Neuropsychologia 50 (2012) 487–498

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Tracking problem solving by multivariate pattern analysis and Hidden Markov Model algorithms

John R. Anderson*

Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 5 April 2011
Received in revised form 30 June 2011
Accepted 20 July 2011
Available online 27 July 2011

Keywords:
Problem solving
Multi-voxel pattern analysis
Hidden Markov Models
Intelligent tutoring systems
Model discovery

ABSTRACT

Multivariate pattern analysis can be combined with Hidden Markov Model algorithms to track the second-by-second thinking as people solve complex problems. Two applications of this methodology are illustrated with a data set taken from children as they interacted with an intelligent tutoring system for algebra. The first "mind reading" application involves using fMRI activity to track what students are doing as they solve a sequence of algebra problems. The methodology achieves considerable accuracy at determining both what problem-solving step the students are taking and whether they are performing that step correctly. The second "model discovery" application involves using statistical model evaluation to determine how many substates are involved in performing a step of algebraic problem solving. This research indicates that different steps involve different numbers of substates and these substates are associated with different fluency in algebra problem solving.

1. Introduction

Psychology has always been challenged by the task of scaling its methods to the complexity of cognition. The methodology of choice for much of its history has been reaction time, which has shown considerable success in analyzing the structure of cognition in the subsecond range. Neuroimaging techniques, particularly ERP with its high temporal resolution, have also had success in that time range. But what happens when a participant is challenged by a difficult problem, thinks about it for a period measured in minutes, and announces the answer? Newell and Simon (1972), when faced with the challenge of understanding such problem solving, decided to tackle this in the most direct way possible and simply ask the participant to tell them what they were thinking. While verbal protocols have been subject to challenges (Nisbett & Wilson, 1977), this methodology has borne considerable fruit (Ericsson & Simon, 1993). Eye movements offer a less intrusive way of tracking thought and also have had some success (Salvucci & Anderson, 2001). However, when problems really get hard, participants will often want to shut up, close their eyes or look away, and think.

The goal here is not to criticize current methods for analyzing brief cognition or extended cognition, but rather to describe a new methodology for tracking the sequential structure of thought. This methodology combines Hidden Markov Models (HMM – Rabiner, 1989) with multi-voxel pattern analysis (MVPA) of fMRI data (e.g., Davatzikos et al., 2005; Haynes & Rees, 2005; Haynes et al., 2007; Hutchinson, Niculescu, Keller, Rustandi, & Mitchell, 2009; Mitchell et al., 2008; Norman, Polyn, Detre, & Haxby, 2006). The paper will illustrate the potential of the methodology with two applications involving an intelligent tutoring system.

1.1. Cognitive Tutors

At Carnegie Mellon University we have developed a successful approach to computerized instruction called *Cognitive Tutors* (Anderson, Corbett, Koedinger, & Pelletier, 1995). These tutors focus on the instruction of mathematics and are widely used. For instance, the Algebra Tutor (Ritter, Anderson, Koedinger, & Corbett, 2007) is currently deployed in over 2600 schools throughout the United States and interacts with approximately 500,000 students each year. They are called Cognitive Tutors because they are built on cognitive models that solve problems in the same way that students do. They individualize instruction by two processes called *model tracing* and *knowledge tracing*. Model tracing uses a model of students' problem solving to interpret their actions. It does this by finding a path of cognitive decisions that produces a match to the observed actions. Given such an interpretation, the tutoring system is able to provide real-time instruction individualized to where a student is in the problem. The second process, knowledge tracing, involves inferring what skills the student has mastered and then selecting new problems and instruction suited to that student's knowledge state.

* Tel.: +1 412 417 7008; fax: +1 412 268 2844.
E-mail address: ja@cmu.edu

0028-3932/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2011.07.025

While positive results have been reported for the tutoring systems, they are limited by two factors:

1. *Diagnosing what a student is thinking*. The only information available to a typical tutoring system comes from the actions that students take in the computer interface. The inference from such surface behavior to underlying thought is perilous. We have made some progress using brain imaging to diagnose the moment-by-moment changes in a student's mental state (Anderson, Betts, Ferris, & Fincham, 2010; Anderson, Betts, Ferris, & Fincham, 2011, in press-a). The first application in this paper will review one example of our work here. It will introduce the principles underlying the approach.
2. *Accuracy of the cognitive models*. The cognitive models in the tutors are quite crude relative to the complexity of the actual cognitive processes. Recently, I have been exploring whether this same approach can be used to refine these cognitive models. The second application in this paper will illustrate how this neuroimaging approach can be used to discriminate among alternative models and discover new models.

