Knowledge Tracing Using the Brain

David Halpern*, Shannon Tubridy, Hong Yu Wang,
Camille Gasser, Pamela Osborn Popp, Lila Davachi, Todd M. Gureckis
Department of Psychology
New York University
New York, NY, 10003

ABSTRACT

Knowledge tracing is a popular and successful approach to modeling student learning. In this paper we investigate whether the addition of neuroimaging observations to a knowledge tracing model enables accurate prediction of memory performance in held-out data. We propose a Hidden Markov Model of memory acquisition related to Bayesian Knowledge Tracing and show how continuous functional magnetic resonance imaging (fMRI) signals can be incorporated as observations related to latent knowledge states. We then show, using data collected from a simple second-language learning experiment, that fMRI data acquired during a learning session can be used to improve predictions about student memory at test. The fitted models can also potentially give new insight into the neural mechanisms that contribute to learning and memory.

1. INTRODUCTION

A shared goal for both cognitive science and educational data mining is the development of accurate models of human learning. On the basic science side, learning and memory are important functions of the human brain that support our ability to flexibly interact with our environment. On the education side, predictive theories of learning may be leveraged by intelligent tutoring systems (ITS) to individually optimize instruction [3, 22].

Perhaps the most influential approach to modeling student learning in the educational data mining literature is "knowledge tracing" [5, 11] whereby the learned mastery of a particular skill or fact is treated as a latent state and the probability that a person's knowledge is in that state is updated in light of observed student behavior. For example, in Bayesian Knowledge Tracing (BKT), each learning unit is assumed to be in one of two discrete states: $\{unknown, known\}$. Each time the student engages in a learning activity, the latent knowledge can transition from the *unknown* to the *known* state with probability l. Performance on a test, quiz, or exercise is conditional on the latent knowledge state, such that being in the *known* state is typically associated with a higher probability of issuing a correct answer than being in the *unknown* state. Using the model, it is possible to infer posterior probabilities of the knowledge state of each learner and skill using Bayes' rule, given the pattern of responses made on various assessments or quizzes. These probabilities are then used to make predictions about learning performance for new students, as well as to design optimized instruction policies.

Research in this area focuses on building more precise models of student learning by, for instance, incorporating factors that reflect individual abilities [42, 23], contextual factors that contribute to errors [6], or models of the exact moment at which a skill is acquired [7]. However, one relatively underexplored question is what types of observable data may be most useful for informing inferences about latent knowledge states during learning. Of particular interest is the idea that many other features besides overt responses might be partially informative. For example, the student's response time to a test question may add additional information about learning alongside correctness [26, 38, 41]. Likewise, patterns of mouse or eye movements during a learning session might help index drifting attention [8, 29].

In this paper we demonstrate that it is possible to integrate indirect neural measurements of brain activity into a cognitive model of learning in a way that 1) can improve prediction of a learner's test performance at a 72 hour delay and 2) allows knowledge tracing without interrupting the learning environment with explicit tests or assessments (which can be distracting or may bias learning).

Although acquiring neural recordings is impractical in most educational settings, the approach of fusing multiple sources of sensor data about individual learners may be a generally useful method for the educational data mining literature. In addition, as we show in our results, such modeling efforts may also feedback to contribute to a better understanding of the neural and cognitive mechanisms that support learning and memory [2, 1, 35, 36]. Finally, as the cost and difficulty of making indirect neural recordings falls (e.g., due to the advent of portable, dry contact electroencephalogram or EEG) the practicality of utilizing such sensors will likely increase (c.f., [15]).

*D. Halpern and S. Tubridy contributed equally to the project and author order was determined arbitrarily.

We begin by reviewing past work in cognitive neuroscience which has attempted to identify predictive signals of learning and memory processes. Next we describe our approach fusing concepts from knowledge tracing with what is known about the cognitive neuroscience of memory. We then describe a dataset collected from human participants performing a simple second-language learning task while undergoing functional magnetic resonance imaging (fMRI). We compare the predictive power of a variety of models against held-out memory recall data at study-test delays ranging from one day to one week. From the fitted model we then extract the neural signals corresponding to learning in the study period.

1.1 Prior work using cognitive neuroscience methods to predict individual learning

The prediction and optimization of human learning has been a long standing goal of cognitive neuroscience research. On the prediction side, a number of studies have explored the "subsequent memory" paradigm [30, 25, 14, 28]. In these experiments, participants study controlled stimuli such as lists of word pairs while brain signals (such as the blood oxygen-level dependent "BOLD" signal measured via fMRI or event-related potentials, ERPs, assessed with EEG) are recorded. Some time later, participants' memory is tested for the material they saw during study. Accuracy on each memory test item is used to back-sort the neural data recordings into brain patterns associated with successful versus unsuccessful later memory. Regions with a reliable difference in brain activation between these two classes are taken to reflect neural correlates supporting lasting memory formation. Across these studies a coherent set of brain regions have been identified as being involved in human memory formation including the hippocampus and medial temporal lobe, which have long been associated with memory formation on the basis of animal and lesion studies [31, 9].

Building on this work, Fukuda et al. (2015) identified two EEG-based subsequent memory signals and used these to classify study trials in a memory experiment as likely to be remembered (*initially well studied*) or forgotten (*intially poorly studied*). In a subsequent session, participants were allowed to restudy half of the items identified as *initially well studied* and half of the items identified as *initially poorly studied*. A final test then assessed knowledge for all of the items. Of particular interest was the finding that the restudy opportunity most benefitted the *initially poorly studied* items compared to the other items. Importantly, the entire prediction about what was or wasn't well studied was based exclusively on indirect neural recordings for each subject rather than any explicit assessment or test.

