudes M_i, and $n + 1$ scale parameters to characterize the gamma-2 distributions. As described in the Appendix, these parameters are estimated to maximize the likelihood of the EEG signal. The Appendix demonstrates the robustness of parameter estimation and stage identification by exploring the method's performance on synthetic data where the precise generating model is known.

We use an HSMM as our tool of choice because it allows us to represent and reason about the ambiguity of how an individual trial may be divided into stages. For an n-bump model, we need to consider all the ways a trial might be divided into $n + 1$ stages. The constraint is $t_1 + 5 + t_2 + 5 + \ldots + 5 + t_{n+1} = T$, where T is the total number of the samples in the trial and the 5 represents the durations of the bumps (5 cs) between the flats. The number of such possible partitions is $(T - 4n)!/[(T - 5n)! \times n!]$, which would be an astronomical number if all partitions were calculated separately. Given a set of parameters, the dynamic programming methods of an HSMM efficiently calculate the summed probabilities of all such partitions. These probabilities are very small and so we typically calculate the log-likelihood of the data from the trial. The expectation maximization algorithm associated with HSMMs estimates a set of parameters that maximize the summed log-likelihood of all the trials in a set rather than maximizing the log-likelihood of each trial separately.

Identifying the Number of Stages and Their Durations

Because every bump takes 50 ms and the shortest observed latency was 410 ms,[7] we could identify as many as eight bumps (and thereby nine stages). Figure 5 illustrates the solutions obtained by increasing the number of stages from one to nine and placing the bumps at their average locations (average trial length is 1,067 ms). The bumps are remarkably stable in scalp profile and location as more stages are added. The solution for n bumps is obtained independently from that of the $n + 1$ bumps and so there is no requirement in the estimation process that the $n + 1$ bumps must include the n bumps. While the procedure will find maximum likelihood estimates for any number of bumps, inclusion of too

[7] The actual value of the shortest latency was 414 ms, but the downsampling algorithm compressed this to 41 samples at 10 ms.

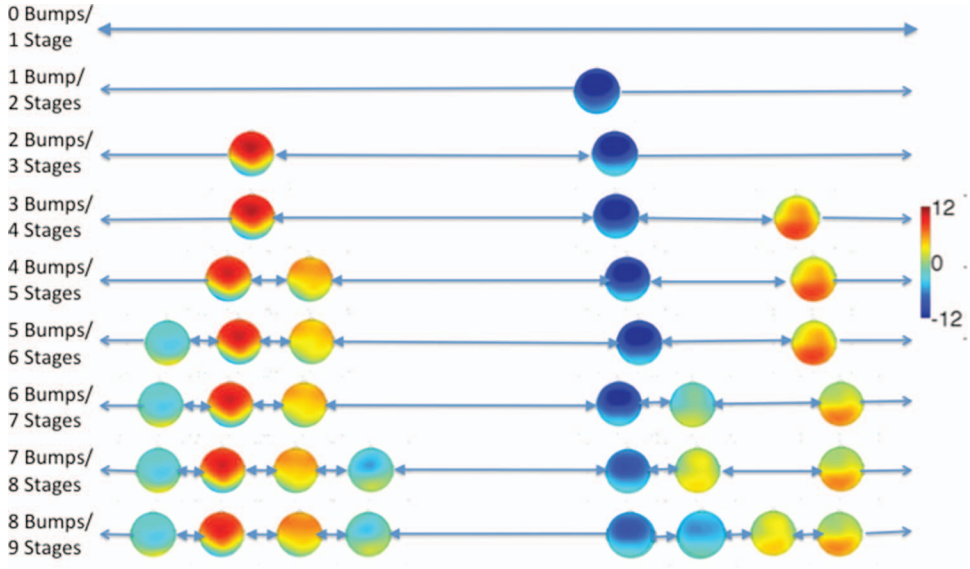

Figure 5. An illustration of the scalp profiles and mean durations between bumps for fits of zero- to eight-bump hidden semi-Markov models (HSMMs) to the data. See the online article for the color version of this figure.

many may lead to overfitting. Some of the later cases seem to involve simply duplicating a bump, creating a pair of nearly identical adjacent bumps.

As with the fMRI data, we performed LOOCV by fitting an HSMM to the data from all but one subject, and then using the HSMM to estimate the likelihood of the remaining subject's data.[8] Figure 6 shows the gain in log-likelihood over a no-bump HSMM for each number of bumps. The log-likelihood rises steadily from one to five bumps. Seventeen of 20 subjects are better predicted by the five-bump HSMM than by any HSMM with fewer bumps ($p <$ .005 by sign test). The log-likelihood increases by an average of 127.0 when going from five to six bumps and is greater for 15 of 20 subjects ($p < .05$ by sign test). However, we are inclined to accept the five-bump solution. This is partially because we will describe a simpler five-bump ACT-R model that has a 124.2 log-likelihood advantage over the six-bump HSMM and fits 12 of the subjects better. The additional bump in the six-bump model occurs during a flat period in the five-bump model that does not have an average of 0 as assumed. The nonzero average reflects the parietal old–new effect and the additional bump in the six-bump captures this elevated flat period. As noted in Anderson and Fincham (2014b), increased variance accounted for does not always correspond to a better explanation of what is happening. Figure 7 provides information on the stability of parameter estimates across subjects:

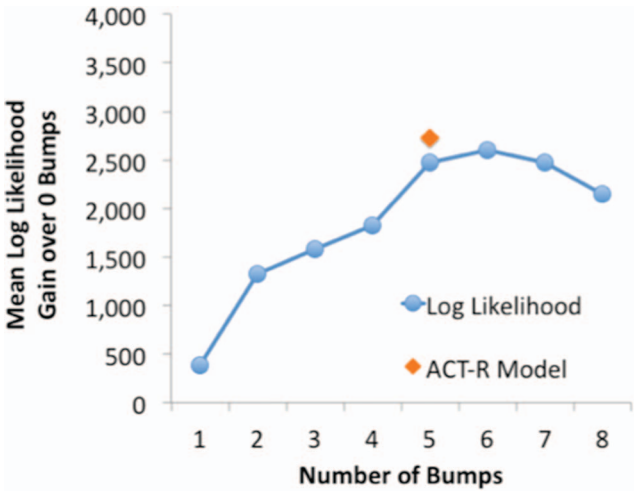

Figure 6. Mean improvement of multibump models (ranging from one to eight) over zero-bump model based on leave-one-out cross-validation. ACT-R = adaptive control of thought–rational. See the online article for the color version of this figure.

Gamma-2 scale (Figure 7a). The duration of each stage is determined by the scale parameter of the gamma-2. Twice this parameter is the mean duration of the stage. To determine the between-subjects variation in this parameter we estimated durations separately for each individual. To keep the definition of stages the same for each subject we fixed the magnitude parameters—that is, the scalp topography of the bumps—to the global estimates. Figure 7a plots the ± 1 *SD* range of the resulting scales. The first three flat durations are much less variable than the last three. Since the estimation procedure is anchored to both stimulus presentation and response, this is not caused by an accumulation of temporal uncertainty. Rather, the later periods are more variable across individuals.

[8] Although there are metrics for penalizing models for their extra parameters like Bayesian information criterion (Kass & Raftery, 1995), they do not extend in simple form to situations where there are so many parameters (Berger, Ghosh, & Mukhopadhyay, 2003) or where observations are not independent as is true of EEG data (Jones, 2011). In contrast, we have found that cross-validation methods offer an effective way to assess models and identify when the extra model complexity is justified (cf. Shiffrin, Lee, Kim, & Wagenmakers, 2008).

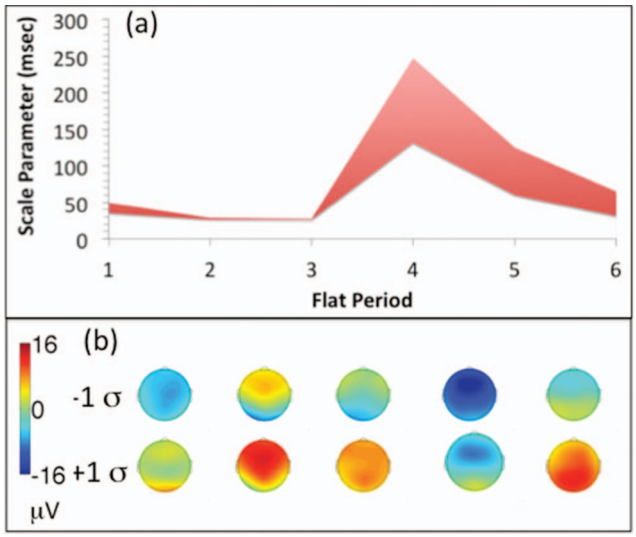

Figure 7. (a) Standard deviation (±1) in estimates of flat durations when fit to individual subjects. (b) Standard deviation (±1) in estimates of scalp profiles when fit to individual subjects. See the online article for the color version of this figure.

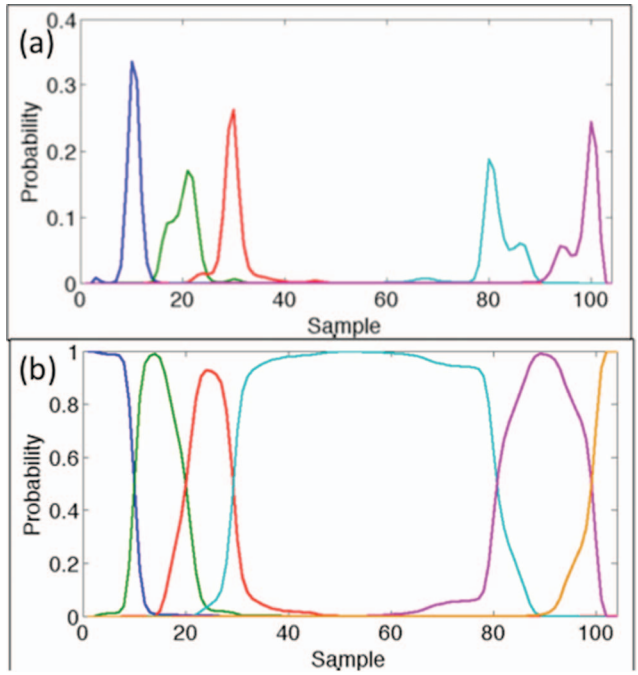

Figure 8. (a) Probability that the five bumps are centered on each time point of a prototypical trial. (b) The probability that various time points fall within each of the six stages for that trial. See the online article for the color version of this figure.

Bump magnitudes (Figure 7b). We fixed the gamma-2 scale parameters to the global scale parameters and estimated the bump magnitudes separately for each individual. Every bump has at least one of the 10 PCA dimensions with a value significantly different from 0 ($p < .005$, corrected for multiple tests). As the PCA values are not meaningful in themselves we reconstructed the scalp profiles from these parameters. Figure 7b displays ±1 *SD* in terms of voltage.

While there are individual differences in the durations of the flats and the scalp profiles of the bumps, there are also general consistencies; for instance, the fourth flat period is the longest for 19 of 20 subjects and the fourth bump shows the most negative voltage for 16 of 20 subjects.

To this point, we have fit the same HSMM to all trials, but behavioral response times varied by condition (see Table 1). This must show up in the durations of some of the stages. The estimation process delivers for each trial a distribution of possible locations of the bumps. Figure 8a shows the distributions of bump locations (center points) for a trial that took 1.04 s (104 samples of 10 ms). From these, we can derive the probability that the subject is in a particular stage at any point in time, as illustrated in Figure 8b. We can then estimate the duration of the stage for that trial as the mean number of samples (area under the curve for that stage in Figure 8b). Averaging these single-trial estimates gives an estimate of the average duration of the six stages for the conditions of the experiment, which are shown in Figure 9. The durations of the first three stages do not vary by conditions, but Stages 4 and 5 show substantial differences among conditions. Additionally, Stage 6 is about 26 ms briefer for new foils than for other probe types.

To determine whether differences among conditions were real, we fit HSMMs with condition-specific gamma-2 distributions for certain stages. A HSMM with condition-specific distributions only for Stages 4 and 5 results in better LOOCV (for all 20 subjects) than the HSMM that produced the estimates in Figure 9. A HSMM with condition-specific distributions for *all* stages only fit eight of 20 subjects better in LOOCV than the HSMM with condition-specific distributions for Stages 4 and 5 only. This supports the conclusion that only Stages 4 and 5 have significantly different durations. Eliminating the differences among conditions for either Stage 4 or 5 results in significantly worse fits (18 of 20 subjects for Stage 4, and 15 of 20 for Stage 5) indicating that these differences are real. Figure 9 further suggests that the duration of Stage 5 might be the same for all targets and repaired foils and different for new foils. Indeed, a HSMM with two distributions for Stage 5 (one for new items and one for all targets and repaired foils) is no worse (nine of 20 subjects worse in LOOCV).

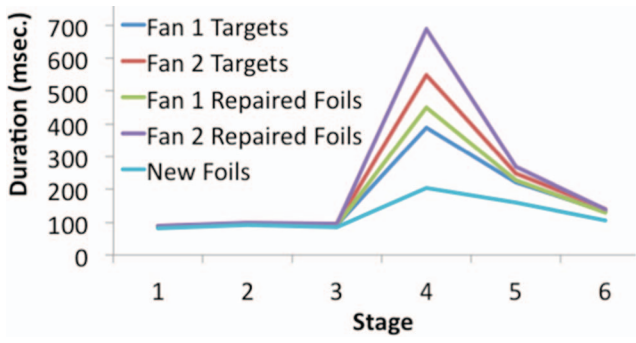

Figure 9. Mean duration of the stages in the various conditions. Note that these values were obtained fitting the five-bump hidden semi-Markov model to all the data (no separate parameter estimation per condition). See the online article for the color version of this figure.