Both applications will use data from an experiment described in detail in Anderson et al. (2011, in press-a). That study followed 16 children going through a sequence of algebra problems with the assistance of the tutor. They worked with the tutor over 6 days (numbered 0–5) and were scanned on Day 1 and Day 5. The study used an experimental tutoring system described in Anderson (2007) and Brunstein, Betts, and Anderson (2009) that teaches a complete curriculum for solving linear equations based on the classic algebra text of Foerster (1990). The tutoring system has a minimalist design to facilitate experimental control and detailed data collection. Nonetheless, it has the basic components of a Cognitive Tutor: instruction when new material is introduced, help upon request, and error flagging during problem solving.

Students use a mouse for all tutor interactions – to select parts of the problem to operate on, to select operations from a menu, and to enter values from a numeric keypad. They go through cycles of four steps in solving an algebra problem. As a simple example consider the four steps in solving the equation "$x - 10 = 17$", illustrated in Fig. 1:

1. *Selecting a transformation*. In this case this involves selecting an operation called "Unwind" and indicating that it applies to the whole equation.
2. *Executing the transformation*. In this case this involves entering "$x = 10 + 17$" by clicking in a keypad.
3. *Selecting evaluation*. In this case this involves selecting "$10 + 17$" as the term to be evaluated.
4. *Executing the evaluation*. In this case this involves entering "27" as the result.

In the example in Fig. 1, solving the problem involves just one cycle of these four steps. More complex problems could involve many cycles of these four steps.

The 16 students did 9 blocks of problems on each day. In a block the children solved 2–7 problems and took anywhere from under 2 min to over 9 min. We collected whole-brain scans every 2 s. Because of problems with the scanner, a block was lost for 5 students on Day 1 and for 4 students on Day 5, leaving 139 blocks on Day 1 and 140 on Day 5. Altogether, they solved 727 problems on Day 1, taking 19,376 2-s scans and 742 problems on Day 5 taking 15,614 scans. Because some problems involve more than one cycle, there were 890 occurrences of the 4-Step cycles on Day 1 and 905 on Day 5. Table 1 gives statistics on durations and error rates for these steps, where an error is defined as selection of wrong part of the equation, selection of a wrong operation, or entry of an incorrect result. The four steps had rather different behavioral characteristics. The two execution steps took much longer than the selection steps while the transformation steps were more error prone than the evaluation steps. Students sped up more than 70% from Day 1 to Day 5 and their error rate almost dropped in half.

2. Application 1: Using brain imaging to track student problem solving

I will begin by describing our approach to the "mind-reading" task of tracking students as they are solving algebra problems in this experiment, focusing on interpreting the student on Day 5. To interpret the behavior of a particular student on Day 5, we combined information from other students on that day and data from the student on Day 1. This is similar to the development and application of Cognitive Tutors, which are deployed with statistics based on pilot students and, as a particular student progresses through the curriculum, they build up a model of that particular student. Our brain imaging data come from *blocks*, which are sequences of about 6 problems that last about 3 min. We used the brain imaging data to determine what problem a student is working on within a block and what step the student is performing within a problem. Determining where the student is will be referred to as the *segmentation* goal. We also used the imaging data to determine whether that step is being performed correctly. This goal will be referred to as the *diagnosis* goal. We chose these goals because there is a hard definition of ground truth here, namely the computer logs of students' progress through these problems. We have developed a novel synthesis of three well-developed methodologies for following the mental states of students, described below:

2.1. Component 1: Hidden Markov Models (HMMs)

We used HMMs to represent the students' non-deterministic progress through the problems. Fig. 2 illustrates part of the HMM state structure used in this application. The figure shows the state structure for a fragment of a block that involves finishing a prior problem, transitioning to a rest state, stepping through one cycle of four steps to solve a problem, and returning to a rest state before the next problem. Each of the steps in solving the problem can be performed correctly or incorrectly. A block of problems was represented with a sequence of separate states for each problem in a block. A typical block of problems might consist of about 50 such states.