The subsequent memory paradigm has been a powerful tool for studying the neural basis of memory. However, the cognitive neuroscience literature does not currently take advantage of the wealth of knowledge about predicting individual learning from the educational data mining and cognitive modeling literatures. For example, classifying brain patterns as forgotten based on a single test fails to account for the possibility of "slippage" (errors in performance of a mastered skill due to chance) which is central to BKT models [11]. Likewise, when an item is not remembered at test it could be for a number of reasons: the item may have been poorly encoded during the study session, or perhaps was well encoded and would have been remembered at an earlier study session but was simply forgotten due to decay or interference. Structured models such as Hidden Markov Models (HMMs) can account for such latent memory dynamics and use them to help improve predictions. The subsequent memory approach is also difficult to apply when learners get repeated study opportunities because of ambiguity about which brain scans should be classified as causally related to the test performance. Finally, the standards for model development within the machine learning and data mining communities is predictive performance on held-out data which is often more difficult than describing statistically reliably patterns within a single data set due to the ability to overfit.

To address these issues, we describe an approach to the simultaneous modeling of behavior and neural recordings in a single knowledge tracing model[1]. Our aim is to demonstrate the value of combining insights from these still somewhat disparate literatures. The approach we take is in some ways similar to work by Anderson and colleagues that has tried to infer from fMRI the mental state of individuals as they engage in complex math problems [2, 1, 4, 43, 34] (see also [35, 36]). While these reports hint at the utility of combining fMRI with probabilistic cognitive models, this prior work does not specifically address the learning and memory issues considered here.

2. THE OMNI DATA SET

The dataset we consider, part of the NSF-funded "Optimizing Memory using Neural Information" (OMNI) project[2], consists of human performance on a cued-recall memory test for a set of Lithuanian-English word translations. The learner's task is to study the word pairs across multiple presentations and then, after a delay, recall the English associate for a presented Lithuanian word.

Starting with a normed set of Lithuanian-English words, we selected 45 translation pairs [21]. During study, participants saw the translation pairs presented one at a time for 4 seconds each with a variable duration inter-trial interval (4s-16s for consistency with event-related MRI timing). Words were presented on a computer screen with the Lithuanian word at the top of the screen and the English translation underneath.

Each word pair was presented five times and no pair was presented for the nth repetition until all words had $n-1$ presentations. Importantly, and in contrast to many psychology studies on the subsequent memory effect, all participants see the same sequence of study items[3]. Immediately following the study session participants gave judgments of

[1]Here we focus on fMRI due to improved spatial resolution, even though other methods (e.g., EEG and skin conductance response), also provide useful signals that correlate with memory performance and could be incorporated into our approach.

[2]http://gureckislab.org/omni

[3]Although the models we apply do not explicitly model inter-item interactions, maintaining a fixed sequence across participants ensures that some of these inter-item effects will be captured in the model parameters we estimate because, for instance, the measured difficulty of a word is always assessed with respect to the other list items.

learning (JOLs, [24]): for each pair participants were presented with the Lithuanian and English word and used the computer mouse to indicate on a scale of 0-100 how likely they were to remember the association in one week.

Participants were given either an immediate recall test (0 hours) or returned to the lab approximately 24, 72, or 168 hours after the initial study session (randomly assigned)[4]. During the recall test, participants saw a Lithuanian word presented on the screen and had to type the associated English word. A trial was coded as correct if participants typed the correct English word (allowing for typographic errors) and all other responses were incorrect.

For more efficient estimation of the different model parameters, we conducted a large behavioral experiment 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 though their MRI data are "missing" and so estimates of their learning are based on the observed JOLs and recall performance). Each participant (N=189) was tested at one of the four study-test delays. Among the behavioral participants (i.e., no MRI data) the group Ns were 20, 49, 60, and 49 in the 0, 24, 72, and 168 hour study-test delay groups, respectively. All MRI participants (N=21) were tested at the 72 hour delay.

MRI participants underwent an identical study-test procedure as the behavioral participants except they were scanned during the study session. MRI data were collected on a Siemens Prisma 3T at the New York University Center for Brain Imaging. Functional Blood Oxygen-Level Dependent (BOLD) data covering the cortex were acquired at a spatial resolution of 2.5 mm^3 with a 1 second repetition time (TR; the temporal resolution of the fMRI data) and anatomical scans were collected at a spatial resolution of .75 mm^3.

To summarize, the final data set consists of a record for each learner that contains: the pattern of recall attempts for each list item, JOLs collected after the study session for each list item, and, for each MRI participant, the 65x77x73 set of voxel measurements across 2936 time-points describing the BOLD signal recorded with MRI.

Figure 1 shows key features of the behavioral data. Across the four different test delays, memory performance generally drops, likely due to forgetting. Participant performance varied widely from 0 to 100 percent correct. In addition, across participants, average JOLs following study were weakly correlated with performance ($r = [0.43, 0.24, 0.31, 0.55]$ and $p = [0.06, 0.10, 0.004, 3.4e{-}5]$ in the 0h, 24h, 72h, and 168h groups, respectively). Pooling across all participants, the mean JOL correlation with final performance is low but significant, $r = .365, p < 1e{-}7$.

[4]Due to schedule difficulties a one subject returned at 48 hours but we still included their data in the modeling. In addition, 9 of the 72 hour subjects were scanned in a different fMRI scanner but we only include their behavioral data here.

