A neurocognitive model for predicting the fate of individual memories

Shannon M. Tubridy, David Halpern, Lila Davachi, & Todd M. Gureckis
Department of Psychology, 6 Washington Place
New York, NY 10003 USA

Abstract

One goal of cognitive science is to build theories of mental function that predict individual behavior. In this project we focus on predicting, for individual participants, which specific items in a list will be remembered at some point in the future. If you want to know if an individual will remember something, one commonsense approach is to give them a quiz or test such that a correct answer likely indicates later memory for an item. In this project we attempt to predict later memory without explicit assessments by jointly modeling both neural and behavioral data in a computational cognitive model which captures the dynamics of memory acquisition and decay. In this paper, we lay out a novel hierarchical Bayesian approach for combining neural and behavioral data and present results showing how fMRI signals recorded during the study phase of a memory task can improve our ability to predict (in held-out data) which items will be remembered or forgotten 72 hours later.
Keywords: memory, joint modeling, cognitive neuroscience

Introduction

A number of approaches in cognitive science and education attempt to use computer models to algorithmically tailor information presentation during study to the needs of individual learners (Smallwood, 1962; Atkinson, 1972; Fu et al., 2006; Ritter, Anderson, Koedinger, & Corbett, 2007; Pavlik & Anderson, 2008; Lindsey, Mozer, Cepeda, & Pashler, 2009; Rafferty, LaMar, & Griffiths, 2015). A core goal of these approaches is to leverage insights about the dynamics of learning and memory in order to predict which materials are likely to be forgotten and which will be remembered at future points in time.

Most of these approaches adapt their recommendations to individuals based primarily on assessments (e.g., performance on quizzes and tests given during or after a learning session)(Atkinson, 1972; Corbett & Anderson, 1995; Khajah, V. Lindsey, & Mozer, 2016). In this paper we attempt to predict an individual's memory (specifically which items will be remembered and which will be forgotten after a delay) without using explicit assessments. Our approach builds on research in cognitive neuroscience which has identified several neuroimaging correlates of successful memory formation (Davachi, 2006). We use functional magnetic resonance imaging (fMRI) to "peer into the minds" of learners while they study and to use the resulting information to improve predictions about their memory tested at a multi-day delay.

We begin by laying out a novel Hidden Markov Model (HMM, Rabiner, 1989) of memory which can learn to utilize fMRI signals as helpful indicators about the mnemonic status {remembered, forgotten} of individual memory traces. We then compare the predictive ability of this model to a number of simpler alternatives which lack access to the neural information but which incorporate other information about individuals such as their own self-reported judgements of learning (JOLs)(Nelson & Dulosky, 1991). To foreshadow, we find that the model utilizing fMRI signals recorded from individuals while they learned allowed us to predict which items they would remember better than alternative approaches which utilized explicit assessments. The success of this approach allows us to determine, without directly asking or testing participants, whether an experienced event is likely to be forgotten in the future and therefore deserving of additional practice.

Prior work on predicting human learning and memory

Atkinson (1972) sought to optimize the acquisition of a foreign language vocabulary using a Markov model of memory which algorithmically chose the sequence of words participants should study on each trial (Atkinson, 1972). In the critical test of the model, word-pairs were selected for study either randomly, by participants themselves, or using the fitted model.

The model-based approach assumed that each individual memory trace – memory for the association between words in a pair – could be in one of three mutually exclusive latent states $S = \{s_U, s_T, s_P\}$ representing (**U**)nknown, (**T**)emporarily stored, and (**P**)ermanently stored memories, respectively. Each time a word was studied there was a probability the memory trace would transition to the more fully learned T or P states according to the study transitions matrix shown in Figure 1. Alternatively, on trials when a word was not studied (e.g., another word was studied) there was a possibility of forgetting (i.e., moving from T to U) as reflected in the decay transitions. Every time an item was presented to the learner in the computer-aided condition, the model probabilistically updated a posterior probability estimate of the state of each memory trace according to the transition probabilities. In this way, the model used the history and dynamics of a trace (e.g., how long ago it was last studied, how many times it has been studied, etc.) to make a prediction about the status of a memory at any point in time.