HMMs offer a powerful way to represent a student model because of the efficient algorithms for assigning probabilities to different possible state sequences. The critical feature of HMMs is their Markov property that the future course of problem solving only depends on the current state and not past history. Models in Cognitive Science are not typically cast in a way that obviously satisfies this Markov property, but nonetheless we have found ways to convert typical cognitive models into Markov state structures (see also Weaver, 2008). In the current example we set the probabilities of state transitions for a particular student on the basis of the behavioral data of other students (see Component 3). In addition to these transition probabilities, the HMM algorithms use the conditional probabilities of the observed data given different states to help determine the sequence of states. In our case these are the conditional probabilities of different brain imaging patterns (see Component 2). Different HMM algorithms (see Rabiner, 1989) can find the probability that a student is in a particular state during any scan (suitable for model tracing), the most probable interpretation of a block of scans (suitable for knowledge tracing), or the probability of a set of data given a particular model (suitable for model evaluation).

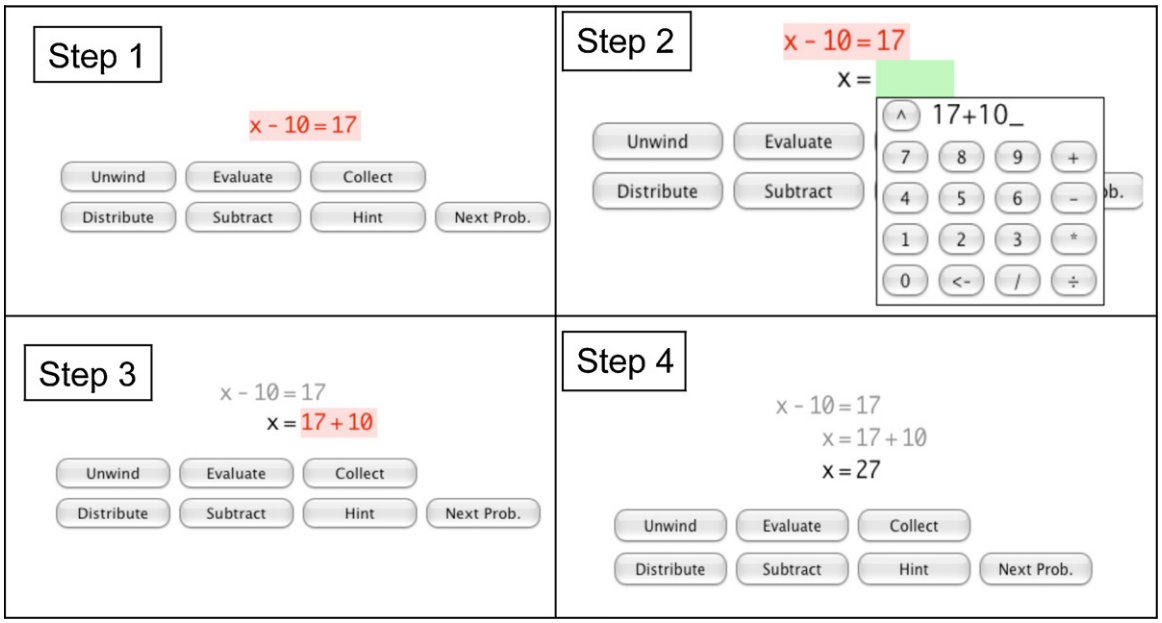

Fig. 1. Each panel illustrates one of the four steps in a problem-solving cycle with the tutor. The subpanels show the states of the tutor within a step. Each step starts with the last state of the previous step. The first panel starts with the initial equation $x - 10 = 17$. *Step 1*: The student selects a transformation to perform on this equation by clicking on the two sides of the equation (resulting in the red highlighting) and choosing "Unwind" from the menu below. *Step 2*: The student expresses the result of the transformation by selecting the green box and entering $17 + 10$. This results in the transformed equation $x = 17 + 10$. *Step 3*: The student specifies that $17 + 10$ is to be evaluated by clicking on this expression (resulting in the red highlighting) and selecting "Evaluate" from the menu below. *Step 4*: The student specifies the result of the evaluation by entering 27. This creates the final answer $x = 27$, which is displayed here.