To summarize the bottom-up analysis of the EEG data, at least five bumps (and six stages) can be reliably identified in the trials. While the spatial profiles of the bumps and the durations of the stages vary somewhat across individuals (see Figure 7), highly significant regularities are apparent as well. The differences in duration across conditions are localized to Stages 4 and 5, with Stage 4 showing differences among all conditions and Stage 5 showing shorter durations for new foils. What is lacking from this analysis is an interpretation of the stages and their corresponding latencies. The process model we describe next provides such an interpretation.

A Model of Associative Recognition

As already anticipated in Figure 6, we have created a HSMM that uses top-down constraints from a cognitive model to yield better LOOCV performance than any of the bottom-up HSMMs considered. This HSMM is based on an existing ACT-R model for classifying targets and repaired foils that has been used to fit numerous behavioral and fMRI data sets (e.g., Anderson & Reder, 1999; Danker, Gunn, & Anderson, 2008; Schneider & Anderson, 2012; Sohn, Goode, Stenger, Carter, & Anderson, 2003). Because the model has only been used to distinguish between targets and repaired foils, we needed to extend it to respond to the distinctive features of new foils. Figure 10 provides an ACT-R swimlane representation of the extended model.

Figure 10a shows the standard ACT-R retrieve-to-reject model for targets and repaired foils. Each probe results in the retrieval of the best matching memory, which is compared to the probe. If the retrieved word pair matches the encoded probe the model responds positively, and if it does not match the model responds negatively. At a high level, this involves encoding the probe, engaging in associative retrieval, deciding if there is a match, and responding. However, as the swimlane representation indicates, more detailed processing takes place in the modules. These module activities are initiated by production rules, which are triggered by the current conditions of the environment (e.g., stimulus present) and the state of the system. Module activities, in turn, change the state of the system and so trigger new productions. To go through the productions one by one,

1. The cortical appearance of the probe (35 ms after its physical appearance) triggers the first rule, which causes the visual module to encode one of the words.

2. The completion of the first encoding triggers the next production rule, which causes the visual module to encode the second word and the retrieval module to retrieve the meaning of the first.[9]

3. The completion of activity in the visual and retrieval modules triggers the third production rule, which causes the retrieval module to request retrieval of the most active associative trace of a word pair that involves the meaning of the first word. If there is a matching pair, that pair will be retrieved. If there is not a matching pair, a pair involving the first word but not the second will be retrieved. The duration of the associative retrieval stage is variable, and depends on fan and whether the word is a target or foil, as described below.

4. The completion of retrieval triggers the fourth production, which causes the imaginal module to store and compare the retrieved memory with the probe. The fourth production also causes the manual module to begin response preparation. Everything about the right-hand response can be prepared except for the finger, which depends on the decision.

5. The completion of the comparison process triggers the final production, which causes the manual module to program the responding finger (index for "yes," middle for "no") and to execute the response.

To further elaborate the EEG linking assumption within the framework of the ACT-R theory: A production evokes a change in neural processing, which produces a phasic neural response characterized by a bump. Thus, within the framework of the ACT-R theory, *an EEG bump marks the firing of a production.* Also, as indicated in Figure 10, the firing of a production initiates a new processing stage.

The timing of bumps is determined by the timing of the various ACT-R modules. With a few exceptions, the timing parameters were based on established conventions for ACT-R modeling:

Productions: Each production takes 50 ms. This is the convention in ACT-R and in other production system models like Executive Process-Interactive Cycle (EPIC) (Meyer & Kieras, 1997).

Visual: The time for visual encoding is conventionally set to 85 ms. Here, we changed the parameter to match the speed of early processing and made it 45 ms.

Retrieval: The words that made up targets and repaired foils were seen dozens of times during initial learning and over the course of the experiment. As a consequence of this practice, we assume that retrieval of the first word's meaning occurs more rapidly than the time to visually encode the second word. The duration of associative retrieval, however, is much slower and is based on the fan model proposed in previous articles (e.g., Schneider & Anderson, 2012). The time to retrieve an association depends on the amount of activation spread to it from the two words in the probe. The amount of activation from word i to its associates (A_i) is determined by its fan,

$$A_i = S - \log(fan).$$

S is the maximum associative activation for the experimental material, which is a free parameter. The activation that can be spread is divided by the number of sources—thus, with two words in the probe pairs, each can spread $[S - \ln(fan)]/2$. Both words will spread activation to the matching trace in the case of targets, and only one word will spread activation to nonmatching traces in the case of repaired foils. The retrieval time for a memory is an exponential function of the amount of activation it receives:

[9] Although we use the term *meaning* throughout this section, what is being retrieved in ACT-R is just a chunk that serves as the link between the letter string on the screen and the stored information about the word.

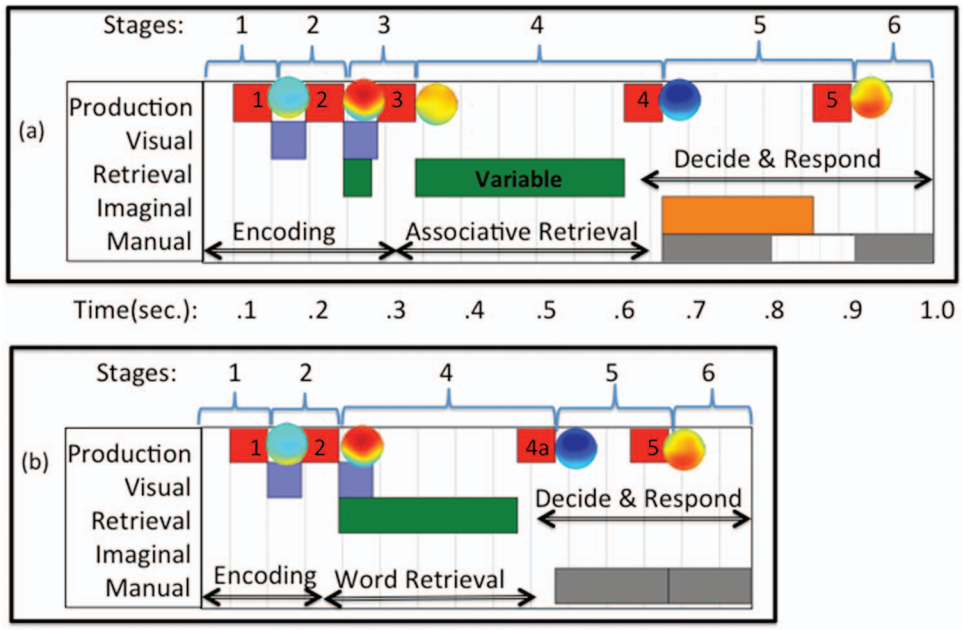

Figure 10. (a) Swimlane representation of the ACT-R retrieve-to-reject model for targets and repaired foils. (b) Swimlane representation of the ACT-R model for new foils. Stage 1 is 85 ms (35 ms to trigger first production plus 50 ms for the production). Stages 2 and 3 are 95 ms (45 ms for visual encoding plus 50 ms for the production). Stage 4 is the retrieval time plus production time. Stage 5 in Panel A is 250 ms in (200 ms for comparison plus 50 ms for the production). Stage 5 in Panel B is 150 ms (100 ms for motor preparation plus 50 ms for the production). Stage 6 is 110 ms (50 ms to complete motor preparation and 60 ms for response execution). ACT-R = adaptive control of thought–rational. See the online article for the color version of this figure.

$$\text{Targets: } t = Fe^{-[s-\ln(\text{fan})]}$$

$$\text{Foils: } t = Fe^{-[s-\ln(\text{fan})]/2},$$

where F is the latency scalar, a second free parameter. S was estimated as 1 and F as 720 ms based on the behavioral data (see Table 1).

Imaginal: Following the convention of fMRI models (Borst, Taatgen, Stocco, & Van Rijn, 2010; Danker, Gunn, & Anderson, 2008; Sohn et al., 2003), the imaginal module takes 200 ms to compare the retrieved item with the representation of the encoded probe.

Manual: The timing of the manual module was set using default ACT-R parameters, which are taken from EPIC (Meyer & Kieras, 1997). The preparation of motor features (style and hand) after the fourth production takes 100 ms. The preparation of the response finger after the fifth production takes 50 ms, and the actual physical key press takes an additional 60 ms.

Figure 10b shows an extension of the model to process new foils (based on a suggestion in Borst & Anderson, 2015). New foils are rejected based on the low experimental familiarity of their words, which are seen only once in the experiment. New foils are processed the same as all other probe types through the firing of the second production, which retrieves the meaning of the first word. Because the first word of a new foil has not been seen in the experiment, it takes longer to retrieve its meaning (estimated to be 256 ms). The model uses the duration of this retrieval to reject the foil, allowing the model to bypass the third production. As with all other probe types, the model must then prepare a motor response (Production 4a) and execute the response (Production 5).

Model-Based Fits to EEG Data

We used the model times to construct an HSMM for the EEG data by setting the gamma-2 distributions to have the same means as the ACT-R stage durations. Thus, the HSMM estimated the magnitudes of the bumps but not the time parameters. Based on the process model, new foils should be accompanied by four bumps (and five stages), and all other probes should be accompanied by five bumps (and six stages). We compared this ACT-R HSMM to the five-bump bottom-up HSMM where a separate time parameter was estimated for each stage of each condition. We also compared the ACT-R HSMM to a five-bump bottom-up HSMM with only four bumps for new foils (and with separate time parameters for each stage of each condition). The bottom-up HSMMs have greater flexibility than the ACT-R HSMM because the average durations of each stage are free parameters. Thus, they fit the data better. However, in LOOCV, the ACT-R HSMM performed slightly better in terms of mean log-likelihood, and fit about half of the subjects better (comparison with five-bump bottom-up HSMM: average log-likelihood 27.3 better and fit nine of 20 subjects better; comparison with four-/five-bump bottom-up HSMM: log-likelihood 2.3 better and fit 10 of 20 subjects better).

Figure 11 compares the mean ACT-R times with the stage times from the four-/five-bump bottom-up HSMM, fit to all subjects.[10] In the bottom-up model we allow all stage durations to vary by conditions, and not just the fourth and fifth stages. Remarkably, the major effect of condition is almost entirely contained in the fourth stage. The general correspondence is extremely close (overall correlation .992). There are small differences in the durations of Stages 5 and 6 in the bottom-up HSMM (Part b) that are not predicted by the ACT-R model (Part a). These differences may not be real, however. As we noted above, a model free to estimate separate durations does no better than the restricted ACT-R HSMM in LOOCV.[11]

Electrode Activity Anchored by Model Events

The bumps reflect points of significant change in information processing. We can use the model-based bumps to align the EEG data according to the onsets of the model's stages. For each trial, we find the times where the probability of the bumps are maximum. For instance, for the trial in Figure 8, these are samples 10, 21, 30, 80, and 100. We then warp the six intervals for a condition (five intervals for new foils) to have the same durations as the average stage durations for that condition by expanding or contracting the electrode activity between the anchors.[12] This warps each trial to have a length equal to the length of the average trial in a condition and with the critical model events occurring at the same times (see Wang, Begleiter, & Porjesz, 2001, for a related application of dynamic time warping to ERPs). Figure 12 shows averages of the warped trials for the frontal (Fz) and parietal (Pz) electrodes. Conditions with longer response times stretch further forward in the stimulus-locked display of electrode Fz and further backward in the response-locked display of electrode Pz. Figure A4 in the Appendix examines the accuracy of bump-locked averages using synthetic data where the bump locations on individual trials are known. The maximum likelihood locations of bumps like those in Figure 5 deviate on single trials from the true locations from under 50 ms (Bump 1) to over 100 ms (Bump 4) on average. Though these are relatively high single-trial errors, the Appendix shows that we have enough trials to ensure that the average plots in Figure 12 are quite representative of the true structure of the trials. The estimated average bump locations are never more than one 10-ms sample from their true locations.

Both electrodes show activity related to all of the bumps. The first bump in Figure 12 resembles the N1. The N1 is typically interpreted as an index of visual attention (Luck, Woodman, & Vogel, 2000), consistent with its role in the ACT-R model of signaling the request to visually process the first word. The second and third bumps are correlated and contribute to the P2 in all conditions except for new foils. Given that both bumps signal word encoding in the model, the correlation between their activity patterns is expected. The midearly time course and anterior distribution of these bumps are consistent with the P2 (Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991). The absence of the third bump for new foils explains why their P2, as seen in the ERP waveforms (see Figure 2), is smaller and briefer.[13] As noted earlier, the N1 and P2 have been observed in other studies of paired associates recognition involving words. Further, research on word reading has shown that frequent words produce a larger P2 (Dambacher, Kliegl, Hofmann, & Jacobs, 2006). The massive repetition of words from targets and repaired foils in our experiment effectively increases their frequency. The P2 has also been related to lexical access (Almeida & Poeppel, 2013), which is consistent with our interpretation of the second bump as reflecting the signal to retrieve word meaning.

The fourth bump produces what at first appears to be an FN400 for new foils in Figure 2. However, the topography of the affect is unusually posterior for an FN400 (Figure 2C; Figure 10). Alternatively, the fourth bump might reflect an N400 for new foils, owing to participants' lower conceptual fluency with those items relative to targets and repaired foils. Although this account explains the topographic distribution of the effect, it does not explain why the fourth bump appears in *all* conditions, just at variable latencies. What are we to make of this bump?