The critical insight from Atkinson is that an explicit model of the dynamics of a memory trace can be exploited by an adaptive computer algorithm to predict human learning. This influential finding inspired a line of work that has utilized *Bayesian Knowledge Tracing* (BKT) to adapt instruction to individual learners (Corbett & Anderson, 1995) and generated a range of subsequent papers and modeling attempts (Pavlik & Anderson, 2008; Lindsey et al., 2009; Khajah, Lindsey, & Mozer, 2014).

Although these models show promise, they have several limitations. In particular, there is usually no information

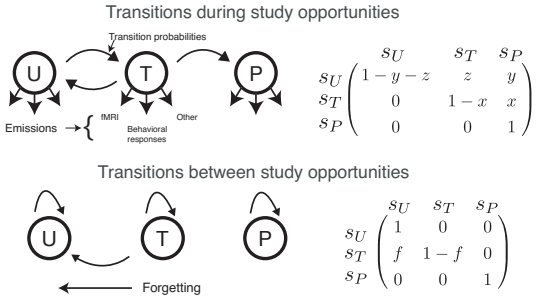

Figure 1: Structure of Atkinson's (1972) three state Markov model showing latent memory states, the allowable transitions between them during learning and forgetting, and the parameters governing the state transitions.

about the knowledge or memory status of some material until the learner completes an explicit assessment (e.g., test or metacognitive rating of learning). However, it has long been recognized that repeated assessments interrupt the learning context and may alter the nature of memory in consequential ways (Anderson, Bjork, & Bjork, 1994). In addition, the focus on explicit assessments runs the risk of ignoring other useful forms of information bearing on the underlying knowledge state of individuals (Anderson, Betts, Ferris, & Fincham, 2010; Anderson, 2012; Turner et al., 2013).

Signals of memory formation in the brain

Following pioneering clinical and animal research (Scoville & Milner, 1957; Mishkin, 1978) it is now common consensus that there are regions in the brain which are critically involved in memory acquisition. One important finding is that neuroimaging signals recorded *during study episodes* differ according to whether those episodes are later remembered compared to those that are forgotten (Paller & Wagner, 2002), a phenomenon known as the subsequent memory effect (SME). Of particular interest in the context of this paper is the fact that these differences are, on average, measurable at the time of the to-be-remembered experience, suggesting that the future memory status of individual episodes could be predicted from the brain response to a single trial. That is, rather than retroactively using later memory performance to identify the neural patterns predicting that performance, we hope to use these signals in a prospective fashion to predict the probability that events have been encoded successfully into memory.

A neuro-cognitive model of memory

Drawing inspiration from both the BKT and the subsequent memory literatures we propose a neurally-informed HMM of paired-associate memory. As a starting point, we adopt the three-state model introduced by Atkinson (1972) and extend it to include observable emissions (i.e., observable data reflective of the hidden states). Although highly schematized, this model provides a characterization of memory traces in terms of latent states and the dynamics of moving through those states over time and can be used to make predictions about future memory performance. A particular strength of this approach is flexibility in incorporating diverse kinds of observable emissions (Rabiner, 1989). Here we focus on BOLD signals measured with fMRI (as well as metacognitive judgments in the form of JOLs) because such signals have, in past work, been been related to future memory performance but the framework could readily incorporate other observations bearing on memory status (e.g., EEG, eye movements).

Our goal of combining neuroimaging data with a cognitive model is part of a broader recent initiative to find ways to link behavioral and neural data together with hypotheses about cognitive processes (Hawkins, Mittner, Forstmann, & Heathcote, 2017; Turner et al., 2013; Anderson et al., 2010; Anderson, 2012; Anderson, Pyke, & Fincham, 2016; Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016; Turner, Forstmann, Love, Palmeri, & Maanen, 2016). Although the research topics in these reports are diverse, they share the perspective that neural data can provide useful information about latent states or cognitive processes beyond that available from behavior alone and that the constraining influence of jointly modeling neural and behavioral data enables more accurate description of cognition.