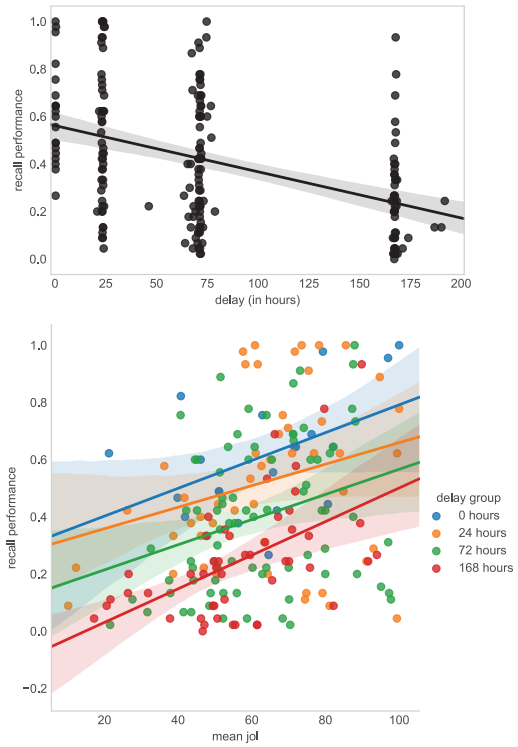

Figure 1: Top: Mean recall performance (% correct) for individuals (dots) at each study-test delay. Bottom: Mean individual participant Judgment of Learning is correlated with individual overall percent recalled within each delay condition.

3. INFERRING KNOWLEDGE STATES FROM BEHAVIORAL AND NEURAL DATA

The following section describes the basic mathematical structure of our models. Similar to BKT, the core of our approach assumes a probabilistic representation of the latent mnemonic status (e.g., *remembered* versus *forgotten*) of each item on the to-be-remembered list and we begin with established two- and three-state models that have shown effectiveness in tracking learning and memory [5, 11]. Where our models differs from past knowledge tracing approaches is that we propose a mapping between these latent mnemonic states and patterns of brain activity that can allow the brain data to inform this inference.

3.1 A Hidden Markov Model of Memory

Like BKT, our approach draws heavily from the structure of HMMs. Each memory trace, i, (i.e., memory for the association between two words) is represented as a non-homogenous, censored Hidden Markov Model with the following properties (notation follows [27]):

3.1.1 States

Each trace can be in one of a number of discrete mnemonic **states**, S. For simplicity we will begin with a two state $S = \{s_U, s_K\}$ model with states corresponding to *unknown* and *known* similar to BKT. However, we also consider a more complex, three-state model first proposed by Atkin-

son [5]. The three-state model has states $S = \{s_U, s_K, s_P\}$ corresponding to *unknown*, *known* (with possibility of forgetting), and *permanently known* (see Figure 2). Across both types of models the s_K and s_P states represent memories that have generally higher recall probabilities (e.g., $\Pr[recall = correct|s_P] > 0$), but the s_K state is susceptible to decay between study events while the s_P state is absorbing[5]. The current state of item i at time t will be denoted q_t^i.

3.1.2 Priors

A **prior**, $\pi_{t=0}$, that captures our initial belief of the memory state of all items. The prior for a particular item memory, i, can be written as $\pi_{t=0}^{i,s} = \Pr[q_{t=0}^i = s]$ for $s \in \{s_U, s_K\}$ (two state) or $s \in \{s_U, s_K, s_P\}$ (three state). With unfamiliar learning materials we assume that the initial memory status is heavily biased towards the unknown state (i.e., $\pi_{t=0}^{i,s_U}$ is much higher than for any other state).

3.1.3 Transitions

A set of **transition probabilities**, A, which determine the likelihood that a memory will move between the different states at each time point. In prototypical HMMs the transition probabilities are stationary and the same transitions are applied at each time step. In our model there are different sets of transition probabilities which are applied at a given time step depend on the type of external "event", e_t^i, that occurs (e.g., a study trial versus a time step between trials; Figure 2). For memory trace i the transition probability of moving from state s to s' after an event of type g will be denoted $a_t^{i,g,s\to s'} = \Pr[q_t^i = s'|e_t^i = g, q_{t-1}^i = s]$ where g indicates the specific event type on trial t.

Event types depend on the particular experiment design but here include "study trial" (*study*), "study with JOL trial" (*study+JOL*), "timestep in which memory decays" (*decay*), and "test trial" (*test*). Generally, during *study* or *study+JOL* events we assume that items tend to transition from a more poorly learned state to a more fully learned state. The probability of transitioning to a new state on a study trial is represented in our three state model by parameters x, y and z and in the two-state model by parameter l (see Figure 2). During *decay*, items in a non-permanent state (s_K) have a probability of transitioning to the *unknown* state with probability f while items in s_P (in the three-state model) remain in the permanently learned state. Decay events are necessary to account for the patterns of forgetting across the study-test delay intervals shown in Figure 1. We assume test trials have no effect on transitions as they appear at the end of the task.

We define an experiment **protocol**, E, as a $\mathcal{N}$ x $\mathcal{T}$ matrix where $\mathcal{N}$ is the number of items being studied and $\mathcal{T}$ is the total number of micro-time steps modeled in the experiment. Each entry of the matrix, e_t^i, codes which of a discrete set of event types occurred on a time step as described above. The protocol captures the dependencies between event sequences

[5]One way for the model to capture the difference in performance at 24 versus 168 hours is to assume different mixtures of the s_K and s_P states following learning. For example, at 168 hours, traces in s_P state may dominate correct responses.

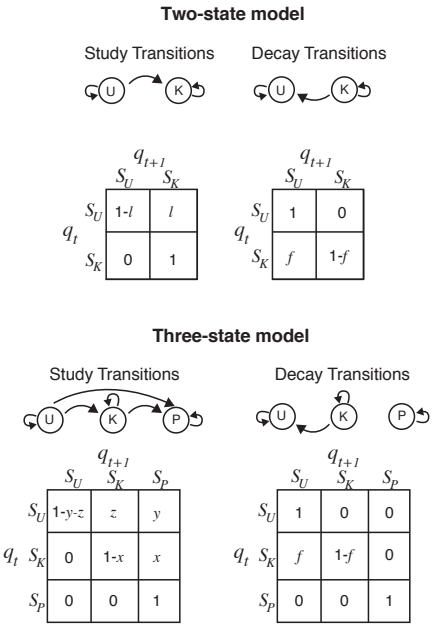

Figure 2: The matrix of transition probabilities for either study or decay events in the two and three state model. The letters within each matrix reflect the transition parameters which are estimated to data. The state labels U are "unknown", K are "known" (with possible forgetting), and P are "permanently known."

that influence different memory traces. For example, if word w is studied on a given trial, then all the other items on the list might undergo a memory decay event during the same time step. This way the protocol enforces the implicit tradeoffs of studying one item over others at a particular point in time.