The frontal distribution of the fourth bump and its negative polarity are consistent with the N2, an ERP component not usually considered in the context of recognition memory paradigms. The N2 is typically seen in two-alternative forced-choice (2AFC) tasks when participants must inhibit a prepotent response. According to the conflict-monitoring theory of the N2 (Yeung, Botvinick, & Cohen, 2004), the anterior cingulate cortex (ACC) monitors for response conflict and increases top-down control upon detecting the coactivation of conflicting responses. Owing to the location of the ACC and the orientation of pyramidal neurons within it, activation of this region produces a frontocentral negativity in the EEG signal. As the fourth bump in the ACT-R model initiates the decision phase, we propose that the bump occurs at the moment of maximum response conflict within the trial. The late and variable latencies of the fourth bump in the different conditions arise from the time to complete retrieval and enter the decision phase. The relatively early and consistent onset of the N2 around 200 ms in 2AFC tasks reflects the fact that participants must simply encode a stimulus before entering the decision phase in those tasks.

In the ACT-R model, the fifth bump reflects the switch to response execution. The flat region preceding this bump at electrode Pz (Figure 12b) differs substantially among conditions. This is the only period that shows effects of condition on electrode values after warping the data to maximum likelihood locations. Given the late, posterior distribution of the fifth bump, the difference between targets and foils likely corresponds to the classic

[10] See the caption of Figure 10 for how the ACT-R times were determined by model parameters.

[11] To the extent that there are not clear signals of the bumps, the estimation process has a tendency to let latency differences in one stage "bleed" into adjacent stages. This causes conditions with longer overall latency to have slightly longer times in multiple stages.

[12] A later bump can have a maximum likelihood position earlier than a previous bump, but this only happened on 2.3% of the trials, which were then excluded.

[13] When we fit new foils using the same five-bump model that was used for the other probes, the warped signal shows two bumps. However, both bumps are weaker than in the other conditions and shifted to positions adjacent to the original bump (Bump 2 is shifted one 10-ms sample forward and Bump 3 is shifted two 10-ms samples backward). In this case, we suspect that the "true" bump is duplicated. In contrast, as displayed in Figure 12a, if we allow just a second bump for the New items, it has the same size as the other conditions and the same time of appearance.

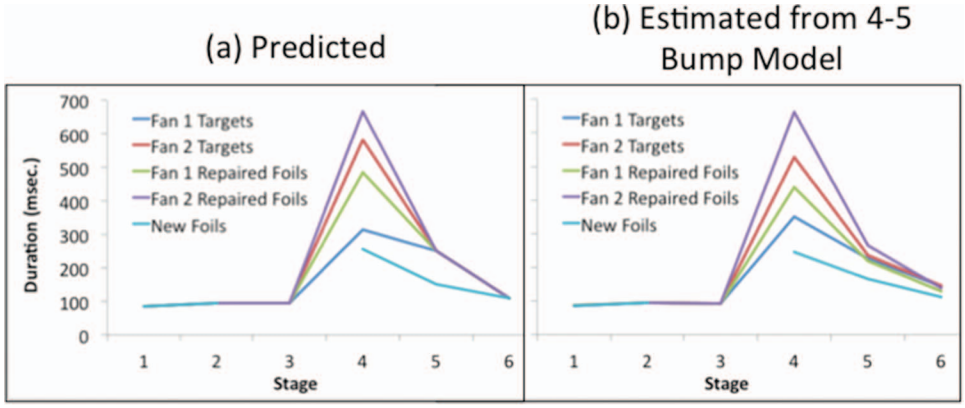

Figure 11. (a) The predicted durations of the stages according to the ACT-R model. See text and Figure 10 for an explanation of these durations. (b) The estimated durations obtained by fitting targets and repaired foils to a five-bump HSMM separately for the four conditions and fitting new foils to a four-bump HSMM (with the third bump excluded). ACT-R = adaptive control of thought–rational; HSMM = hidden semi-Markov model. See the online article for the color version of this figure.

parietal old–new effect (Curran, 2000; Düzel et al., 1997). Fan also has an effect, with Fan 1 items producing a greater positivity. To assess statistical significance, we measured the voltage at Pz during each trial at the sample identified as the peak of the last bump. We also measured the voltage at the sample identified as the middle of the flat preceding the last bump. A 5 (Condition) × 2 (Middle of Flat, Middle of Bump) within-subjects ANOVA showed significant effects of condition, $F(4, 76) = 15.74$, $p < .0001$, and of point, $F(1, 19) = 213.14$, $p < .0001$. These factors did not significantly interact, $F(4, 76) = 2.25$, $p > .05$, indicating that differences among conditions were stable during this period. The mean electrode values for the five conditions were 6.7 μV for Fan 1 targets, 5.5 μV for Fan 2 targets, 4.6 μV for Fan 1 foils, 3.5 μV for Fan 2 repaired foils, and 3.6 μV for new foils. All pairwise differences were significant (all $p < .05$) except for the differences between new foils and Fan 2 repaired foils. A 2 (Probe Type) × 2 (Fan) ANOVA on targets and repaired foils showed a highly significant effect of probe type, $F(1, 19) = 36.36$, $p < .0001$, and fan, $F(1, 19) = 23.72$, $p < .0005$, with no interaction, $F(1, 19) = 0.28$.

In the ACT-R model, the stage coinciding with the parietal old–new effect involves processing of the memory trace after its retrieval. We suggest that the effect is a sustained response reflecting the different activations of the retrieved memories in the various conditions. The mean voltage at Pz during this period is strongly correlated ($r = .896$) with model-based activation values (1.00 for Fan 1 targets, 0.30 for Fan 2 targets, 0.50 for Fan 1 repaired foils, 0.15 for Fan 2 repaired foils, and 0.00 for new foils). The parietal old–new effect has been linked to the finding from fMRI studies that parietal activation differs following the presentation of old versus new items (Wagner, Shannon, Kahn, & Buckner, 2005). The engagement of the parietal cortex in these circumstances may relate to its role in the maintenance and binding of recovered information from episodic retrieval (Vilberg & Rugg, 2008). These considerations indicate that the parietal old–new effect may reflect a

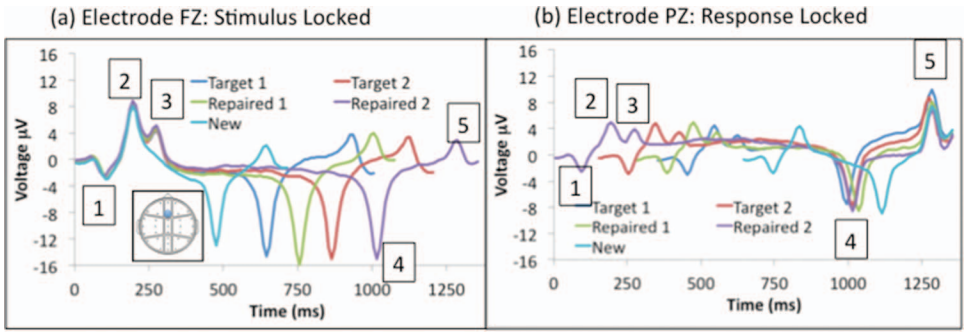

Figure 12. Average electrocardiographic data after warping every trial so that the maximum likelihood locations of the bumps correspond to the average locations for that condition. (a) The Fz electrode with all conditions starting from stimulus presentation. (b) The Pz electrode with all conditions ending with the response. The locations of the bumps are noted for the repaired-2 condition. Fz = frontal; Pz = parietal. See the online article for the color version of this figure.

sustained difference, rather than a transient change as signified by a bump. The fifth bump, then, reflects a switch from such maintenance processes to response execution.

Generalization to a Sternberg Working Memory Task

Sternberg's (1966) working memory task initiated much of the research on latency-based analysis of mental processing stages. Therefore, this seemed the natural task to test how well our analysis and model of EEG generation would generalize. We chose to apply the method and model to an EEG data set reported in Jacobs et al. (2006).[14] In the task participants were asked to judge whether a particular consonant was a member of a set of two, four, or six studied letters. As in our fan experiment, the task involved visual presentation of alphanumeric information and a manual response. The task also involved recognition memory, if of a rather different character. The classic effect in this paradigm is a near linear increase in response latency with set size. According to the classic explanation offered by Sternberg (1969; see Figure 13a) the effect of memory set size was due to a serial comparison of the probe with the elements of the memory set held in working memory.

An existing ACT-R model for the task (Anderson, Bothell, Lebiere, & Matessa, 1998) offers a different explanation and attributes the set size effect to retrieval from declarative memory. This model predates the integration of visual, imaginal, and manual modules into ACT-R and simply estimates an intercept to capture perceptual, decision-making, and motor factors. According to the model, items in the set are stored in memory. When the probe appears, the model attempts to retrieve the probe from the memory set; that is, the probe serves as the retrieval cue. The decision is made using a retrieve-to-reject strategy as in the fan experiment. An item is always retrieved from the memory set but it will not match the probe if the probe is a foil. Similar to the fan model, the activation of the retrieved item will vary with the number of alternatives. When more items are in the list, less activation will spread to each item (the fan effect) and judgment times will be slower.

Figure 13b shows the natural extension of the ACT-R model from the fan task to the Sternberg task. Because one letter is presented, there is only the need for a single visual encoding. Based on the fact that ACT-R model has four productions, we would expect that four bumps would yield the best fit to the EEG data. As Figure 13 indicates, we can map the four stages of the Sternberg task onto the four module activities of the ACT-R model. In effect, the Sternberg model also predicts a four-bump model of the EEG data. Both models predict that the effect of set size will be in the third stage, although their interpretations of that stage differ.

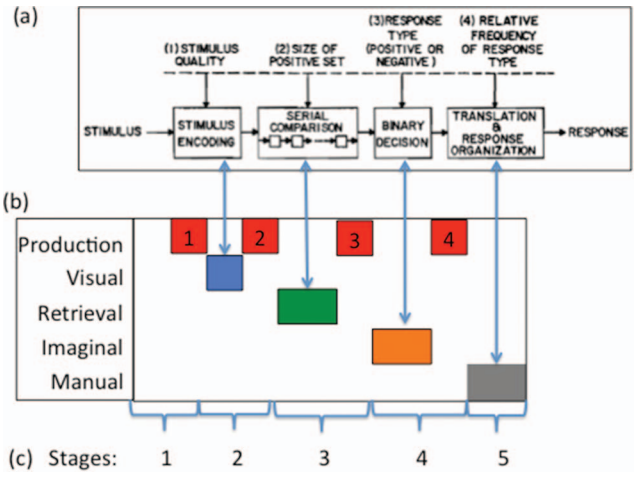

Figure 13. (a) Sternberg's model of the working memory task. (b) The ACT-R model. (c) How they map onto the five stages of a four-bump model. ACT-R = adaptive control of thought–rational. See the online article for the color version of this figure.

Methods and Results

The Jacobs data set involved 18 participants each performing 576 trials. After rejecting trials according to the same criteria used in the fan experiment, 8,412 trials remained. For these data, there was a clear effect of set size on latency (621, 667, and 719 ms for sizes two, four, and six), $F(2, 34) = 34.12$, $p < .0001$. Foils were not significantly slower than targets (678 vs. 654 ms), $F(1, 17) = 2.78$, $p > .1$.

The EEG was recorded during the presentation of the memory set and the test probe, though we focus only on the test portion of each trial as we did for the fan experiment. This is the portion of the trial where both the Sternberg and ACT-R models apply. The EEG was recorded using a 129-channel geodesic sensor net (for additional recording details, see Jacobs et al., 2006). We excluded six exogenous electrodes located around the eyes and analyzed the signal from the remaining 123 electrodes. As the fan data was referenced to the average of the right and left mastoids, we referenced Jacobs et al.'s data with respect to the average of electrodes E56 and E106, which are closest to the reference electrodes in our recording montage.[15]

As in our analysis of the fan experiment, we down-sampled the data to 100 Hz, applied a PCA to the data from the electrodes, and retained the top 10 components. These accounted for 88% of the variance in the data. We applied the same HSMM analysis as in the fan experiment to these data to look for 50-ms bumps. As the fastest trials were 300 ms, we fit models with up to six bumps. Figure 14a shows the mean improvement of models with increasing numbers of bumps over a no-bump model based on LOOCV. In line with the expectations from both the Sternberg and the ACT-R models, four bumps provide the best fit. The four-bump model clearly outperforms the one-, two-, and six-bump models (fitting 16, 14, and 16 of participants better). The four-bump model fits 11 of 18 participants better than the five-bump model and is more parsimonious. The four-bump model also fits 13 of 18 participants better than the three-bump model, which is marginally significant ($p < .1$, two-tailed sign test).

[14] This data set and others are available from the Kahana Laboratory at http://memory.psych.upenn.edu/Electrophysiological_Data. We thank Joshua Jacobs for his help in understanding details of the data and the experiment.

[15] Jacobs et al. used an average reference. Our basic conclusions do not change when we use an average reference; however, using a simulated mastoid reference facilitates comparison of the bump profiles between experiments.

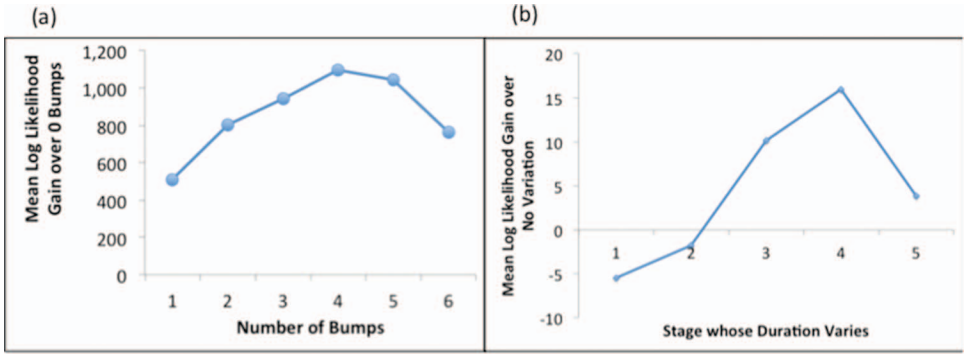

Figure 14. (a) The mean gain in log-likelihood per participant in leave-one-out cross-validation (LOOCV) for multibump models (ranging from zero to six) over a zero-bump model. (b) The mean gain in LOOCV by allowing each stage of the four-bump model to vary with set size versus a model in which all stage durations are constant across conditions. See the online article for the color version of this figure.