Structure of the model

We assume that a learner studies a list of N word pairs or pair-wise associations. We characterize each memory trace (i.e., each word pair) as a HMM with the structure shown in Figure 1. For computational reasons, we assume that individual memory traces are independent but in future work this assumption can be easily relaxed to model inter-item interactions.

The core of each HMM is the set of discrete latent memory states $S = \{s_U, s_T, s_P\}$; a prior, $\Pi_{t=0}$, over the initial states (before study) for each word pair; the transition probabilities, T, which determine the probability of a trace moving between the different states at each point in time as set by the parameters x, y, z, and f (Figure 1A); and the emission distributions, E, defining the probability of observing data (behavioral or physiological) given a latent mental state and an external event eliciting an observable signal.

We propose two sets of transition probabilities (Figure 1), the application of which depends on the type of event, $e_t = \{\text{study,decay,test}\}$, occurring at time t for a particular word pair. If, at time t, word pair i is presented for study, the study transitions are applied, increasing the probability that the latent trace for pair i has moved into an accessible memory state. At timesteps when item i is not presented for study (e.g., another item is studied), the decay transitions are used, reflecting the possibility that learned items might fall back into the U state. The event type occurring at t also defines whether and which observable emissions can be expected. As in BKT, a behavioral response on a test event is an emission whose likelihood is dependent on the underlying state ($P[\text{correct response}|q = s]$). Similarly, presenting an item for study while a person is undergoing fMRI scanning results in an observed BOLD signal that may be related to later memory performance (which in turn is assumed to depend on the

latent memory state of the item).

Given a model and a protocol R specifying the order of events for a particular participant (e.g., the order of words studied and the time between presentations) we can infer, at any point in the study sequence, the most likely mnemonic status for each word pair. The posterior probability of each state for a particular word pair at time t can be obtained by Bayes' rule:

$$p[q_t = s' | o_t, e_t = g, q_{t-1} = s] = \frac{b_t^{g,s'} a_t^{g,s \to s'} \pi_{t-1}^{s}}{\sum_k b_t^{g,s_k} a_t^{g,s \to s_k} \pi_{t-1}^{s}} \quad (1)$$

The likelihood of observed data for a memory trace at time t conditional on event type g and a particular latent state s is given by $b_t^{g,s} = P[o_t | e_t = g, q_t = s]$. Some state-event combination are "silent" in the sense that if, for example, word pair j is presented for study there my be no observable behavior or other signal that bears on the status of item i. The transition probabilities for an item moving from state s to s' given event g – which determines the transition functions to be used – are represented as $a_t^{g,s \to s'}$, and π_{t-1}^{s} encodes the prior distribution over states for an item as provided by the initial state prior or the previous time step. When there are no observable emissions this reduces to application of the appropriate transition matrix to the posterior over states from the last time step.

Memory task - behavioral

The behavior that we seek to predict is performance on a cued-recall memory test for a set of Lithuanian-English word pairs. Participants' task is to study the word pairs and then, given a Lithuanian word, recall the associated English word. Starting with a normed set of Lithuanian-English words (Grimaldi, Pyc, & Rawson, 2010), we selected 45 word pairs to be learned. During study, participants see the pairs presented one at a time for 4 seconds each with a variable duration ITI (4s-16s) between trials (for consistency with event-related MRI timing). Each word pair is presented five times and no pair is presented for the nth repetition until all words have $n-1$ presentations. Importantly, and in contrast to many studies, all participants see the same sequence of study items. Although the model we use is simple and doesn't explicitly model factors like inter-item interactions during study, keeping a fixed protocol across people ensures that some of these effects will be captured in the acquisition and forgetting parameters we estimate.

Immediately following the study session participants give judgments of learning (JOLs): for each pair, participants use the mouse to indicate on a scale of 0-100 how likely they think they are to remember the association. Participants then return to the lab for a recall test either 24h, 72h or 168h (1 week) after the initial study session. During the recall test, participants are given a cued recall task in which they see a Lithuanian word presented on the screen and have up to 12 seconds to type in the associated English word. Recall performance for each trial was considered correct if participants typed the correct English word and all other responses were considered failures of recall.