3.1.4 Observable signals

The mapping between brain and behavior is made through a set of **observation distributions**, B, which define the probabilities that, on event type g at time t, an observable random variable of data type d, $\mathbf{o}_t^{g,d}$ takes on a value $v_k^{g,d}$ from a (potentially infinite) alphabet $v^{g,d}$. For each memory trace i, we can write the probability of its associated observables as $b_t^{i,g,s,d}(v_k^{g,d}) = \Pr[o^{g,d} = v_k^{g,d}|e_t^i = g, q_t^i = s]$. Observation distributions in effect define the full generative model that links both behavior and neural information to underlying knowledge states. Here we consider three types of observations: behavioral assessments (*recall*), JOLs (*JOL*), and hemodynamic fMRI measurements (*MRI*). However, this approach can easily incorporate many other measures including response time, pupil dilation, EEG measurements, or alternative fMRI signals.

Behavioral Assessments. At certain points during the experiment the protocol might define a memory test event. On these types of trials the subject might be asked to recall a studied item from memory or to recognize it from a list of alternatives. The response given on these trials is treated as an observation associated with this particu-

lar type of event. Specifically, the alphabet is $v^{test,recall} \in \{correct, incorrrect\}$ and $v^{g,recall} \in \emptyset$ for $g \neq test$, reflecting the absence of any recall response on non-test events. The distribution of test question answers about memory trace i from state s at time t, is then $b_t^{i,test,s,recall}(correct) = p_{\text{recall}_s}$ and $b_t^{i,test,s,recall}(incorrect) = 1 - p_{\text{recall}_s}$ where p_{recall_s} is defined (or fitted) for each memory state. For other trial types, i.e. $g \neq test$, $b_t^{i,g,s,recall}(\emptyset) = 1$. So the update to state posterior probabilities on those events is driven by the state transitions. The parameters governing the probability of issuing a correct response conditioned on the latent memory state are equivalent to the "guess" and "slip" parameters in BKT.

Judgments of Learning. JOL responses were only given on the last study trial (a *study+JOL* event). JOL data were included in the model as the raw response/100 to each JOL trial for each person, i.e. $v^{study+JOL,JOL} \in [0,1]$ and null for other trial types. We model the distribution of JOLs as a truncated Gaussian distribution in the range 0 to 1, i.e. $b_t^{i,study+JOL,s,JOL} = TN(\mu_{JOL_s}, \sigma_{JOL_s}, 0, 1)$ with μ_{JOL_s} and σ_{JOL_s} defined independently for each state s.

Hemodynamic fMRI measurement. Functional MRI scans provide time-series data for each of a large set of 3-dimensional voxels tiling the imaged volume (e.g., the brain). In studies measuring fMRI activation levels at specific timepoints it is common to estimate the activation level within voxels and then average voxels within spatial clusters, whether spatially contiguous (regions of interest, or ROIs) or sets of spatially disjoint but functionally related voxels showing similar response profiles (e.g., independent components). Due to the central limit theorem we can expect that the mean activation within a set of such voxels will be approximately normal. We also expect, based on prior work, that there will be a mean shift in the fMRI activation levels of various brain regions during study trials that are later remembered compared to those that are later forgotten [14, 28]. We collect fMRI data for each *study* trial. The fMRI observation consists of N_{fMRI} features. Therefore, $v^{study,MRI} \in \mathbb{R}^{N_{\text{fMRI}}}$ and null otherwise. We model the fMRI state observation distributions as independent Gaussians for each feature n_i, i.e. $b_t^{i,study,s,MRI_{n_i}} = N(\mu_{MRI_{n_i}s}, \sigma_{MRI_{n_i}s})$.

3.1.5 Inference

The full model is specified by a protocol, E, a set of priors over the states, $\pi_{t=0}$, a set of transition probabilities, A, and a set of observation distributions associated with each state-event pair, B. Using Bayes' rule, the posterior probability that a memory trace on trial t is in state $s' \in S$ is:

$$\pi_t^{i,s'} = \frac{b_t^{i,g,s',d} a_t^{i,g,s \to s'} \pi_{t-1}^{i,s}}{\sum_j b_t^{i,g,s_j,d} a_t^{i,g,s \to s_j} \pi_{t-1}^{i,s}} \quad (1)$$

3.1.6 Illustrative calculation

To illustrate the impact of hypothetical fMRI observations, consider Figure 3 which shows the protocol, E, for the timing of study events for two memory traces (Panel A): item 1 (black) and item 2 (white). On time points where item 1 is studied the protocol has a black cell (and similarly for

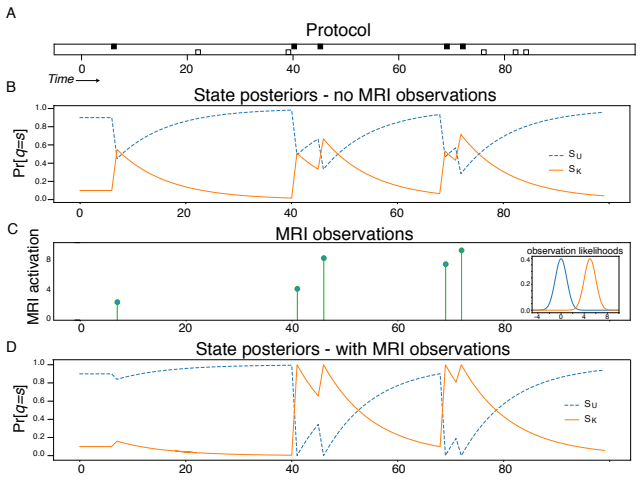

Figure 3: Example illustration of the effect of fMRI observations on inferences about latent knowledge in a two state-model. A) Protocol showing the timing of study events for item 1 (black boxes) and item 2 (white boxes). B) State posterior estimates for item 1 obtained from a hypothetical setting of the two-state model parameters (dashed blue = S_U, solid orange = S_K). C) Hypothetical "observed" fMRI signal on each study trial for item 1 (inset shows the probability density function over MRI observation values conditioned on the state). D) State posteriors for item 1 after incorporating the observation likelihoods from study trials for this item. The inferred state probabilities are dramatically altered by the incorporation of the MRI observation (see text).