Table 1
Behavioral statistics for the four steps.

	Day 1			Day 5		
	Mean scans	Stdev scans	Percent errors	Mean scans	Stdev scans	Percent errors
1. Select transformation	2.31	2.31	20.80%	1.47	0.94	10.50%
2. Execute transformation	7.10	5.05	20.60%	4.62	2.85	10.40%
3. Select evaluation	1.24	0.85	3.30%	0.76	0.64	0.80%
4. Execute evaluation	3.68	2.58	6.60%	2.99	2.37	6.60%

In many applications such as the current one, we challenge the HMM algorithms to infer the boundaries between states as well as the states themselves. This requires using a special variant of HMM called semi-Markov models because the duration in a state is variable. The same HMM algorithms can be extended to semi-Markov models (Yu, 2010). We also use the behavioral data of other students (Component 3) to infer a probability distribution of durations in the different states.

2.2. Component 2. Multi-voxel pattern analysis

In applications like this tutoring study, there are tens of thousands of whole-brain images from different students going through different problems. We train multi-voxel classifiers to associate different brain patterns with different mental states. These classifiers deliver the conditional probabilities that a given brain pattern comes from particular states, which is what is needed for the logic of a HMM. There are a number of distinctive features of our approach to pattern classification:

a. *State abstraction.* The number of states in the model can be large. The current example involves about 50 states for each block of problems, and each block of problems is given its own set of states. Even though there are tens of thousands of images, there is not enough data to recognize each state. Therefore, a necessary step is to find some abstraction of the specific states into a smaller number of states. This application used 9 abstract states – one corresponded to the rest period between problems and the

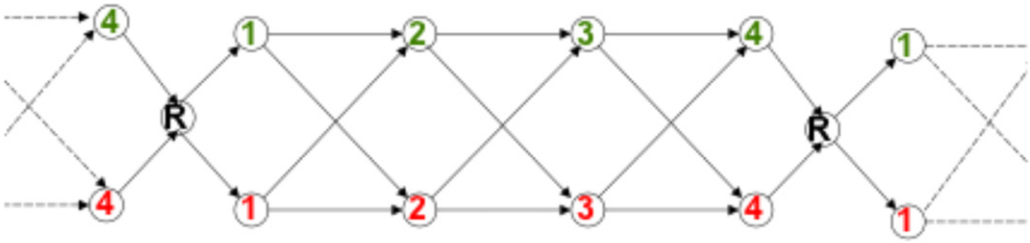

Fig. 2. The behavioral model as a semi-Markov process. States correspond to steps (green correct, red incorrect) and rest period (R). While correct and incorrect states are numbered identically to indicate the step, each state is distinct within the HMM.

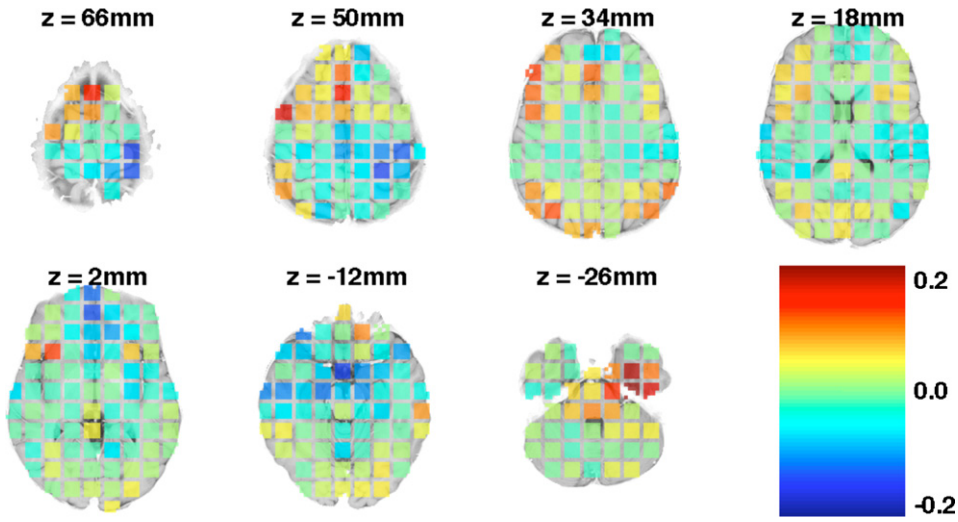

Fig. 3. Differences between weights associated with error steps and correct steps. The 408 ROIs were created by evenly distributing $4 \times 4 \times 4$ voxel cubes over the 34 slices of the 64×64 acquisition matrix. Between region spacing was 1 voxel in the *x*- and *y*-directions in the axial plane, and one slice in the *z*-direction. The final set of regions was acquired by applying a mask of the structural reference brain and excluding regions where less than 70% of the region's original 64 voxels survived.