Due to the high cost of fMRI data acquisition we took the approach of collecting a large behavioral dataset outside of the MRI scanner and combined those data with additional observations from participants who performed the same task during MRI scanning (under this view all participants are equally useful but purely behavioral subjects are treated as through their fMRI data are "missing"). Each behavioral participant (N=150) was tested at one of the three study-test delays. Including participants at each of three delays provides help in estimating the forgetting rate for each word in a way that allows separation between the T and P states. Both states are associated with successful recall, so including multiple delays allows us to separate those memory traces that are more likely to be recalled at shorter delays than longer delays (T state at end of study session) from those that are likely to be recalled at all delays (more likely to be P state traces).

Memory task - MRI

MRI participants (N=20) underwent the same study-test procedure except they were scanned during the study session and all MRI participants were tested at the 72h delay. MRI data were collected on a Siemens Prisma 3T. Functional data covering the cortex were acquired at 2.5 mm^3 with a 1 second TR (multi band factor 4) and anatomical scans were acquired at $1mm^3$.

Identifying fMRI emissions

After standard MRI preprocessing (Danker, Tompary, & Davachi, 2017), we selected data for inclusion in the model. We reduced the dimensionality of the fMRI data using group spatial independent components analysis (ICA) using the ICASSO algorithm as implemented in the GIFT ICA toolbox (http://mialab.mrn.org/software/gift/) (Calhoun, Adali, Pearlson, & Pekar, 2001; Van Maanen et al., 2011). This procedure, which is blind to trial information, results in a set of 60 independent components that are characterized by a particular spatial and temporal profile for each participant. Components that were unstable across estimations (ICASSO) and components associated with signal from ventricles or motion were discarded leaving 43 independent components for inclusion as model features. Individual trial activations were summarized as the mean of of timepoints encompassing 4-6 seconds post-stimulus onset (to account for the temporal lag in the BOLD response), giving us one activation value for each trial in each component for each MRI participant.

Predicting behavior

To assess whether our proposed model can accurately predict performance in the task we fit three variants: a model fit to trial timing and recall performance (the binary recall success scores for each word) (*Recall*); a model fit to trial timing, recall performance, and JOL emissions (*Recall+JOL*); and a model fit to trial timing, recall performance, and fMRI emissions (*Recall+MRI*). In each case the training data included

data from all of the behavioral data and a subset of the MRI participant data (see **Model evaluation** below).

The parameters to be estimated for all three models are the x, y, and z values controlling transitions between states during study opportunities and the f parameter determining forgetting rates (Figure 1). In the *Recall+JOL* and *Recall+MRI* models we also estimate the parameters for distributions of emission likelihoods (i.e., probability of fMRI signal or JOL ratings conditioned on the latent states of an item).

For all words we set the initial state priors, π_0^s, at $[0.99, 0.005, 0.005]$ for U, T, and P, respectively, as none of the participants in our study had prior experience with Lithuanian. We also fixed the probabilities of giving the correct behavioral response as $[.01, .9, .9]$ for latent memory states U, T, and P. This reflects the assumption that it is very unlikely that one would guess the correct answer in a cued recall test without any memory ($s = U$) and that, as in Atkinson, the primary difference between T and P states is the susceptibility to decay over time rather than the availability of a memory to recall (via the influence of the f parameter).

To get better estimates of the parameters, we used a hierarchical Bayesian model that used group-level priors over the parameters to regularize the estimates. Each x_w was drawn from a Logit-Normal(x, σ_x) where x itself was drawn from a Normal(0, 6) and σ_x was drawn from a Truncated-Normal(0, 1). The model for the f_w parameters was exactly the same. The simplices zy_w were generated using the following procedure: z and y were drawn from a Normal(0, 6). z_w and y_w were drawn from Normal(z, σ_z) and Normal(y, $sigma_y$) respectively with σ_z and σ_y both drawn from a Truncated-Normal(0, 1). Finally, zy_w was set to $softmax([0, z_w, y_w])$. This can be thought of as a multivariate generalization of the Logit-Normal with a diagonal covariance matrix.