Figure 15 shows the electrode activity for Bumps 1, 2, 3, and 4. The mapping of Figure 13b onto the ACT-R model in Figure 10 suggests that these bumps correspond to Bumps 1, 2, 4, and 5 from the fan experiment. Taking the 32 electrodes from the geodesic sensor net closest to the electrodes from the Neuroscan Quick-Cap used in the fan experiment, we computed the mean correlations between different pairs of bumps from the two experiments (Table 2a) and the mean deviations in the voltages of the electrodes (Table 2b). Bumps 2, 3, and 4 of this experiment clearly match Bumps 2, 4, and 5 of the fan experiment. The correspondence between the first bumps in each model is less clear, in part because neither shows much variation from zero. When looking at which bump in the fan model matches the first bump in the Sternberg model most closely, the first bump has lowest mean deviation and second highest correlation. Alternatively, when looking at which bump in the Sternberg model matches the first bump in the fan model most closely, the first bump has lowest mean deviation and third highest correlation.

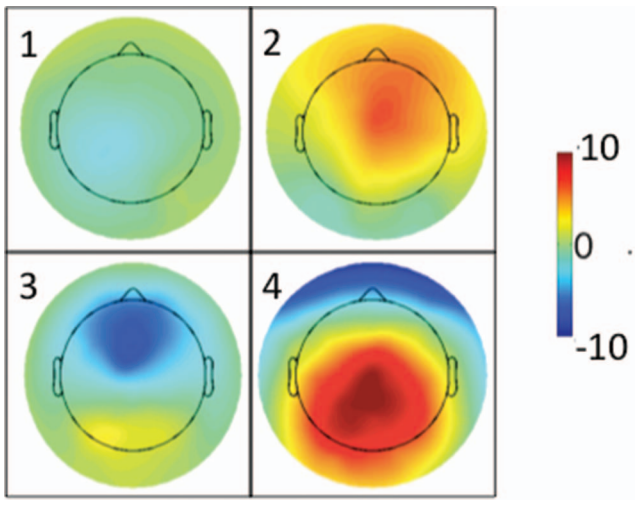

Figure 15. Mean electrode activity reconstructed for the four bumps by averaging the observed voltages at the time of the maximum-likelihood samples for each bump and during each trial. See the online article for the color version of this figure.

Given the superiority of the four-bump model, we asked which of the resulting five stages showed effects of set size. Figure 14b shows the results of a LOOCV analysis where we allowed each of the stages, one by one, to vary in duration. Contrary to the Sternberg and ACT-R models, the best performing model is one in which the fourth stage varies with set size. The model is significantly better than models that allow Stages 1, 2, or 5 to vary (performing better for 16, 16, and 17 participants), and is marginally better than a model that allows Stage 3 to vary (13 of 18 participants, $p < .1$). Allowing both the third and fourth stages to vary does not result in a better average prediction as measured by log-likelihood, and only fits nine of 18 participants better.

Figure 16 shows the estimated stage durations as a function of set size with the large effect of set size concentrated in Stage 4. Besides the fact that set size impacts the second-to-last stage, there are two other differences from the fan experiment (Figure 11b) involving the durations of the first and last stages. First, Stage 1 is estimated to be 35 ms shorter than in the fan experiment. We are not sure whether this discrepancy is real, or caused by differences in how stimulus presentation is synchronized with screen updates

Table 2
Correspondence Between Fan Bumps and Sternberg Bumps

	Fan bump				
Sternberg bump	1	2	3	4	5
	Correlation				
1	.210	−.082	.274	.271	−.729
2	−.726	.894	.780	−.795	−.394
3	.681	−.854	−.684	.863	.478
4	.027	−.146	−.413	−.099	.898
	Mean deviation				
1	2.396	7.791	6.067	7.504	8.374
2	8.988	2.181	3.521	16.098	5.685
3	4.872	12.177	10.257	4.602	10.846
4	11.414	8.477	8.419	18.479	4.838

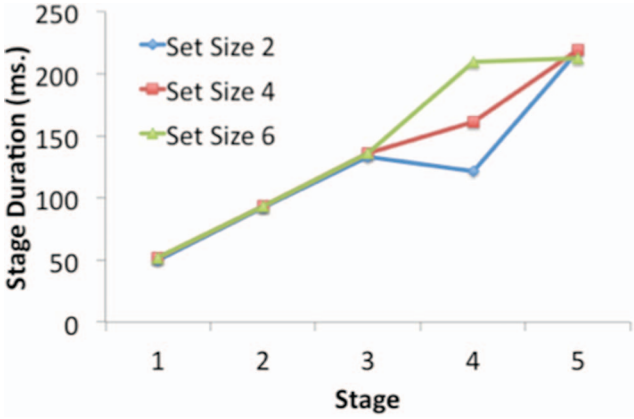

Figure 16. Mean stage durations for the Sternberg task as a function of set size. See the online article for the color version of this figure.

and recorded in the two different laboratories. Second, the final stage is about 80 ms longer in the Sternberg experiment than in the fan experiment (217 ms vs. 134 ms). The ACT-R model predicted a brief response stage of 110 ms for the fan experiment because participants could prepare the response hand, but not the response finger in advance. In contrast, in the Sternberg experiment, participants responded with both hands. This means that they could not prepare the response hand or finger in advance. Therefore, the fastest prepared response in this experiment would be 160 ms according to the ACT-R theory. The actual estimated response stage times are slightly longer in both tasks, suggesting that subjects were not always fully prepared.

Figure 17 presents the effects of probe type and set size on activity at the Fz and Pz electrodes. Parts (a) and (b) present conventional averages comparable to Figure 2a while Parts (c) and (d) present the bump-warped data. Conventional averaging rapidly obscures the structure of the trial as one moves from stimulus onset and temporal variability increases. Focusing on the bump-warped representation, we see that the third bump shows greater negativity at the Fz electrodes in the case of foils. Based on our interpretation of that bump as reflecting the N2, this suggests that participants experienced somewhat greater response conflict before rejecting probes as foils. Consistent with this interpretation, participants showed slightly, albeit not significantly, longer response times to foils. The fourth bump shows greater positivity at the Pz electrode for targets, typical of the parietal-old new effect. This supports our interpretations of the last bump in the fan experiment. Figure 17d shows bump-warped waveforms at these electrodes for different set sizes. The fourth bump at the Pz electrode decreases with set size, mirroring the fan effect we found in the corresponding fifth bump of the fan experiment.

Stage-Locked Modulations in Theta Activity

The original Jacobs et al. (2006) report focused on theta activity over the left parietal region and found that it was correlated with how well an item was remembered. Figure 18 shows theta power at one of the electrodes analyzed in their article (the left parietal electrode 53, which is near P3 in the 10–20 system). As in Jacobs et al., we calculated and *z*-transformed theta power for each subject. We then warped theta activity in each trial based on the maximum likelihood locations of the bumps in that trial. Consistent with their report, there is an effect of both set size (Figure 18a) and probe type (Figure 18b). The circles in Figure 18 mark the positions of the four bumps in each condition. We performed an ANOVA looking at how theta activity varied with set size and position in Figure 18a, and with target–foil and position in Figure

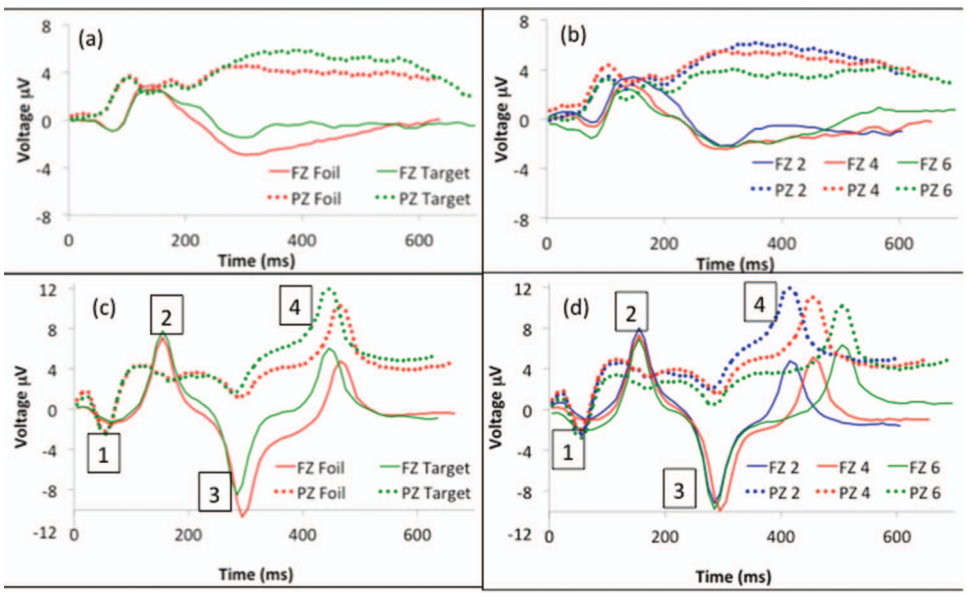

Figure 17. Average Fz and Pz activity in the Sternberg experiment as a function of probe type (a) and set size (b). Fz and Pz activity after warping every trial so that the maximum likelihood locations of the bumps correspond to the average locations for that condition of probe type (c) and set size (d). The bumps are numbered in (c) and (d). Fz = frontal; Pz = parietal. See the online article for the color version of this figure.

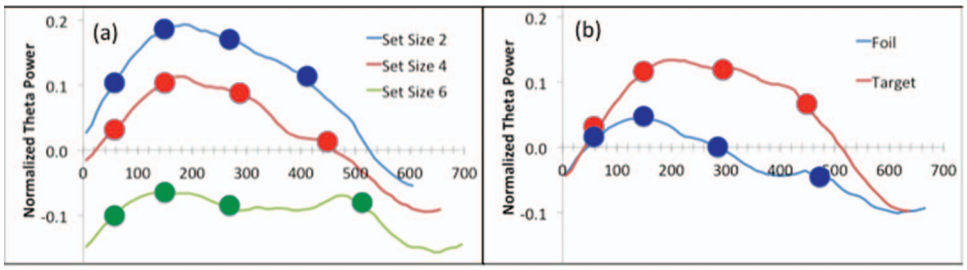

Figure 18. Normalized theta power at electrode P3 in the Sternberg experiment after warping every trial so that the maximum likelihood locations of the bumps correspond to the average locations for that condition. (a) Effect of set size. (b) Effect of target versus foil. Locations of the bumps in each condition are indicated by circles. See the online article for the color version of this figure.

18b. We looked at the six points defined by the beginning of the trial, the four bumps, and the end of the trial.

Set size. The effects of set size, $F(2, 34) = 22.55$, $p < .0001$ and position, $F(5, 85) = 5.21$, $p < .0005$, were significant, as was their interaction, $F(10, 170) = 2.58$, $p < .01$. Focusing on the interaction, there was no change in the effect of set size before Bump 4, $F(8, 34) = 1.72$, $p > .05$, after which differences among set sizes decreased from Bump 4 to the end, $F(2, 34) = 8.56$, $p < .001$. Thus, it appears that the effect of set size diminished once the decision was made and response preparation began.

Target-foil. The main effect of probe type was not significant, $F(1, 17) = 2.31$, $p > .1$, but the effect of position, $F(5, 85) = 5.21$, $p < .0005$, and the interaction between probe type and condition were significant, $F(5, 85) = 2.98$, $p < .05$. As seen in Figure 18, the effect of probe type was absent from the beginning point to Bump 1, and again at the end point, $F(1, 17) = 0.01$, $p > .1$. From Bump 1 through Bump 4, there was a significant effect of target–foil, $F(1, 17) = 7.59$, $p < .05$, which did not interact with position, $F(2, 34) = 0.91$, $p > .1$. The early absence of a target–foil effect is sensible given that participants did not begin knowing the type of probe. The theta-related effects of target–foil appeared when retrieval started with Bump 2 and ended when response preparation began with Bump 4.

In summary, the method has generalized robustly to another data set. Despite the differences in the participant populations, the paradigms, and the recording equipment, we recover similar bumps that show similar effects. The major unexpected result in the Sternberg experiment is the appearance of the set size effect in the fourth decision stage rather than the third retrieval stage. Given the short-term nature of the task, the small effect on the retrieval stage is perhaps not surprising (and it is estimated to be much briefer than in our fan experiment). We think the effect in the decision stage points to where the ACT-R theory needs further development. Currently, guided by fMRI results, we have simply assumed a 200-ms comparison process in all cases. However, fMRI data do not provide the temporal resolution needed to study the precise duration of the decision stage, and a constant comparison time seems somewhat unlikely. The data suggest that we should elaborate this decision stage, perhaps to involve a comparison as in the original Sternberg proposal. Finally, the stages appear to provide a meaningful interpretation of theta power modulation.

Discussion and Relationship to Other Analyses and Theories

We have shown in two experiments that our HSMM-bump analysis can identify trial-by-trial markers of cognitive processing stages in EEG data. Using this information, it is possible to identify whether and how specific stages are affected by different experimental factors. Our approach combines modern statistical and neural imaging techniques with Sternberg's additive factors logic. In doing so, it allows us to acquire latency measures for individual processing stages rather than just the cumulative duration of all stages.