item 2 using white). Panel B shows hypothetical evolutions over time for the two-state posterior probabilities $\{s_U, s_K\}$ for item 1 obtained by applying the study and forgetting transitions as shown in Figure 2 but without other observable information (i.e., a Markov model). In this example we set the l transition parameter applied on study events to 0.4 and the f parameter governing decay to 0.1.

At time point 1 the priors reflect the fact that before any study attempts a person is unlikely to know the item (e.g., $\pi_{t=0}^{i=1,s_U} = .9$). At time point 6, item 1 is presented for study for the first time and the posterior probabilities of each state are updated by applying the study transition probabilities to the state posteriors on time $t-1$. Immediately after this study event, Panel B shows that there is now an increased probability of item 1 being in state s_K (solid orange line). However, between time point 6 and 40, item 1 is not presented again and so for each time step between we apply the decay transitions leading to gradual forgetting.

The addition of observable signals that are probabilistically related to latent memory states alters these predictions. The inset figure in Panel C shows how the mean response from a set of voxels in the human brain might result in Gaussian-distributed summed BOLD signals that overlap but differ depending on the state of the memory (e.g., signal being stronger for s_K, orange, than for the s_U, blue, state). Panel C illustrates a hypothetical sequence of fMRI measurements that could be made about item 1 during the study trials

(i.e., samples from the Gaussian distributions from the inset plot).

Panel D shows the posterior estimates of item 1's state at each time point obtained through combination of the transition dynamics **and** MRI observations (i.e., using the Hidden Markov Model). As can be seen comparing panel B and D, the addition of observations that are probabilistically associated with latent states can lead to different inferences about the posterior probabilities over those states. Until item 1 is presented at time point 6 the posterior estimates are the same in the Markov and Hidden Markov Models. However, at time point 6 we observe a fMRI signal of a particular magnitude which in turn has a likelihood of originating from each of the two underlying states. If we take into account the observed signal, our estimates of the posterior over states change, since a fairly small signal was observed and the likelihood of such a signal is substantially larger for state s_U than s_K. Consequently, our belief that the item is in state s_K is lower when we include the observation in our estimates than when we simply use the transition probabilities.

Similarly, at time point 40 item 1 is presented for a second study opportunity. Without observations our best estimate of the state probabilities suggests we should be indifferent between s_U or s_K, but the larger MRI observation observed is unlikely to have emerged from the unknown state and so the observation-constrained posterior estimates are weighted much more heavily towards the s_K state. By including the Markov dynamics characterizing the likely temporal evolution of memories, we can adjudicate between otherwise ambiguous neural signals by appropriately dealing with uncertainty in measurement.

3.1.7 Model Evaluation and Fitting Procedure

The following section details the model evaluation, comparison, and feature selection strategies we used.

Model parameterization. Partially due to identifiability concerns [37, 17], some parameters were fixed to semantically coherent values [16], while others were estimated from the data.

For all words we fixed the initial state priors, $\pi_{t=0}$, as $[.99, .01]$ or $[0.99, 0.005, 0.005]$ for s_U, s_K in the two-state model or s_U, s_K, and s_P in the three-state model, respectively. This was motivated by the fact that none of the learners in our dataset had prior experience with Lithuanian. We also fixed the probabilities of giving the correct test response, $\mathbf{p}_{\text{recall}}$ as $[.01, .9]$ and $[.01, .9, .9]$ for latent memory states s_U and s_K (two-state model) or s_U, s_K, and s_P (three state model, see below), respectively. This reflects the assumption that it is very unlikely that one would guess the correct answer in a cued recall test without any memory ($s = s_U$) and that, as in [5], the primary difference between s_K and s_P in the three-state model is the susceptibility to decay over time rather than the availability of a memory to recall (via the influence of the f parameter; see Figure 2).

Fitted parameters include those determining the transition probabilities and observation distributions within each model. Both the two- and three-state models have transition probabilities to fit for each word pair w (summarized in Figure 2). In the two-state model these are the l_w and f_w parameters controlling memory strengthening and decay, respectively. For the three-state models, the x_w, y_w, and z_w values control transitions between states during study opportunities and the f_w parameter determines forgetting rates.

Although the learning trajectories for each word pair were instantiated in separate HMMs, 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.

When fitting models that incorporated JOLs or MRI data we also estimated the means and variance parameters for the Gaussian (truncated for JOLs) observation likelihood from each latent state. For the JOL distributions, each μ_{JOL_s} was drawn from a Normal(.5, .5) and each σ_{JOL_s} was drawn from Inverse-Gamma(1, 2). Similarly, for each fMRI feature n_i (see below) in state s, $\mu_{MRI_{n_i}s}$ was drawn from a Normal(0, 1) and $\sigma_{MRI_{n_i}s}$ was drawn from an Inverse-Gamma(1, 2).

fMRI feature selection. After standard MRI preprocessing [12], 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/) [10, 39]. This procedure, which is blind to trial information and memory outcome, resulted in a set of 60 independent components that are characterized by a particular temporal (the time-course of activation) and spatial (the loading of each component on fMRI voxels) 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 for each identified component were summarized as the mean of timepoints encompassing 4-6 seconds post-stimulus onset (to account for the temporal lag in the BOLD response), resulting in one activation value for each trial in each component for each MRI participant.