other 8 were the 4 basic steps (see Fig. 1) performed correctly or incorrectly.

b. *Coarse, whole brain activation patterns*. While we have some success using specific predefined regions in some of our other work, we throw away information in such a complex task if we do not use the activation over the full brain. We take the brain activity over about 400 relatively large regions, each a little more than a centimeter cubed. Using smaller regions does not yield much more information and the sheer number of such regions leads to serious problems of overfitting. Fig. 3 illustrates the regions and their weights of association with being in an error state (averaged over the four steps). Warmer colors indicate regions that are more strongly associated with error states.

c. *Linear Discriminant Analysis (LDA)*. We have examined a number of methods sometimes associated with improved performance in the literature such as support vector machines (SVMs) with radial basis functions and other kernels. However we get the best results with LDA. Hsu, Chang, and Lin (2009) note linear classifiers are much more efficient and do not have accuracy disadvantages relative to SVMs when the number of features and instances are large. This is our situation and LDA produces the conditional probabilities that are required by the HMM

d. *Scan lag*. We train the classifier to associate states with the brain activity that follows 4–5 s after the student is in that state. This delay gives us optimal performance, which this is not surprising given the lag of the hemodynamic response. The current example uses the activity 2 scans (4 s) later to classify the state the student was in during a the time of particular scan. We have tried using multiple scans rather than a single scan for classification, but this typically results in overfitting the data.

e. *Merging group and individual data*. The best results come from combining imaging patterns both from other students and from the particular student (in this case from Day 1) to train the classifier. The data from the specific student are useful because each student's activation patterns have their own idiosyncrasies. However, there is not enough data from individual students to reliably train student-specific classifiers. In the current example we used 1/3 weightings of the Day 1 imaging data from a particular student, the Day 1 imaging data from the other 15 students, and the Day 5 imaging data from the other students.

2.3. Component 3. Student modeling

The probabilities of various state transitions and durations of residence within a state can be estimated for a student from other students and past behavior of this student. In contrast to multi-voxel pattern analysis, where abstract states are needed to avoid overfitting, we use no state abstraction with this behavioral data. Indeed, the best results come from using estimates of durations in specific states and transition probabilities between these states. So, each step for each problem is its own state. This allows us to capture the large differences in difficulty among individual problems. In the current example we obtained, from other students, statistics on state durations and error probabilities for each step of each problem. We also estimated for that student their mean latencies and errors for a step averaged over problems. We used a 2/3 weighting of the problem-specific measures from other students and a 1/3 weighting of the student-specific measures, to generate an estimate of the latency and accuracy of this student for this step of a particular problem. Fig. 4 shows the distribution of times aggregated into three categories of problem difficulty. Note the continuous latency distributions fit to these data, which allow us to calculate the probability of any duration for any step of any problem.

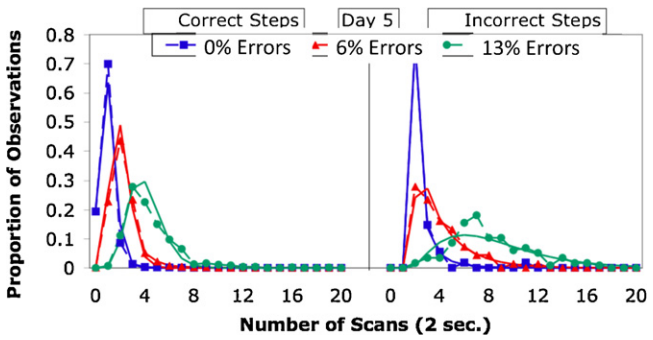

Fig. 4. Distribution of correct and error step times for a student on Day 5 as a function of the difficulty other students experienced with that step. The points connected by dotted lines are the proportions of observations with different number of scans. The smooth lines are fitted gamma functions.