We want to emphasize that the goal of this effort is to identify cognitive stages and not to fully explain the EEG signal. Our theory fails to completely explain the EEG signal on at least three scores. First, it does not address the generators of these bumps. Source localization techniques could be applied to the bumps in order to reveal likely neural generators. Second, our work with synthetic data in the Appendix suggests that the bumps we are recovering account for only about 5% of the variance in the EEG data. Of course, we account for much more of the traditional averaged signal with its ERP components. For instance, the synthetic data averaged in Figure A8 correlates about .8 with the corresponding averages of the actual data. Third, sustained activity cannot be modeled with bumps. In our own data set, an example of such sustained activity is the parietal old–new effect. Nonetheless, our approach identifies the boundaries (i.e., the bumps) directly preceding and following the activity in each trial. In this way, we can identify the processing stages where such sustained activity occurs.

In discussing the relationship to other research in the literature, we start with a comparison to something quite close—the HSMM modeling of Borst and Anderson (2015). We then turn to more general implications for understanding associative memory and the use of EEG for identifying cognitive processing stages.

Comparison With Borst and Anderson (2015)

Borst and Anderson (2015) directly applied the HSMM approach developed for fMRI to EEG. This approach identified *states* defined by constant patterns of activity (brain signatures, as in the fMRI analysis). Figure 19a shows the results of their analysis in terms of the voltage displays of the discovered states.

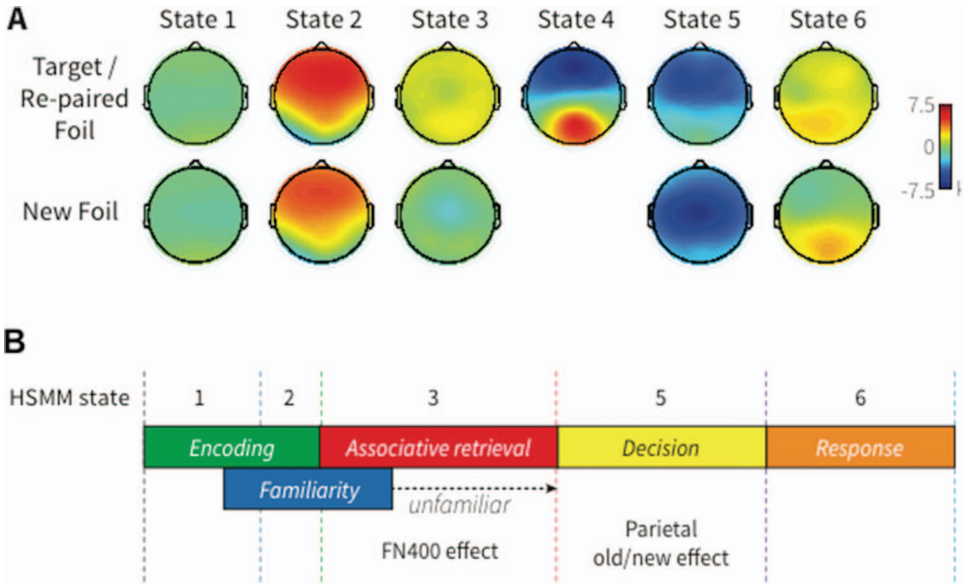

Figure 19. (a) The scalp profiles corresponding to the states identified in Borst and Anderson (2015). (b) The final model proposed by Borst and Anderson (2015). HSMM = hidden semi-Markov model. See the online article for the color version of this figure.

They found six states for targets and repaired foils and five for new foils. State 4 was quite brief and only observed with some frequency for 8 of 20 subjects. It was discounted in the final model. Ignoring that state, our five-bump HSMM gives a similar solution. The states in Borst and Anderson (2015) roughly map to the bumps and flats in the current analysis in the following way:

State 1 = Flat 1, Bump 1, first half of Flat 2

State 2 = second half of Flat 2, Bump 2, first half of Flat 3

State 3 = second half of Flat 3, Bump 3, first half of Flat 4

State 5 = second half of Flat 4, Bump 4, first half of Flat 5

State 6 = second half of Flat 5, Bump 5, Flat 6

The first three flats do not vary by condition, corresponding to Borst and Anderson's (2015) finding that the first two states also did not vary by condition. The fourth flat does vary by condition, corresponding to their finding that States 3 and 5 varied by condition. Finally, Borst and Anderson found a difference in the duration of State 6, corresponding to the differences we observed in Stage 5 and 6 when estimating separate durations for each condition (see Figure 11). However, as noted, performance was no worse in a LOOCV analysis when the last flat period was constrained to have a constant duration for all conditions.

Given these similarities, it is not surprising that Borst and Anderson (2015) proposed a final account of their data that is very similar to the current ACT-R model in Figure 10. Figure 19b shows their interpretation. As in the current model there are four basic stages and the processing of target and repaired foils is similar. To deal with new foils, Borst and Anderson (2015) proposed a familiarity process that starts in parallel with encoding the second word (State 2). This was implemented in the current ACT-R model as a slow retrieval of the meaning of the words. Thus, the speed of retrieval indicates familiarity—the faster the retrieval, the more familiar the word. This implementation is different from the classic familiarity stage proposed by dual-process theories, in which a continuous index of familiarity is given but no information is retrieved (e.g., Yonelinas, 2002). However, this is a natural way of implementing such a process in ACT-R. When words were deemed unfamiliar (in parallel to the associative retrieval of targets and repaired foils—State 3), Borst and Anderson's (2015) model proceeded directly to the decision stage, and the associative retrieval was skipped. This is exactly what happens in the current ACT-R model, where the associative retrieval only occurs for targets and repaired foils (compare Figure 10a to 10b). Finally, the decision and response stages are directly comparable, except that Borst and Anderson proposed that the duration of these stages was modulated by the difficulty of the decision. Given that the current analysis did not provide conclusive evidence for these duration differences, the more parsimonious explanation is that these stages' durations are constant.

The convergence in the outcomes of these two approaches supports the robustness of their conclusions. However, it does raise the question of whether one should conceive of cognitive stages as periods of constant activity, as in Borst and Anderson (2015), or as phasic bursts of EEG, as considered here. Statistically, discriminating between these accounts is difficult because both analyses imply that the other approach would find corresponding regularities. The advantages of the current approach are that it connects with existing conceptions of the EEG signal and that it promotes precision in the detection of the boundaries of cognitive events. Furthermore, it generalized robustly to the Sternberg task, and provides a theoretically grounded connection between the HSMM

method and ACT-R process models—a connection that was lacking in the analysis of Borst and Anderson.

Implications for Theories of Memory

Behavioral and model correlates. The ACT-R model accounts for the fan data in terms of a variant of a dual-process theory. As in a dual-process theory (e.g., Diana, Reder, Arndt, & Park, 2006; Yonelinas, 2002), the model makes judgments in two ways. New foils are rejected on the basis of the long latency in retrieving a representation of the word. This could be interpreted as familiarity-based judgment. Schooler and Hertwig (2005) also proposed an ACT-R model that judged familiarity in terms of speed of retrieval. Elsewhere, researchers have argued that latency is too variable to account for the accuracy of recognition memory (Conroy, Hopkins, & Squire, 2005; Poldrack & Logan, 1997). However, those studies involved single-word recognition (without the massive repetition of the words for targets and repaired foils that occurred in this experiment), where foils are typically rejected more slowly than targets. Our study involved significantly faster rejection of new foils. Additionally, the ACT-R model uses retrieval time, rather than total response time, to make judgments. We calculated how discriminable new foils would be from Fan 1 targets (the briefest of the other probes) according to times from the ACT-R model with gamma-2 distributions. The discriminability in total response time (d') was .94, but the discriminability in retrieval time was 2.62.

Many behavioral studies of associative recognition have found evidence consistent with new foils being rapidly rejected on the basis of a familiarity-based mechanism and repaired foils being rejected on the basis of recollection (e.g., Gronlund & Ratcliff, 1989; Rotello & Heit, 2000). The standard dual-process model for recollection (e.g., Hintzman & Curran, 1994) assumes a recall-to-reject process for foils in which the reject decision is based on a mismatch between the retrieved memory and the probe.

Electrophysiological correlates. ERP data, and particularly the FN400 and parietal old–new effect, have played a significant role in arguing for dual-process theories of memory (Rugg & Curran, 2007). The FN400 is thought to reflect familiarity or conceptual priming, whereas the parietal-old new effect is thought to reflect recollection. We did not observe an FN400 in the fan experiment. The absence of a difference between targets and repaired foils might relate to the facts that (a) the individual words that made up each pair had been studied and were thus highly familiar, (b) both the targets and the repaired foils repeated throughout the test phase, and (c) recognition memory for nonunitized word pairs might depend on recollection rather than familiarity (Bader et al., 2010).

A fourth bump appeared for new foils at the time when the FN400 was expected. This bump produced a more negative signal for new foils in Figure 2. However, as seen in Figure 12, the same negative response occurred in the other conditions as well, just after a substantial delay. For targets and repaired foils, the position of the fourth bump varied more from trial to trial. Greater variability in the latency of the bump in these conditions shifted and reduced the peak of the negativity in the conventional averaging of Figure 2.

We interpret the fourth bump as the N2 component of the ERP. According to one influential theory, the N2 arises from the anterior cingulate cortex and tracks response conflict (Yeung, Botvinick, & Cohen, 2004). In line with that theory, the fourth bump occurred at the onset of the decision stage and before a response was selected in the fan experiment. The same bump appeared in the Sternberg experiment where familiarity-based processing would not be expected to contribute to task performance because the consonants used as stimuli were repeatedly used through all memory sets in the experiment. Interestingly, the bump appeared at the onset of the decision phase in the model of that task as well. The N2 has traditionally been studied in 2AFC tasks. The implication of our results is that the N2 may occur more generally in tasks where people must decide and respond. However, the N2 may be obscured owing to latency variability in standard averaging methods.

We did observe a parietal old–new effect (see Figure 2). At the time of the parietal old–new effect, the ACT-R model is processing the memory trace that was just retrieved. We suggest that the effect is a sustained response related to the different activations of the memories in the various conditions. The mean voltage at Pz during this period is strongly correlated ($r = .896$) with model-based activation values (1.00 for Fan 1 targets, 0.30 for Fan 2 targets, 0.50 for Fan 1 repaired foils, 0.15 for Fan 2 repaired foils, and 0.00 for new foils). In the Sternberg task, the portion of the EEG signal corresponding to the parietal old–new effect also showed an effect of set size consistent with the proposal that activation decreases with set size.

fMRI studies of memory processing consistently reveal a retrieval-success effect in the lateral parietal cortex; activation is greater for items that are successfully retrieved than for items that are not (Gilmore, Nelson, & McDermott, 2015; Vilberg & Rugg, 2008; Wagner, Shannon, Kahn, & Buckner, 2005). The topography of the parietal old–new effect seen in the EEG signal is consistent with a source in the lateral parietal cortex. Additionally, experiment manipulations that produce a parietal old–new effect also yield greater activation in lateral parietal cortex, as revealed by fMRI studies (Vilberg & Rugg, 2008). The convergence of results suggests that the scalp-recorded parietal old–new effect originates from a source in the lateral parietal cortex.

Several theories have been proposed to account for the role of lateral parietal regions in memory processing, three of which are relevant here. The *output buffer hypothesis* proposes that parietal regions transiently maintain retrieved information in a form accessible to decision-making processes (Wagner et al., 2005). A somewhat related idea is that the parietal cortex itself does not maintain retrieved information, but that it focuses attention on the contents of working memory stored elsewhere (Vilberg & Rugg, 2008). The *accumulator hypothesis* proposes that the level of parietal activation corresponds to evidence accumulated toward an eventual "old" response (Wagner et al., 2005). Lastly, the *bottom-up attention hypothesis* generalizes the idea of "attention-grabbing" from external perceptual stimuli to information retrieved internally from memory (Cabeza, Ciaramelli, & Moscovitch, 2012).

In our model, the time period corresponding to the parietal old–new effect comes after retrieval, when retrieved information is compared to the probe. The postretrieval locus of the effect is consistent with all three theories. Our analysis indicates that the parietal old–new effect is a sustained difference rather than a transient change. If it were a transient change, it would appear as an isolated bump rather than an elevated flat. This outcome is consistent with the first and second theories, but not the third. We find effects of fan and set size on the amplitude of the parietal old–new effect. This could be consistent with the output buffer

hypothesis if one assumed that more information was retrieved and subsequently maintained for items with higher activation. This result could also be consistent with the accumulator hypothesis if one assumed that during the postretrieval decision process, items with higher activation had a higher accumulation rate.

The possibility that parietal activation—and the parietal old–new effect—reflects evidence accumulation is reminiscent of a recent account of the P300. Twomey, Murphy, Kelly, & O'Connell (2015) proposed that the P300 encodes a dynamic decision variable. Upon reaching a critical threshold, a response occurs. They conducted an auditory oddball detection task with varying levels of discrimination difficulty. The P300 rose most quickly when discrimination difficulty was low, but rose to the same maximum in all conditions as revealed by response-locked waveforms. Twomey et al. proposed that the same account could be applied to other late positive potentials, such as the centroparietal positivity and the parietal old–new effect.

Use of EEG for Identification of Cognitive Stages

Single-trial EEG signals may seem hopelessly noisy. This has motivated the typical practice in EEG research of averaging data across many trials. The method described in this article takes advantage of the statistical power of combining large numbers of trials without abandoning the structure of individual trials. Rather than computing an average signal from all trials, our method estimates the parameters of the process that drives this signal. The critical process is the appearance of phasic bumps that mark changes in cognitive processing. Estimating the magnitudes and locations of bumps allows us to return to individual trials in order to interpret the variable time course of cognitive events (see Figure 8). Averaging the data according to the maximum likelihood locations of the critical events (Figures 12 and 17) reveals regularities that are lost by other methods of averaging.