In the models incorporating JOLs or MRI data we also estimated the mean and variance parameters for the Gaussian (truncated for JOLs) emission likelihood from each latent state. As with the transition parameters, the individual MRI components were treated as independent observations but the emission likelihood priors were hierarchical.

For estimation in this model, we used MCMC sampling via the NUTS algorithm as implemented in STAN to estimate the posterior over the parameters (4 chains of 200 iterations; 100 per chain discarded as burnin; 400 total samples per parameter). HMM models like this can be difficult to sample since parameters can be highly correlated and getting 200 samples for the fMRI models took 12-36 hours of compute time per model per fold. While hierarchical parameters were sometimes noisily estimated, to ensure convergence, we checked that estimates of the probability of recall had low $\hat{R}$ values (Stan Development Team, 2016).

Model evaluation In order to compare models, we want to know how well our models will predict new, unseen data. A common metric of model fit in cognitive science is the log likelihood of the data. Many approximation methods have been proposed for computing the expected log likelihood of new data such as AIC (Akaike, 1974) and WAIC (Watanabe, 2010). However, it is generally agreed that the generalization method with the fewest assumptions is K-fold cross validation and this is preferred when sufficient data and computational resources are available (Vehtari, Gelman, & Gabry, 2017). Our goal is to assess the utility of incorporating MRI signals into a memory model so we use K-fold cross validation where the folds were defined over the 20 fMRI subjects. We divided up the data from these subjects into ten equally sized folds. We then trained ten versions of each of the three model types where the training set consisted of all of the data from behavior-only subjects and nine of the ten folds of the fMRI subjects. On the held-out test set, we used the identity of the words and the trial timings (and JOL or fMRI observations, where appropriate) to generate the posterior probability of recall for each held out word at the time of test.

In addition, we evaluated a "baseline" fMRI model that predicted recall using just fMRI activations without the contributions of the three-state cognitive model. These *fMRI-baseline* predictions were generated by training an L2 regularized logistic regression model using the same cross validation regime as described above. The predictors were the fMRI activations measured on each study trial in each of the independent components used in the three-state model and the output was the probability of recall in the test sets.

As we are primarily interested in our ability to classify a new piece of data as successfully recalled or not rather than the log likelihood of the trial under the model, we adopted a cross-validated area under the ROC curve metric (ROC-AUC). The ROC-AUC can be interpreted somewhat like an accuracy measure where 0.5 represents chance prediction and higher values indicate better predictive performance of the model. Using ROC-AUC allows us to compare the held-out predictive performance of models with varying numbers of parameters while providing a metric of model performance that is relatively insensitive to class imbalance and does not prioritize one kind of error over another (e.g., trading off Hits versus Misses). The model ROCs were defined by calculating, in each cross validation fold, the proportion of predicted as remembered trials that were recalled correctly (*Hits*) and the proportion of predicted as remembered trials that were not (*False Alarms*) at each level of posterior recall probability given by the model.

Results

The *Recall* model, which is trained and evaluated using the timing of study and test trials and the observed recall performance, demonstrated above chance prediction performance in each of the cross validation folds (Figure 2, mean across fold AUC=0.64; sem=0.02). This result, consistent with prior work, establishes the ability of the structure of our three state model to predict memory performance given only the trial timings and identity of held-out word pairs.

The *Recall+JOL* model, which adds judgments of learning to the *Recall* model, improved our held-out prediction,

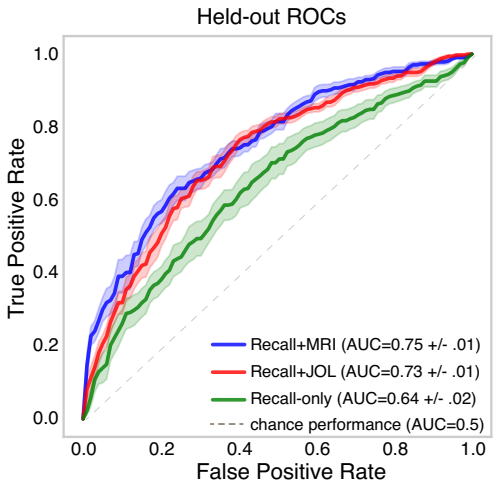

Figure 2: ROC curves for held-out predictions in each model variant. The curves show the mean ± sem across each of the cross validation folds.

achieving a mean held out ROC-AUC of 0.73 ($\pm$0.01), indicating that although knowing the basic information used in the *Recall* model (trial timing, recall accuracy for each word) is useful, additional observations in the form of JOLs can be used to refine predictions about held-out performance.