2.4. Application and evaluation

With this framework we can calculate the probability of any interpretation of a sequence of scans. For example, consider a situation where a student is solving a sequence of problems that involves m scans, going through r steps, which are a subset of s states. In the current experiment, for a specific block of problems and a particular student, m would be on the order of 150 scans, r would be on the order of 30 steps, and s on the order of 50 states.[1] An interpretation, I, of the m scans is an assignment of the scans to a subset of the s states. Using a naïve Bayes rule, the probability of any such interpretation I can be calculated as the product of prior probability determined by the behavioral model and the conditional probabilities of the fMRI signals given the assignment of scans to states:

$$p(I|fMRI) \propto \left[S_r(a_r) \prod_{k=1}^{r-1} p_k(a_k) * t_{k,k+1} \right] * \left[\prod_{j=3}^{m+2} p(fMRI_j|I) \right]$$

The first term in the product is the prior probability (based on the behavioral model) and the second term involves the conditional probabilities (based on the LDA of the imaging data). The term $p_k(a_k)$ in the prior probability is the probability that the kth interval is of length a_k and $S_r(a_r)$ is the probability of the rth interval surviving at least as long as a_r. The term $t_{k,k+1}$ is the probability of transitioning from state k to k+1. The second term contains $p(fMRI_j|I)$, which are the probability density values for the fMRI signal on scan j+2 given I's assignment of scan j to a state. This is what the fMRI conditional probabilities provide. While the number of such interpretations is astronomical, HMM algorithms enable efficient calculation using dynamic programming techniques. This can be done either for real-time interpretation (finding the most probable state for the current scan), post-block reconstruction (the most probable sequence of all the scans in a block), and model selection (the probability of the imaging data given a model). This application will use the Viterbi algorithm (Rabiner, 1989).

Interpreting a student can be operationalized as identifying the state in the HMM diagram. The computer logs of student behavior provide a definition of ground truth. We performed separate evaluations of the segmentation goal (identifying what step a student is performing) and the diagnosis goal (determining whether that step is being performed correctly). Fig. 5 illustrates the algorithm's performance on these two dimensions[2]:

1. *Segmentation goal.* Fig. 5a illustrates the success of the algorithm at identifying where the student is in a block of problems. The algorithm assigns each scan to a step of some problem and the figure plots the mean difference between the assigned step and the true step. The model correctly classifies most of the scans and is off by just a single step on most of the remaining cases. Most of these errors are right at the boundaries between steps. The figure also shows how well the algorithm can do using just statistics based on fMRI data or just behavioral model, and both.[3] Performance is much better using behavioral data and fMRI data together than using either alone. This illustrates the boosting effect one gets by combining behavioral and fMRI data.
2. *Diagnosis goal.* Fig. 5b illustrates the success of the approach on the diagnosis goal. One can vary the criterion for classifying a step as an error and so generate a curve giving the probability of a hit (classifying an error step as an error) as a function of the probability of false alarming (classifying a correct step as an error). A measure of classification accuracy is the area under the curve, which is 0.5 for chance classification and 1.0 for perfect classification. The level of performance in the figure is a high 0.946.

To summarize this application, the combination of MVPA and a behavioral model of the student can yield a fairly accurate diagnosis of where a student is in problem-solving episodes that last many minutes. Moreover, prediction accuracy using both information sources was substantially greater than using either source alone. The performance in Fig. 5 should not be taken as the limit of what can be achieved. Performance could be improved by enhancing the imaging data, by adding other data sources, or by improving the behavioral model. These are topics of our current efforts.

3. Application 2: Using brain imaging to build and evaluate cognitive models

The model whose state space is illustrated in Fig. 2 is typical of the models built into Cognitive Tutors. Their temporal grain size corresponds to the student actions on which feedback can be meaningfully generated. These models attempt to represent all the correct and incorrect transitions that the student might be making. As illustrated in Fig. 4, these states correspond to highly variable periods of time that take many seconds. Perhaps the student is actually going through multiple mental substates in the period of time occupied by a step. This second application will illustrate how model evaluation techniques based on HMMs and can determine if a step does consist of substates. While this example continues with the tutoring domain, it illustrates methods that potentially could generalize to many model evaluation and discovery problems.