Model estimation. We used MCMC sampling via the NUTS algorithm as implemented in Stan [32] to estimate the posterior over the parameters (4 chains of 200 iterations; 100 per chain discarded as burnin; 400 total samples per parameter). To ensure convergence, we checked that estimates of the probability of recall had low $\hat{R}$ values (a measure of whether the sampling chains are converging to similar estimates) [33, 19].

Model evaluation. In order to compare models, we want to evaluate how well our models will predict new, unseen

data. It is generally agreed that the generalization method with the fewest assumptions is leave-one-out cross validation, which is preferred when sufficient data and computational resources are available [40]. To conserve on computational resources, here we use K-fold cross validation, setting K to 10. Because our goal is to assess the utility of incorporating MRI signals into a memory model, the held-out data only included data from the 20 fMRI subjects. We divided up the data from these subjects into ten equally sized folds. We then trained ten versions of each model 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.

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.

Model Comparison. We fit three variants of each of the two- and three-state models: a *Recall* model fit to trial timing and recall performance (the binary recall success scores for each word); a model fit to trial timing, recall performance, and JOL observations (*Recall+JOL*); and a model fit to trial timing, recall performance, and fMRI observations (*Recall+MRI*). In each case the training data included data from all of the behavioral participants and a subset of the MRI participant data, and models were evaluated on held-out data. The logic of these comparisons is to see if the models incorporating additional observations (*Recall+JOL* and *Recall+MRI*) provide a better basis for prediction than do the purely behavioral models. In addition, we are interested in whether the model incorporating MRI observations is able to outperform the model incorporating JOLs. This would suggest that the brain data contains more information relevant about memory performance than do people's own self-reports about their memory fidelity. While we are ultimately interested in held-out predictive performance, the models do differ in model complexity. In raw numbers, for the two-state models, the *Recall* model had 2 x 45 word parameters and 4 hyperparameters, the $Recall + JOL$ model added 4 parameters, and the $Recall + MRI$ model added $4N_{\text{fMRI}}$ parameters. For the three state models, the *Recall* model had 4 x 45 word parameters and 7 hyperparameters, the $Recall + JOL$ model added 6 parameters, and the $Recall + MRI$ model added $6N_{\text{fMRI}}$ parameters. However, due to the hierarchical nature of these models, the effective number of parameters may have differed depending on the amount of regularization done by the hierarchical prior.

Table 1: Cross validated Area Under the Curve of the Receiver-Operating Characteristic (ROC-AUC) with ± standard error (in parentheses) across folds.

	two-state model	three-state model
Recall	0.64 (.02)	.64 (.02)
Recall+JOL	0.73 (.01)	.73 (.01)
Recall+MRI	0.72 (.02)	.75 (.01)

4. RESULTS

4.1 Two-state model

For each variant of the two-state model (*Recall*, *Recall+JOL*, *Recall+MRI*) we computed the ROC-AUC for predictions of recall accuracy in held-out trials for the MRI participants. The *Recall* model, trained on the timing of study and test trials and recall performance, achieved a mean (across held-out folds) ROC-AUC of 0.64 (±.02), providing an above chance baseline model against which to evaluate the utility of JOL and fMRI observations (Figure 4A).

The *Recall+JOL*, which adds judgments of learning to both the training and evaluation of the *Recall* model, achieved a mean held-out ROC-AUC of .73 (±.01), improving our predictions relative to the *Recall* model. This shows that metacognitive judgments collected from individuals at the end of a learning session can be used to refine predictions about held-out recall performance.

We next assessed whether fMRI signals recorded during study events could be leveraged to make predictions about held-out performance. The *Recall+MRI* model yielded a held-out ROC-AUC of 0.72 (±.02). Although the held-out performance did not surpass the *Recall+JOL* model, this result indicated that there may be information in the MRI measurements that could be used to make predictions about held-out memory recall performance.

4.2 Three-state model

We next considered whether a more elaborated model of memory could leverage more subtle dynamics of the fMRI data.[6] The held out ROC-AUCs for the *Recall* and *Recall+JOL* three-state models did not differ from those observed in the two-state model (Figure 4B). However, the three-state MRI model boosted the held-out AUC to .75 (±.01) which was an improvement compared to the original two-state *Recall+MRI* model. This was also, in terms of held-out predictions, the most successful model we considered in these comparisons (but see Conclusions), building confidence in the utility of incorporating neural signals into knowledge tracing models.

[6]Although our primary interest in this work is evaluating the held-out predictions of our models, we note that complexity of the three-state model means that three-state *Recall* or *Recall+JOL* variants may not be identifiable due to the sparseness of observations (a single recall outcome or the recall outcome and a single JOL) [37, 17] However, for the MRI participants we have data for every trial, enabling estimation of a three-state *Recall+MRI* model.

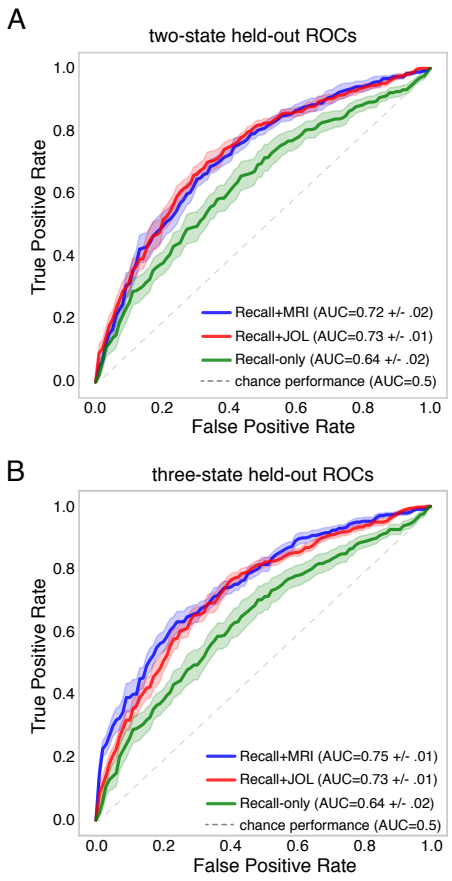

Figure 4: ROC curves for held-out predictions in each of the two-state (panel A) and three-state (panel B) model variants (*Recall*, *Recall+JOL*, *Recall+MRI*). The curves show the mean ± sem across each of the cross validation folds.