We next assessed whether we could use fMRI signals to make accurate predictions on held out trials. We trained the *Recall+MRI* model using the piecds in the *Recall* model and added fMRI observations for each MRI participants' study trials. The training set included the trial-level activations in each of forty-three independent components. The held out data included the trial timings for the held out word pairs as well as the study trial MRI observations from each of the components. This model achieved a mean ROC-AUC across folds for held out trials of 0.75 ($\pm$0.01). This result demonstrates that once trained, our model can predict trial-level memory performance given only the identity of a word pair, the timing of trials, and fMRI signal from study events and that this predictive accuracy surpasses that provided by the *Recall* and *Recall+JOL* models.

We also examined whether the fMRI data alone, without the structure provided by the models, could be used for prediction. The result of this analysis was a mean held-out ROC-AUC of 0.60 ($\pm$0.05) in the *fMRI-baseline* logistic regression model, indicating a benefit of joint cognitive and fMRI modeling relative to fMRI data alone.

Examining emission likelihoods

The *Recall+MRI* model included activation from a number of independent components as neural features. After estimating the emission parameters we can assess which components provided information about the latent model states. Used in this way, the joint model can be used as a tool for a richer understanding of how complex cognitive dynamics, especially those that might not be apparent in a more conventional analysis (e.g., a traditional subsequent memory analysis that only

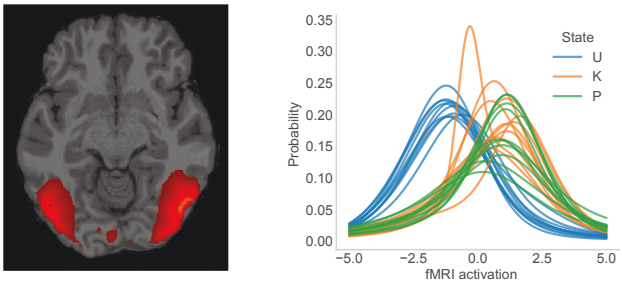

Figure 3: Topography and posterior predictive distributions for MRI emissions from most informative component. Individual traces show the distributions for each fold of the cross validation.

considers activation at the time of study and performance at the time of test), are instantiated in the brain. Figure 3 shows the voxel loadings and posterior predictive distributions for component activation conditioned on model state for the most informative component in our model. This component, associated primarily lateral occipital and fusiform gyrus regions involved in processing complex visual inputs, showed stronger activation for items in the T or P states relative to U.

Discussion

This paper introduced a neuro-cognitive model of memory that jointly models both brain and behavior within a single hierarchical Bayesian framework. Building off the three state model of memory developed by Atkinson (1972), we designed a discrete Hidden Markov Model of memory capable of learning to incorporate informative fMRI (or other) signals. The model is part of a growing movement towards joint modeling of brain and cognition (Turner et al., 2013; Anderson, Fincham, Schneider, & Yang, 2012). The advantage of this approach is that information from the the brain can help to constrain inferences about behavior, while inferences about behavior can help to constrain the interpretation of brain signals.

Although this work is preliminary and based on a relatively small number of fMRI subjects (N=20) exposed to a fixed trial sequence, we were able to make above-chance predictions on held out recall performance using only the timing and identity of individual study trials (*Recall* model). Incorporating observations in the form of individual metacognitive judgments of learning (JOLs) or MRI signal recorded during the study session led to improved predictions, with the *Recall+MRI* model achieving the best held out prediction performance.