This method relies on the power of the dynamic programming techniques underlying HSMMs to search through all the ways that a trial could be parsed and to combine the likelihood of each. The method tames the combinatorial problem of considering all possible interpretations of a trial. However, it requires that stages be identified and that one specify how the durations of stages vary by condition. The space of such HSMMs is itself a combinatorial challenge, and brute force explorations (Figures 5 and 6) can only take us so far. As Sternberg recognized in his additive factors method, one needs a theoretical framework to guide postulation of the stages and to identify factors that might affect the duration of each. The general class of dual-process, retrieve-to-reject models served that role and pointed to a better five-bump HSMM for the fan experiment, despite the initial evidence suggesting a six-bump HSMM.

To make rigorous connection between such a framework and the data requires a parameterized model and a precise linking assumption. ACT-R and the bump hypothesis provide such a model and link. Production rule firing drives changes in processing in ACT-R and the appearance of bumps in the corresponding electrophysiological model. An HSMM based on the existing ACT-R retrieve-to-reject model outperformed any of the HSMMs discovered using a purely bottom-up search of the data. The ACT-R model also indicated that new foils should have one fewer bumps than all other probe types, explaining their weaker P2 in the original analysis (see Figure 2).

Historically, the term *event-related potential* has referred to the synchronization of neural activity with overt experiment events, such as the presentation of a stimulus or the commission of a response. This conception is fundamentally limited insofar as significant cognitive events occur with variable latencies. The method developed in this article aligns the EEG signal with the onset of multidimensional bumps. In this way, "events" take on richer meaning as latent changes in cognitive processing. The method developed in this article provides a path for any theory that proposes discrete changes in task-related information processing. Like the well-practiced additive factors method, this method can test whether factors affect specific stages in the ways predicted by a theory. Further, this method can identify the actual durations of each stage rather than just the differences in durations between conditions.

References

Almeida, D., & Poeppel, D. (2013). Word-specific repetition effects revealed by MEG and the implications for lexical access. *Brain and Language, 127,* 497–509. http://dx.doi.org/10.1016/j.bandl.2013.09.013

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195324259.001.0001

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38,* 341–380. http://dx.doi.org/10.1006/jmla.1997.2553

Anderson, J. R., Carter, C. S., Fincham, J. M., Qin, Y., Ravizza, S. M., & Rosenberg-Lee, M. (2008). Using FMRI to test models of complex cognition. *Cognitive Science, 32,* 1323–1348. http://dx.doi.org/10.1080/03640210802451588

Anderson, J. R., & Fincham, J. M. (2014a). Discovering the sequential structure of thought. *Cognitive Science, 38,* 322–352. http://dx.doi.org/10.1111/cogs.12068

Anderson, J. R., & Fincham, J. M. (2014b). Extending problem-solving procedures through reflection. *Cognitive Psychology, 74,* 1–34. http://dx.doi.org/10.1016/j.cogpsych.2014.06.002

Anderson, J. R., Fincham, J. M., Schneider, D. W., & Yang, J. (2012). Using brain imaging to track problem solving in a complex state space. *NeuroImage, 60,* 633–643. http://dx.doi.org/10.1016/j.neuroimage.2011.12.025

Anderson, J. R., Lee, H. S., & Fincham, J. M. (2014). Discovering the structure of mathematical problem solving. *NeuroImage, 97,* 163–177. http://dx.doi.org/10.1016/j.neuroimage.2014.04.031

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128,* 186–197. http://dx.doi.org/10.1037/0096-3445.128.2.186

Assecondi, S., Bianchi, A. M., Hallez, H., Staelens, S., Casarotto, S., Lemahieu, I., & Chiarenza, G. A. (2009). Automated identification of ERP peaks through dynamic time warping: An application to developmental dyslexia. *Clinical Neurophysiology, 120,* 1819–1827. http://dx.doi.org/10.1016/j.clinph.2009.06.023

Bader, R., Mecklinger, A., Hoppstädter, M., & Meyer, P. (2010). Recognition memory for one-trial-unitized word pairs: Evidence from event-related potentials. *NeuroImage, 50,* 772–781. http://dx.doi.org/10.1016/j.neuroimage.2009.12.100

Basar, E. (1980). *EEG brain dynamics: Relation between EEG and brain evoked potentials*. Amsterdam, The Netherlands: Elsevier.

Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference, 112,* 241–258. http://dx.doi.org/10.1016/S0378-3758(02)00336-1

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K. R. (2011). Single-trial analysis and classification of ERP components: A tutorial. *NeuroImage, 56,* 814–825. http://dx.doi.org/10.1016/j.neuroimage.2010.06.048

Boring, E. G. (1929). *A history of experimental psychology*. New York, NY: Century.

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage, 108,* 60–73. http://dx.doi.org/10.1016/j.neuroimage.2014.12.029

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. R. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS ONE, 10,* e0119673. http://dx.doi.org/10.1371/journal.pone.0119673

Borst, J. P., Schneider, D. W., Walsh, M. M., & Anderson, J. R. (2013). Stages of processing in associative recognition: Evidence from behavior, EEG, and classification. *Journal of Cognitive Neuroscience, 25,* 2151–2166. http://dx.doi.org/10.1162/jocn_a_00457

Borst, J. P., Taatgen, N. A., Stocco, A., & van Rijn, H. (2010). The neural correlates of problem states: Testing FMRI predictions of a computational model of multitasking. *PLoS ONE, 5,* e12966. http://dx.doi.org/10.1371/journal.pone.0012966

Cabeza, R., Ciaramelli, E., & Moscovitch, M. (2012). Cognitive contributions of the ventral parietal cortex: An integrative theoretical account. *Trends in Cognitive Sciences, 16,* 338–352. http://dx.doi.org/10.1016/j.tics.2012.04.008

Conroy, M. A., Hopkins, R. O., & Squire, L. R. (2005). On the contribution of perceptual fluency and priming to recognition memory. *Cognitive, Affective & Behavioral Neuroscience, 5,* 14–20. http://dx.doi.org/10.3758/CABN.5.1.14

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28,* 923–938. http://dx.doi.org/10.3758/BF03209340

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084,* 89–103. http://dx.doi.org/10.1016/j.brainres.2006.02.010

Danker, J. F., Gunn, P., & Anderson, J. R. (2008). A rational account of memory predicts left prefrontal activation during controlled retrieval. *Cerebral Cortex, 18,* 2674–2685. http://dx.doi.org/10.1093/cercor/bhn027

D'Avanzo, C., Schiff, S., Amodio, P., & Sparacino, G. (2011). A Bayesian method to estimate single-trial event-related potentials with application to the study of the P300 variability. *Journal of Neuroscience Methods, 198,* 114–124. http://dx.doi.org/10.1016/j.jneumeth.2011.03.010

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134,* 9–21. http://dx.doi.org/10.1016/j.jneumeth.2003.10.009

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological, 39,* 1–38.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review, 13,* 1–21. http://dx.doi.org/10.3758/BF03193807

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica, 30,* 412–431. http://dx.doi.org/10.1016/0001-6918(69)90065-1

Düzel, E., Yonelinas, A. P., Mangun, G. R., Heinze, H.-J., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America, 94,* 5973–5978. http://dx.doi.org/10.1073/pnas.94.11.5973

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (Eds.). (2007). *Statistical parametric mapping: The Analysis of functional brain images*. London, United Kingdom: Academic Press.

Gilmore, A. W., Nelson, S. M., & McDermott, K. B. (2015). A parietal memory network revealed by multiple MRI methods. *Trends in Cognitive Sciences, 19,* 534–543. http://dx.doi.org/10.1016/j.tics.2015.07.004

Gratton, G., Kramer, A. F., Coles, M. G., & Donchin, E. (1989). Simulation studies of latency measures of components of the event-related brain potential. *Psychophysiology, 26,* 233–248. http://dx.doi.org/10.1111/j.1469-8986.1989.tb03161.x

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858. http://dx.doi.org/10.1037/0278-7393.15.5.846

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33,* 1–18. http://dx.doi.org/10.1006/jmla.1994.1001

Jacobs, J., Hwang, G., Curran, T., & Kahana, M. J. (2006). EEG oscillations and recognition memory: Theta correlates of memory retrieval and decision making. *NeuroImage, 32,* 978–987. http://dx.doi.org/10.1016/j.neuroimage.2006.02.018

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine, 30,* 3050–3056. http://dx.doi.org/10.1002/sim.4323

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2001). Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping, 14,* 166–185. http://dx.doi.org/10.1002/hbm.1050

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences, 18,* 203–210. http://dx.doi.org/10.1016/j.tics.2014.01.002

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences, 4,* 432–440. http://dx.doi.org/10.1016/S1364-6613(00)01545-X

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science, 295,* 690–694. http://dx.doi.org/10.1126/science.1066168

Mäkinen, V., Tiitinen, H., & May, P. (2005). Auditory event-related responses are generated independently of ongoing brain activity. *NeuroImage, 24,* 961–968. http://dx.doi.org/10.1016/j.neuroimage.2004.10.020

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330. http://dx.doi.org/10.1037/0033-295X.86.4.287

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104,* 3–65. http://dx.doi.org/10.1037/0033-295X.104.1.3

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320,* 1191–1195. http://dx.doi.org/10.1126/science.1152876

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10,* 424–430. http://dx.doi.org/10.1016/j.tics.2006.07.005

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage, 45,* S199–S209. http://dx.doi.org/10.1016/j.neuroimage.2008.11.007

Poldrack, R. A., & Logan, G. D. (1997). Fluency and response speed in recognition judgments. *Memory & Cognition, 25,* 1–10. http://dx.doi.org/10.3758/BF03197280

Poli, R., Cinel, C., Citi, L., & Sepulveda, F. (2010). Reaction-time binning: A simple method for increasing the resolving power of ERP averages. *Psychophysiology, 47,* 467–485. http://dx.doi.org/10.1111/j.1469-8986.2009.00959.x

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77,* 257–286. http://dx.doi.org/10.1109/5.18626

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 6539–6544. http://dx.doi.org/10.1073/pnas.0812589106

Rhodes, S. M., & Donaldson, D. I. (2008). Electrophysiological evidence for the effect of interactive imagery on episodic memory: Encouraging familiarity for non-unitized stimuli during associative recognition. *NeuroImage, 39,* 873–884. http://dx.doi.org/10.1016/j.neuroimage.2007.08.041

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience, 14,* 611–653.

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28,* 907–922. http://dx.doi.org/10.3758/BF03209339

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11,* 251–257. http://dx.doi.org/10.1016/j.tics.2007.04.004

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology, 64,* 127–160. http://dx.doi.org/10.1016/j.cogpsych.2011.11.001

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–628. http://dx.doi.org/10.1037/0033-295X.112.3.610

Schweickert, R., Fisher, D. L., & Goldstein, W. M. (2010). Additive factors and stages of mental processes in task networks. *Journal of Mathematical Psychology, 54,* 405–414. http://dx.doi.org/10.1016/j.jmp.2010.06.004

Shah, A. S., Bressler, S. L., Knuth, K. H., Ding, M., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2004). Neural dynamics and the fundamental mechanisms of event-related brain potentials. *Cerebral Cortex, 14,* 476–483. http://dx.doi.org/10.1093/cercor/bhh009

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284. http://dx.doi.org/10.1080/03640210802414826

Sohn, M. H., Goode, A., Stenger, V. A., Carter, C. S., & Anderson, J. R. (2003). Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 100,* 7412–7417. http://dx.doi.org/10.1073/pnas.0832374100

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654. http://dx.doi.org/10.1126/science.153.3736.652

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30,* 276–315. http://dx.doi.org/10.1016/0001-6918(69)90055-9

Sternberg, S. (2011). Modular processes in mind and brain. *Cognitive Neuropsychology, 28,* 156–208. http://dx.doi.org/10.1080/02643294.2011.557231

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage, 62,* 451–463. http://dx.doi.org/10.1016/j.neuroimage.2012.04.048

Tuan, P. D., Möcks, J., Köhler, W., & Gasser, T. (1987). Variable latencies of noisy signals: Estimation and testing in brain potential data. *Biometrika, 74,* 525–533. http://dx.doi.org/10.1093/biomet/74.3.525

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *The European Journal of Neuroscience, 42,* 1636–1643. http://dx.doi.org/10.1111/ejn.12936

Van Petten, C. V., Kutas, M., Kluender, R., Mitchiner, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience, 3,* 131–150. http://dx.doi.org/10.1162/jocn.1991.3.2.131

Vilberg, K. L., & Rugg, M. D. (2008). Memory retrieval and the parietal cortex: A review of evidence from a dual-process perspective. *Neuropsychologia, 46,* 1787–1799. http://dx.doi.org/10.1016/j.neuropsychologia.2008.01.004

Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences, 9,* 445–453. http://dx.doi.org/10.1016/j.tics.2005.07.001

Wang, K., Begleiter, H., & Porjesz, B. (2001). Warp-averaging event-related potentials. *Clinical Neurophysiology, 112,* 1917–1924. http://dx.doi.org/10.1016/S1388-2457(01)00640-X

Wiegand, I., Bader, R., & Mecklinger, A. (2010). Multiple ways to the prior occurrence of an event: An electrophysiological dissociation of experimental and conceptually driven familiarity in recognition memory. *Brain Research, 1360,* 106–118. http://dx.doi.org/10.1016/j.brainres.2010.08.089

Woody, C. D. (1967). Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. *Medical & Biological Engineering, 5,* 539–554. http://dx.doi.org/10.1007/BF02474247

Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology, 41,* 822–832. http://dx.doi.org/10.1111/j.1469-8986.2004.00239.x

Yeung, N., Bogacz, R., Holroyd, C. B., Nieuwenhuis, S., & Cohen, J. D. (2007). Theta phase resetting and the error-related negativity. *Psychophysiology, 44,* 39–49. http://dx.doi.org/10.1111/j.1469-8986.2006.00482.x

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959. http://dx.doi.org/10.1037/0033-295X.111.4.931

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. http://dx.doi.org/10.1006/jmla.2002.2864

Yu, S. Z. (2010). Hidden semi-Markov models. *Artificial Intelligence, 174,* 215–243. http://dx.doi.org/10.1016/j.artint.2009.11.011

(*Appendix follows*)

Appendix

Definition of Likelihood and Robustness of Parameter Estimation

Defining Likelihood

Our data are the sequences of 10 PCA values for every 10-ms sample for every trial. An *n*-bump HSMM calculates probability of these data summing over all ways of placing the *n* bumps to break the trial into $n + 1$ flats. Except for the first stage, each stage is taken as the duration of the bump that initiates it and the following flat. Bumps are always five samples but the flats have variable durations.[16] The probability of a flat of length *t* is calculated as the normalized[17] density at point $t + .5$ in a gamma distribution with shape parameter *2* and scale parameter *b*, denoted $g_2(t,b)$.