In addition, whereas the *Recall* and *Recall+JOL* models did not discriminate between the two- and three-state models, the fMRI data enabled better predictions using the three-state model, highlighting the utility of neuroimaging data in selecting between cognitive models.

4.3 Relating model dynamics to the brain

In addition to the improvements in memory prediction afforded by joint modeling of behavioral and neural data, our approach also allows for examination of fMRI data in light of the estimated models. Figure 5 presents two example analyses in this vein.

Figure 5A shows the contrast map resulting from regressing the change in posterior probability of s_K associated with each study trial (as estimated in the two-state *Recall* model) against the fMRI time-series in each voxel. Using the estimated two-state *Recall* model parameters, we extracted the state posteriors on each study event for the MRI participants based on the sequence and timing of study trials. We then calculated the change in predicted state posterior from just before to just after a study trial and used this change as the predictor for brain activations. This analysis is related

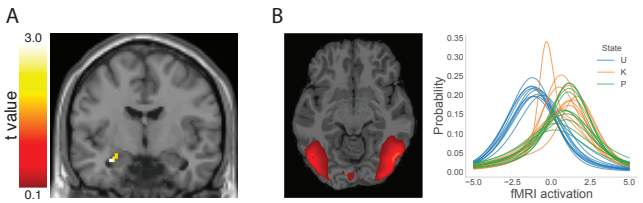

Figure 5: Examples of using estimated model to analyze the brain. A) Coronal slice showing left anterior hippocampal voxels tracking the change in s_K state posterior for each study trial. B) Topography (left; axial slice) and posterior predictive distributions (right) for MRI activations from most informative component in the three-state model. Individual traces show the distributions for each fold of the cross validation

to the General Linear Model approach often used in the subsequent memory literature, except that rather than using binary regressors that coded for *remembered* or *forgotten* outcomes as determined by a recall test, we used the estimated continuous state posteriors from the two-state model.

Using a knowledge tracing model in this way to provide estimates of when a particular item is learned during a study sequence with multiple repetitions allows for more sensitive analyses of the brain's relationship to cognitive processes unfolding over extended time. Interestingly, we found that the voxels significantly correlated with the change-in-state-posterior regressor were a cluster in left anterior hippocampus, consistent with the hypothesized role for this region in encoding new information into memory [13].

An alternative way to use the fitted models is to examine the estimated fMRI features' observation likelihoods for each latent knowledge state. The *Recall+MRI* model included activation from a number of independent components as candidate neural features. After estimating the model, the fMRI observation parameters can be used to assess which components provided information about the latent model states. Used in this way, the joint model can be used as a tool for understanding 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 considers activation at the time of study and performance at the time of test), are instantiated in the brain. The most informative component in our model was associated with voxels in lateral occipital and fusiform gyrus regions involved in processing complex visual inputs, as shown in an axial slice through the brain (anterior/posterior of the brain is up/down in the image) in figure 5B. The posterior predictive distributions for component activation conditioned on model state are also shown in figure 5B, and these estimated distributions showed stronger activation for items in the K or P states relative to U.

5. CONCLUSIONS

We evaluate a framework for integrating neuroimaging recordings into a knowledge tracing model. Our approach builds upon recent reports showing robust memory-related signals in the brain. We collected a medium-sized data set of human participants performing a second-language acquisition

task both inside and outside a scanner. We then compared a variety of models on their ability to predict held out data for the MRI participants. Our most predictive model was a three-state hidden Markov model that incorporated neural measurements. This is interesting because this model was more predictive than alternative approaches that leveraged participants' self-assessment of their learning (JOLs). One conclusion from this analysis is that there seem to be measurable signals in the brain that index the quality of memory with higher fidelity than people's own introspective access.

We also observed that the use of fMRI measurements enabled discriminating between models that were equivalent when using behavioral data (recall or JOL) alone. Whereas the held-out performance of the two- and three-state models was the same for the *Recall* and *Recall+JOL* model variants, using fMRI data to inform the model estimation revealed an improvement for the three- compared to the two-state model. This result points to the ways in which joint modeling of behavioral and neural data can afford insights into cognitive dynamics that might not be available to researchers focusing on more restricted kinds of data.

Although the results are promising, our assumptions about the fMRI data at this stage are simplistic. For example, our model assumed that the distribution of fMRI signals was stable across time. However, it is well known that fMRI signals often show a pattern of *repetition suppression* [20] where the measured BOLD signal is systematically lower on subsequent presentations of an item. A more sophisticated analysis of the brain may lead to improvements in our models. Another particularly interesting direction is to attempt to model individual learner abilities (c.f., [42, 23]) on the basis of patterns of brain activity given the large variance in overall performance across participants (see Figure 1).

Modifications to the model structure might also improve predictions. As an example, in ongoing work we estimated the three-state *Recall+MRI* model but modeled the fMRI observations as arising from transitions between states rather than from the states themselves (i.e., each fMRI component has a distribution of activations associated with *staying* in a state and another distribution associated with *switching* between states). The three-state version of this *Recall+MRI-Transition* model yielded a held out AUC of 0.77 ($\pm$.02), which is our best performing model to date. This shows that there is certainly more signal we can exploit from the data by improving our generative model of the fMRI signal. Attempts to improve the fMRI modeling and explore different model structures are continuing.