Besides showing a framework for integrating cognition and brain measures in a single model, the predictions from our model can easily be incorporated into assistive learning technology (e.g., automated tutors). The estimated probability that a learner will remember some material can be used within optimization frameworks to design an optimal schedule of practice (Atkinson, 1972; Pavlik & Anderson, 2008; Rafferty et al., 2015). One advantage of using our approach is that we can identify the probability of future remembrances

without interrupting the learning process to perform explicit assessments. In addition, although our model summarizes a number of memory phenomena in a fairly abstracted discrete state model, it is possible to examine the posterior parameter estimates for the neural emissions to begin understanding the neural contributions to a dynamic, hidden set of latent cognitive processes.

Acknowledgements Supported by NSF grant DRL-1631436. We thank Camille Gasser and Victor Wang for assistance in data collection. Correspondence should be addressed to Shannon Tubridy (shannon.tubridy@nyu.edu) or Todd Gureckis (todd.gureckis@nyu.edu).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*(6), 716-723.

Anderson, J. (2012). Tracking problem solving by multivariate pattern analysis and Hidden Markov Model algorithms. *Neuropsychologia*, *50*(4), 487–98.

Anderson, J., Betts, S., Ferris, J., & Fincham, J. (2010). Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(15), 7018–23.

Anderson, J., Fincham, J., Schneider, D., & Yang, J. (2012). Using brain imaging to track problem solving in a complex state space. *NeuroImage*, *60*(1), 633–43.

Anderson, J., Pyke, A., & Fincham, J. (2016). Hidden Stages of Cognition Revealed in Patterns of Brain Activation. *Psychological Science*, *27*(9), 1215–1226.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063-1087.

Atkinson, R. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96*, 124-129.

Calhoun, V., Adali, T., Pearlson, G., & Pekar, J. (2001). A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapping*, *14*(3), 140-151.

Corbett, A., & Anderson, J. (1995). Knowledge tracking: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278.

Danker, J., Tompary, A., & Davachi, L. (2017). Trial-by-trial hippocampal encoding activation predicts the fidelity of cortical reinstatement during subsequent retrieval. *Cerebral Cortex*, *27*, 3515-3524.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Curr Opin Neurobiol*, *16*(6), 693–700.

Fu, W., Bothell, D., Douglass, S., Haimson, C., Sohn, M., & Anderson, J. (2006). Toward a real-time model-based training system. *Interacting with Computers*, *18*(6), 1215-1241.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* Chapman and Hall/CRC.

Grimaldi, P., Pyc, M., & Rawson, K. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for lithuanian-english paired associates. *Behavior Research Methods*, *42*, 634-642.

Hawkins, G., Mittner, M., Forstmann, B., & Heathcote, A. (2017). On the efficiency of neurally-informed cognitive models to identify latent cognitive states. *Journal of Mathematical Psychology*, *76*, 142-155.

Khajah, M., Lindsey, R., & Mozer, M. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *TopiCS in Cognitive Science*, *6*, 157-169.

Khajah, M., V. Lindsey, R., & Mozer, M. (2016). How deep is knowledge tracing?

Lindsey, R., Mozer, M., Cepeda, N., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the ninth annual conference on cognitive modeling (iccm).* Manchester, UK.

Mishkin, M. (1978). Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus. *Nature*, *273*(5660), 297–298.

Nelson, T., & Dulosky, J. (1991). When peoples judgments of learning (jol) are extremely accurate at predicting subsequent recall: The delayed-jol effect. *Psychological Science*, *2*, 267-270.

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

Pavlik, P., & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101-117.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications to speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Rafferty, A., LaMar, M., & Griffiths, T. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*, 584-618.

Ritter, S., Anderson, J., Koedinger, K., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin and Review*, *14*(2), 249-255.

Scoville, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry*, *20*(1), 11–21.

Smallwood, R. (1962). *A decision structure of teaching machines.* Cambridge, MA: MIT Press.

Stan Development Team. (2016). *PyStan: the python interface to Stan.* (Version 2.14.0.0)

Turner, B., Forstmann, B., Wagenmakers, E., Brown, S., Sederbefg, P., & Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*, *72*.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Maanen, L. V. (2016). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of eeg, fmri, and behavioral data. *NeuroImage*, *128*, 96 - 115.

Van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488–17495.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571-3594.