The durations t_i of the $n + 1$ flats for a trial must satisfy the constraint $t_1 + 5 + t_2 + 5 + \ldots + 5 + t_{n+1} = T$, where *T* is the total number of the samples in the trial. The probability of any placement of bumps that satisfies this constraint is the probability of the flat durations times the probability of the signal given the bump locations:

$$P(t_i, t_2, \ldots t_{n+1}) = \left(\prod_{k=1}^{n+1} g_2(t_k, b_k)\right)\left(\prod_{d=1}^{m} \Pr(Samples_d)\right).$$

The first product concerns the probability of the $n + 1$ flat durations. The second product concerns the probability of the PCA values for the trial on the *m* dimensions (in our case $m = 10$). Because the dimensions are uncorrelated (and assumed orthogonal) we treat this as a product of the probabilities of the *T* samples for each dimension *d*.

We will now focus on the probability of the samples on a single dimension for a particular placement of the bumps. This probability is determined by the values during the flats and the values during the bumps. There is a strong correlation between adjacent samples, which rapidly drops of with distance (Figure A1). Reflecting this drop off, we treat the samples in one bump as independent from any other bump, the samples in any flat as independent from any other flat, and the samples in any flat as independent from the samples in any bump. This approximation leaves us to deal with the local sequential correlations within flats and bumps. This gives the following expression for the probability of the samples on a PCA dimension:

$$\Pr(Samples) = \left(\prod_{k=1}^{n+1} \Pr(Flat_k)\right)\left(\prod_{k=1}^{n} \Pr(Bump_k)\right)$$

where $Flat_k$ are the samples in the *kth* flat and $Bump_k$ are the samples in the *kth* bump. Assuming that the log-likelihood is linear with the squared difference between the observed values, we have

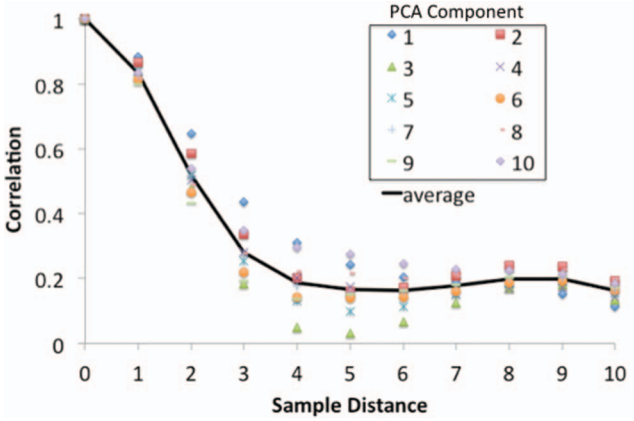

Figure A1. The correlation between samples at different distances for the 10 principal component analysis (PCA) components. See the online article for the color version of this figure.

$$In(\Pr(Samples)) = K - \frac{\left[\sum_{k=1}^{n+1}\sum_{j=1}^{t_k} S_{kj}^2 + \sum_{k=1}^{n}\sum_{j=1}^{5}(S_{kj} - B_{kj})^2\right]}{V},$$

where the first sum of squares are the deviations of the values S_{kj} in the $n + 1$ flats (with t_k being the duration of the *kth* flat) from their expected values of zero[18] and the second term squares the deviation of the values S_{kj} in the bumps from their expected values B_{kj}. The other two parameters are *K*, a normalizing constant, and *V*, a measure of variability addressed below. Because the PCAs have variance 1, this expression can be rewritten as

$$In(\Pr(Samples)) = K - \frac{T}{V} + \frac{\left[\sum_{k=1}^{n}\sum_{j=1}^{5}(S_{kj}^2 - (S_{kj} - B_{kj})^2)\right]}{V},$$

where *T* is the summed variance of the *T* samples. Log probability will be maximized by maximizing the last term in this expression, which is a measure of how much the bumps reduce the variance at the locations where they are placed.

[16] In the exposition to follow, the number of samples can be zero. In the actual implementation it is bounded below by 1, but the last sample of a flat is the first sample of the following bump.

[17] Normalized so that the sum of probabilities of samples from 0 to the maximum length of a trial is 1.

[18] To the extent that a flat has a nonzero mean this analysis ignores systematic variability. This is the case in the period of the parietal old–new response.

(*Appendix continues*)

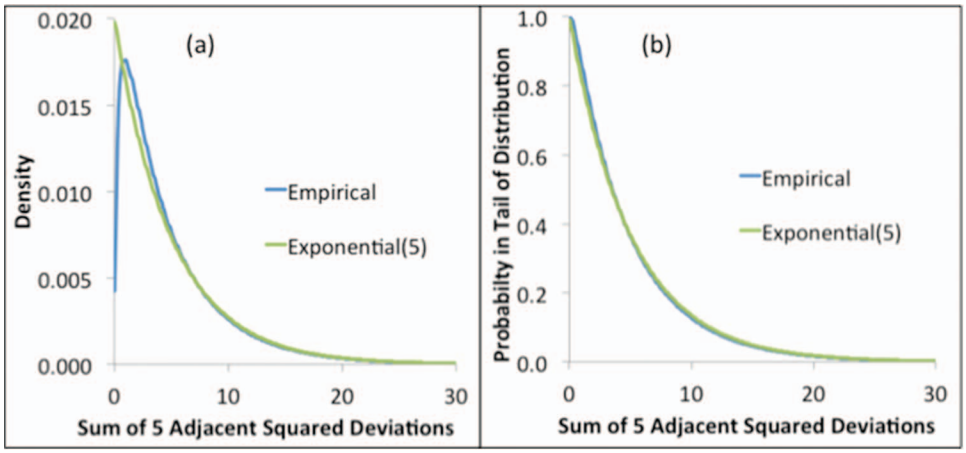

Figure A2. (a) Comparisons of the densities for sum of squared deviations of five adjacent samples from the grand mean of 0 and exponential-5 probability distribution. (b) Survival distributions for the sum of squared deviations and exponential-5. See the online article for the color version of this figure.

To combine this measure of signal variability with the gamma probabilities of the flat durations we need to determine the divisor V in order to have the defined probability of the samples being a density function. A convenient way of doing this is to see if the empirical distribution of the sums of five adjacent PCA squared signals can be approximated to a known parameterized distribution. Figure A2a shows the distribution of sums of five adjacent PCA squared signals. As Figure A2a reveals, the distribution is approximated by an exponential distribution with scale parameter 5 (the correlation with the exponential is .966). The source of deviation from an exponential-5 is that the empirical distribution has few instances of very small sums.[19] The approximation is close as shown in Figure A2b, which compares the survivor functions for the sum of squares and the exponential distribution. These survivor functions reflect how much extreme deviations from prediction are penalized in estimating log-likelihood. Treating the distribution of sums of 5 deviations as an exponential-5 is equivalent to setting V to 5.

Robustness of Parameter Estimation

The article identifies a five-bump HSMM with a particular set of bump profiles and flat durations. We used both nonparametric and parametric bootstrapping to explore how likely the same conclusions would be given another sample of data. With the parametric bootstrapping methods, we were able to also explore whether our conclusions be affected by different potential complications in the data. Here, we report a summary of our explorations but a more thorough report is available at http://act-r.psy.cmu.edu/?post_type=publications&p=17655.

Nonparametric Bootstrapping

We created 100 new data sets from the original data set by resampling with replacement—thus, a particular trial can appear zero, one, or more times in the resampled data. We did this either with the constraint that the number of trials for each of the 20 subjects was fixed, or that the number of trials for each of the five conditions was fixed. In either case, the 95% confidence intervals on the flat durations were not more than 50 ms for any flat. The bump profiles were similarly recovered with high precision. This reflects the fact that we have more than ample trials for stable estimation.

Parametric Bootstrapping

While nonparametric bootstrapping has the virtue of working with the real trials, it is limited in that it cannot explore temporal variability in the samples within a trial. It can only take the trials that exist and create new mixtures of whole trials. To address within-trial variability and explore other issues, we generated synthetic data. This involves generating trials by superimposing phasic peaks of signal (bumps) at points determined by the gamma-2 distributions for the flats on noise characterized by the same power spectrum for the human EEG used in the simulations of Yeung et al. (2007).

[19] The best fitting gamma distribution is a gamma with shape 1.2 and scale 4.1.

(*Appendix continues*)

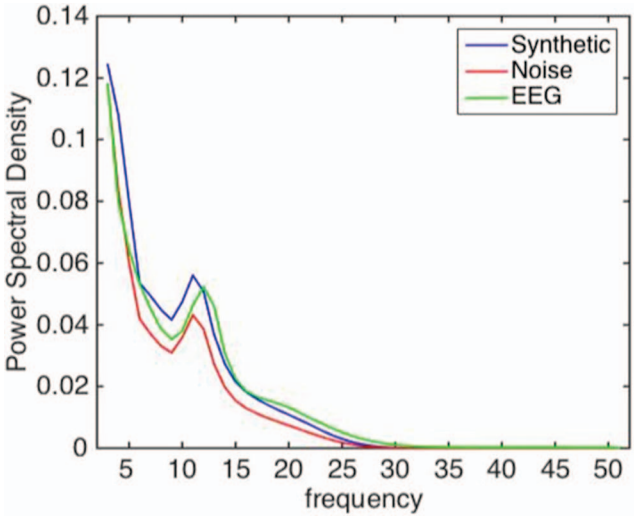

Figure A3. Power spectrum of actual electroencephalographic (EEG) data, the Yeung et al. (2007) noise generator, and the synthetic data. See the online article for the color version of this figure.

Exploration of a range of signal-to-noise ratios (SNRs) indicated that a SNR of .05 (−13.01 dB) results in estimating HSMMs from the synthetic data with log-likelihoods similar to those of the real data. Therefore, this is the value used. Figure A3 compares the power spectrum of the noise, the synthetic data with the model-specified bumps added, and the power spectrum of our EEG data. Consistent with the nonparametric bootstrapping, the duration of the flats and the magnitudes of the bumps were recovered with high precision from the synthetic data. In addition, we were also able to explore a number of issues about the robustness of the estimation procedure as discussed below.

Number of bumps. To explore our ability to identify the number of bumps, we took the parameters from the data when we estimated one to eight bumps (see Figure 5) and generated synthetic data with these parameters. As with the actual data we performed LOOCV on 20 synthetic subjects and looked at the mean log-likelihood of the left-out subject. Each panel in Figure A4 shows fits to data generated with a particular number of bumps. Within the panel we see the result of fitting that data with different numbers of bumps. In each case, the likelihood was maximal for the number of bumps from which the data was generated. In all cases, at least 19 of the 20 synthetic subjects were best fit in LOOCV with the correct number of bumps.

Shape parameter. In the main section of the article we assumed that the gamma distribution for the flat durations had a shape parameter of 2. We explored the question of what shape parameter would give the best maximum-likelihood fit to the data. The answer varied with the number of bumps we assumed. We focused on five-bump and six-bump models because they were most likely for our data. The best fitting shape parameter for a five-bump model was, in fact, 2, while the best fitting parameter

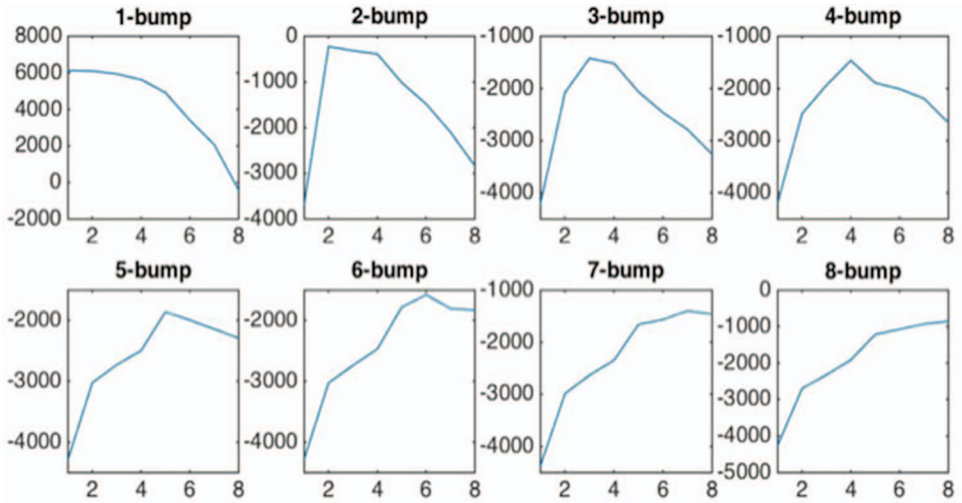

Figure A4. The panels are for synthetic data sets generated with one to eight bumps. The graphs within panels are the mean likelihood in leave-one-out cross-validation of 20 synthetic subjects fitted with different numbers of bumps. See the online article for the color version of this figure.