We have also illustrated several ways in which this kind of simultaneous modeling approach might feedback to our understanding of the role of the brain in supporting learning and memory. Using a model-based regressor coding for the change in posterior probability of latent knowledge states, we identified a significant effect in a left anterior hippocampus region that is known to be involved in memory formation on the basis of past studies [13]. The similarities between this novel analysis approach and past cognitive neuroscience studies give converging evidence about the hypothesized role of these regions. We also used our estimates of the fMRI observation distributions to examine the relationship between fMRI activation arising from different neural components and the latent knowledge states instantiated in the model(s), which is a novel approach to understanding the way psychological mechanisms or processes may be implemented in the brain.

While we acknowledge the practical limitations of acquiring neuroimaging data in an educational setting – although advances in EEG technology and the established ability to measure subsequent memory signals with EEG may enable such use in restricted settings [18, 15] – overall we believe this work represents an encouraging first step for knowledge tracing approaches that utilize indirect neural information as opposed to explicit tests.

6. ACKNOWLEDGMENTS

This research was supported by NSF grant DRL-1631436 and seed funds from the NYU Dean for Science. We thank Mike Mozer for advice.

7. REFERENCES

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

[2] J. Anderson, S. Betts, J. Ferris, and J. Fincham. 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, 2010.

[3] J. Anderson, A. Corbett, K. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2):167–207, 1995.

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

[5] R. Atkinson. Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, 96:124–129, 1972.

[6] R. Baker, A. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 2008.

[7] R. Baker, A. Goldstein, and N. Heffernan. Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, 21:5–25, 2011.

[8] D. Bondareva, C. Conati, R. Feyzi-Behnagh, J. Harley, R. Azevedo, and F. Bouchet. 2013. *Artificial Intelligence in Education*, pages 229–238, 2013.

[9] M. W. Brown and J. P. Aggleton. Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2(1):51–61, January 2001.

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

[11] A. Corbett and J. Anderson. Knowledge tracking: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4:253–278, 1995.

[12] J. Danker, A. Tompary, and L. Davachi. Trial-by-trial hippocampal encoding activation predicts the fidelity of cortical reinstatement during subsequent retrieval. *Cerebral Cortex*, 27:3515–3524, 2017.

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

[14] L. Davachi, J. Mitchell, and A. Wagner. Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proc Natl Acad Sci U S A*, 100(4):2157–2162, 2003.

[15] S. Dikker, L. Wan, I. Davidesco, L. Kaggen, M. Oostrik, J. McClintock, J. Rowland, G. Michalareas, J. Van Bavel, M. Ding, and D. Poeppel. Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380, 2017.

[16] S. Doroudi and E. Brunskill. The misidentified identifiability problem of bayesian knowledge tracing. In *Proceedings of the 10th International Conference on Educational Data Mining*, 2017.

[17] J. Feng. *Essays on learning through practice*. PhD thesis, The University of Chicago, 2017.

[18] K. Fukuda and G. Woodman. Predicting and Improving Recognition Memory Using Multiple Electrophysiological Signals in Real Time. *Psychological science*, pages 0956797615578122–, 2015.

[19] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7(4):457–472, 11 1992.

[20] K. Grill-Spector, R. Henson, and A. Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Science*, 10(1):14–23, 2006.

[21] P. Grimaldi, M. Pyc, and K. Rawson. Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for lithuanian-english paired associates. *Behavior Research Methods*, 42:634–642, 2010.

[22] K. Koedinger, E. Brunskill, R. Baker, E. McLaughlin, and J. Stamper. New potentials for data-drive intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[23] R. Liu and K. R. Koedinger. Towards reliable and valid measurement of individualized student parameters. In X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, editors, *Proceedings of the 10th International Conference on Educational Data Mining*, pages 135–142, 2017.

[24] T. Nelson and J. Dulosky. When people's judgments of learning (jol) are extremely accurate at predicting subsequent recall: The delayed-jol effect. *Psychological Science*, 2:267–270, 1991.

[25] K. Paller, M. Kutas, and A. Mayes. Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography and clinical neurophysiology*, 67(4):360–71, 1987.

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

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

[28] C. Ranganath, M. Johnson, and M. D'Esposito. Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, 41(3):378–389, 2003.

[29] M. Rau and Z. Pardos. Adding eye-tracking aoi data to models of representation skills does not improve prediction accuracy. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 622–623, 2016.

[30] T. Sanquist, J. Rohrbaugh, K. Syndulko, and D. Lindsley. Electrocortical Signs of Levels of Processing: Perceptual Analysis and Recognition Memory. *Psychophysiology*, 17(6):568–576, 1980.

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

[32] Stan Development Team. PyStan: the python interface to Stan, 2017. Version 2.17.0.0.

[33] Stan Development Team. Stan modeling language users guide and reference manual, 2017. Version 2.17.0.0.

[34] C. Tenison, J. M. Fincham, and J. R. Anderson. Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive Psychology*, 87:1–28, 2016.

[35] B. Turner, B. Forstmann, E. Wagenmakers, S. Brown, P. Sederbefg, and M. Steyvers. A bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*, 72:193–206, 2013.

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

[37] B. van De Sande. Properties of the bayesian knowledge tracing model. *JEDM-Journal of Educational Data Mining*, pages 1–10, 2013.

[38] W. Van Der Linden. Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3):247–272, 2009.

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

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

[41] Y. Wang and N. Heffernan. Leveraging first response time into the knowledge tracing model. *Educational Data Mining*, pages 176–179, 2012.

[42] M. Yudelson, K. Koedinger, and G. Gordon. Individualized bayesian knowledge tracing models. *Artificial Intelligence in Education*, 2013.

[43] Q. Zhang, J. Anderson, and R. Kass. Consistency in brain activation predicts success in transfer. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX, 2016.