(*Appendix continues*)

for the six-bump model was 1.2. We simulated data where the gamma distributions for the flats had different shape parameters and determined what shape parameters would result in the best maximum likelihood fits to the synthetic data. While the best-fitting parameter increased with the generating parameter, they were not always the same. However, for a five-bump model, the best fitting parameter for synthetic data generated with a shape parameter of 2 was in fact 2, matching the real data. For a six-bump model, synthetic data generated with a true shape of 2 resulted in a best-fitting shape of 1.6, close to the 1.2 obtained with the real data.

Single-trial bump location. We investigated the accuracy with which we could recover the bump locations on a trial-by-trial basis using two ways of estimating location on a trial: maximum likelihood location and mean location. Using maximum likelihood, the root-mean-square deviations in the trial-by-trial locations for the five bumps were 54 ms, 57 ms, 69 ms, 142 ms, and 70 ms. As might be expected, using mean location resulted in somewhat smaller deviations: 49 ms, 53 ms, 62 ms, 122 ms, and 60 ms.

Using either estimation procedure, the accuracy of locating bumps on single trials is only modest given the SNR of .05. However, even with this much single trial uncertainty, average bump location and mean electrode activity can be estimated quite accurately. Figure A5 displays Bump 4 from the Fz electrode stimulus-locked for Fan 1 and Fan 2 targets. The 500- to 750-ms range is where the Fan 1 bump occurs and 750- to 1,000-ms range is where the Fan 2 bump occurs. These are reconstructed from the synthetic trial PCAs. The blue lines are the data warped using knowledge of where the bumps were on each trial, the red lines are based on the mean locations, and the green lines on the maximum likelihood. These inferred locations are representative of all the bumps in that there is at most one 10-ms-sample difference with the true locations. The mean estimates give a much wider distribution, which is why Figure 12 uses maximum likelihood locations. The maximum likelihood bump magnitudes average about 12% larger than the actual bump magnitudes, reflecting a small bias in the estimation process.

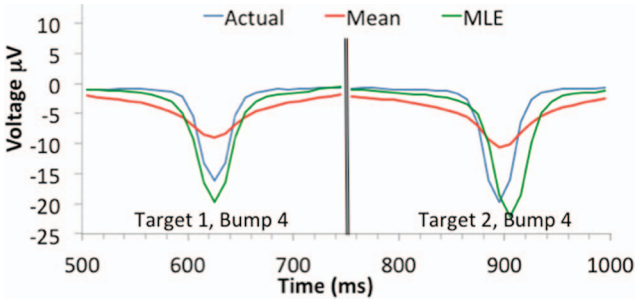

Figure A5. Bump 4 locations and estimates for synthetic data for Fan 1 and Fan 2 targets. MLE = maximum likelihood estimate. See the online article for the color version of this figure.

Some additional features of these results with simulated data are relevant for interpretation of the results in Figure 12 with the actual data. First, note that, as in the actual data, the bump appear to be larger for the longer Fan 2 condition, even for the actual signal. This is a result of the *z* scoring of each trial. The variability of the data is less on longer trials because the bumps contribute less to the signal. When these longer trials are rescaled to have a variance of 1, the bumps get amplified relative to shorter trials. Second, while the average bumps based on true location are 50 ms wide, the bumps based on maximum likelihood locations are 10 to 20 ms wider. This is because of uncertainty in their localization. However, the maximum-likelihood bumps in the real data (see Figure 12) are a further 10 to 20 ms wider than the maximum-likelihood bumps in the simulated data. This suggests that the actual bumps in the data might be wider than the 50 ms assumed.

Width of bumps. As noted above the actual bumps may be wider than our assumed 50 ms. To investigate the robustness of the model when the width assumption is not met, a number of synthetic datasets are generated with different underlying bump widths. We then applied the same model that assumes the bump width to be 50 ms and compared how accurately it could still recover the stage durations. Figure A6a gives examples of the first PCA component for synthetic trials generated with bumps of different widths before any noise is added. Then we added noise assuming our standard SNR of 0.05. The correct number of bumps can be recovered when the bump width ranges from 30 to 110 ms. An additional bump provided the best fit with 130 and 150 ms, basically fitting two bumps to cover the span of the fourth bump, which is widely separated from adjacent bumps. Figure A6b shows the estimated stage durations when a five-bump model, assuming 50 ms durations, is fit to the data when the true bumps are of different widths. As can be seen, the stage durations are recovered with quite high accuracy. In generating the data we also allowed the bumps to overlap. Despite the fact that the analysis does not allow for this (for purposes of computational tractability) we see that the true locations of the events still are recovered.

The effect of a period of sustained activity. The parietal old–new effect appears in our data as a period of sustained activation that spans the fifth stage between the fourth and fifth bumps. This contrasts with the assumption of our analysis that all flat periods have 0 mean activity. While it is nice that the analysis can identify the existence of such periods despite the assumption that they are not there, one can wonder whether the presence of such sustained periods might distort conclusions about the number of bumps and their timing.

(*Appendix continues*)

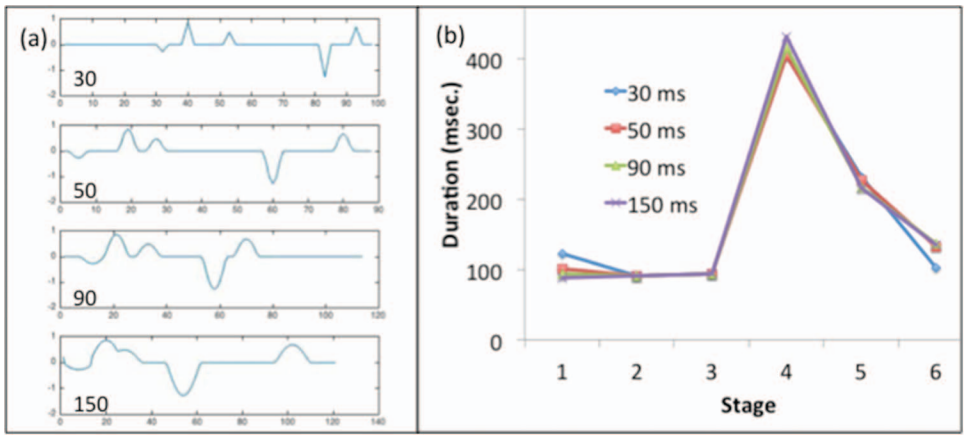

Figure A6. (a) Examples of synthetic trials with bumps of 30-, 50-, 90-, and 150-ms widths. (b) Estimates of stage durations given synthetic bumps of different widths. See the online article for the color version of this figure.

To explore this issue, we looked at how the conclusions would vary if the fifth flat did have a nonzero mean activity like we found for that flat. We generated synthetic data where the height of the fifth flat is varied from a ratio 0.1 to 0.9 of the fifth bump. A ratio of 0.3 was chosen because it best matched the flat obtained from the original EEG dataset. Figure A7 shows the first simulated PCA component warped around the location of the bumps like Figure 12. The locations of the bumps are quite similar with or without a nonzero flat.

Mapping the synthetic data back to the electrodes. Our work with synthetic data has involved generation of the PCA components. How does this relate to electrode activity, as it is traditionally presented, response-locked or stimulus-locked (e.g., Figure 2)? Given the coefficients obtained in the PCA, one can map the 10 synthetic PCA components back to electrode activity on each trial. Averaging the trials together will result in some loss of alignment but should correspond to these traditional presentations. Figure A8a shows the stimulus locked synthetic data mapped to electrode Fz (to be compared with Figure 2a) and Figure A8b shows the synthetic response-locked Pz activity (to be compared with Figure 2b). The correspondences between the real and synthetic electrodes, while not perfect, are quite apparent.

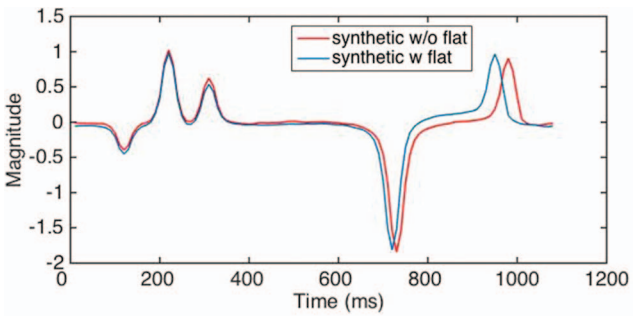

Figure A7. The time-warped first principal component analysis component with and without an elevated flat between the fourth and fifth bumps. See the online article for the color version of this figure.

The fourth bump. The existence of the fourth bump in all the conditions of the fan experiment is an interesting discovery of the HSMM-MVPA analysis. It is not apparent in either the stimulus-locked or response-locked traditional representations (see Figure 2). Rather, it lies buried in the middle of the trial obscured by variability in the stage durations. One can wonder whether this is real or some artifact of the analysis. In the long lag between the early positive second and third bumps and the late fifth bump, there will tend to be a period when the signal is most negative by chance, and perhaps Bump 4 is just capturing these random moments. To address this question, we generated data with only Bumps 1, 2, 3, and 5 and then fit a four-bump and a five-bump model. The four-bump model fits the synthetic data generated with four bumps slightly better than the five-bump model. Figure A9 shows the result of the five-bump model for the Fz and Pz electrodes, warped by the bump locations. If one forces a five-bump model on data that does not have the fourth bump, it does identify a negative fourth bump but not the one we obtain from the actual data. The fourth bump force-fit in this synthetic data without a real fourth bump is much weaker and much earlier in the interval between the third and fifth bump.

(*Appendix continues*)

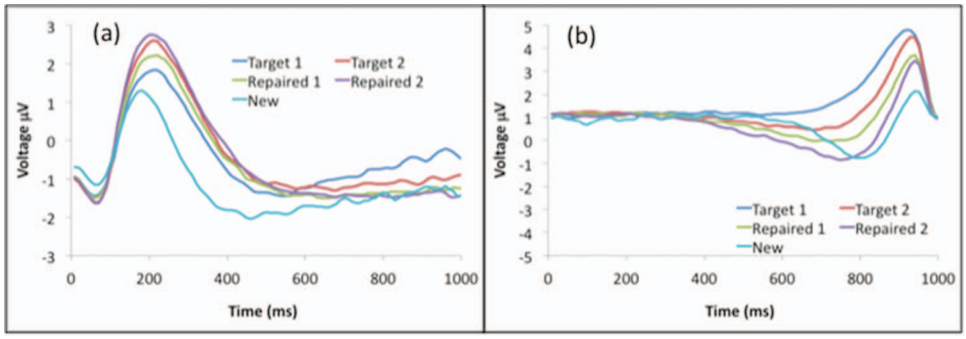

Figure A8. Synthetic data: (a) stimulus-locked Fz electrode; (b) response-locked Pz electrode. These are generated adding a parietal old–new effect to the fifth flat. See the online article for the color version of this figure. Fz = frontal; Pz = parietal. See the online article for the color version of this figure.

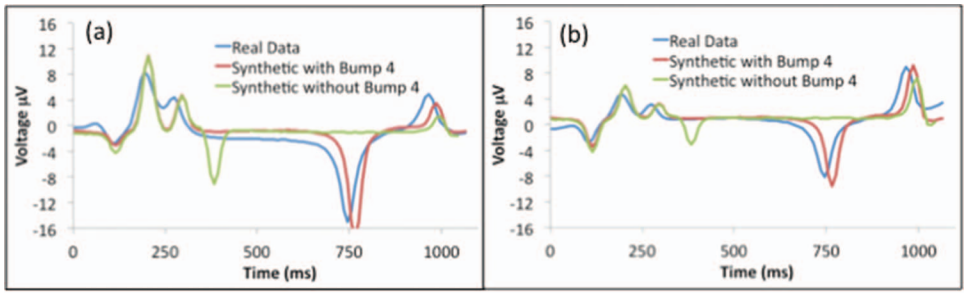

Figure A9. (a) Fz electrode and (b) Pz electrode for data warped according to a five-bump model fit to synthetic data with and without a fourth bump as well as the actual experimental data. For convenience, the fit is to all the data not separating out the conditions. Fz = frontal; Pz = parietal. See the online article for the color version of this figure.

Nonevent locked bumps. EEG data can plausibly have some nonevent locked peaks with meaningful topographies, perhaps in response to random internal thoughts or body sensations. One can wonder whether the presence of such components in the EEG activity might distort our conclusions about the existence of event-locked bumps. To investigate whether the nonevent locked bumps will have an effect on the estimation of the HSMM, we generated a synthetic dataset, with the model-specified bumps, but also 0 to 2 nonevent locked bumps on every trial (with equal probabilities). A random magnitude profile was determined for these bumps in the same range of magnitudes as the event-locked bumps. When this bump occurred it always had this same random profile (i.e., it was determined randomly once, then kept constant for all trials). The actual position of the bump was randomly placed within the trial. A five-bump model was still recovered as the best fit to the data and the location of the bumps was closely reproduced. Because these random peaks were not aligned in any way with the structure of the trial, the bump model was unable to pick up their existence. As far as the analysis software is concerned, these bumps are just part of the random noise.

Variable bump magnitudes. The analysis assumes that the magnitudes of the bumps are constant on every trial. If the true data had trial-to-trial variability in bump magnitude, would this introduce any error into our conclusions? To address this issue, we generated synthetic data where the magnitude varied independently on each PCA dimension for each bump according to a normal distribution. We were still able to accurately recover the parameters of the model adding normal noise at what we thought was the extreme plausible variation (90% of the trials within true mean plus or minus the average magnitude, 10% of the trials beyond). Basically, such variability just adds more noise to the data slightly worsening the quality of the fit.

Received May 31, 2015
Revision received January 20, 2016
Accepted January 28, 2016 ■