The specific goal will be to decide whether a step is best described as a specific sequence of 1, 2, 3, or 4 substates. Unlike the previous example, there are not computer logs to assess the ground truth – we don't know how many substates there are or where they begin and end. Rather, goodness of fit to the data will be used to infer the number of substates and their segmentation. We will assume that the duration t_i of residence in any substate i is variable, described by a 2 parameter gamma function $g(t, v_i, a_i)$, where v_i and a_i are the parameters of the gamma function for substate i. The durations were discretized such that each 2-s scan gets associated with a single substate. Gamma functions were similarly used in the previous example to describe the duration in a step (see Fig. 4). Each substate i will have its own distribution of times and its own pattern of activation over the 408 voxels. This means each substate will have 410 parameters associated with it (2 for the latency distribution and 408 for the activation values).

The Appendix A describes how predictions are derived from the different substate models for a step and how parameters are estimated. This results in a calculation of the likelihood of the data under the various models. Because the models are nearly nested,[4] it is to be expected that the model with more substates will have a

[1] As noted earlier these approximately 50 states are abstracted to 9 for use of the imaging data, but are represented distinctly in the HMM so that state-specific behavioral data can be combined with the more abstract imaging data.

[2] Visit http://act-r.psy.cmu.edu/actrnews/index.php?id=34 to see a demonstration of the performance of our system on one block of algebra problems, predicting the actual mouse clicks in the problem.

[3] Performance with just fMRI is achieved by making all transitions and durations in the HMM equiprobable. Performance with just behavioral model is achieved by making the conditional probabilities of the fMRI data the same for all states.

[4] Pure nesting would mean that the more complex model is guaranteed to produce at least a good a fit (larger log likelihood) as the simpler model. The nesting is not perfect because the discrete approximation to the gamma removes the guarantee that the more complex model will provide a better fit to the distribution of scan lengths.

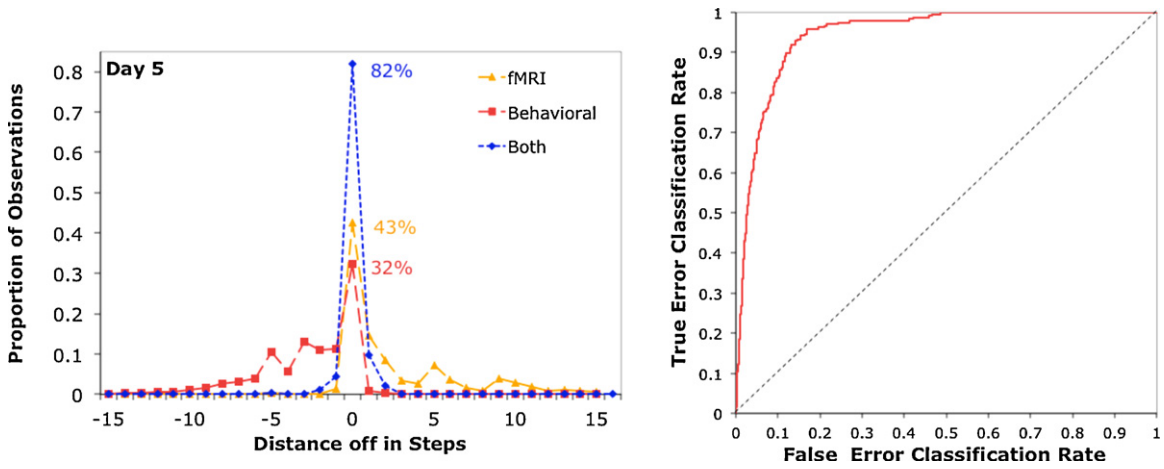

Fig. 5. (a) Performance on segmentation. (b) Performance on diagnosis.

Table 2
Comparison of models for the different steps.

	Step 1	Step 2	Step 3	Step 4
(a) Chi square differences				
2 over 1	1093.5	3491.5	424.9	2186
3 over 2	883.7	2313	198.8	1086.8
4 over 3	518.6	2007.6	−51.8	918.5
(b) Chi square differences with permuted data				
2 over 1	599 ± 243	690 ± 76	531 ± 204	633 ± 23
3 over 2	552 ± 112	608 ± 341	193 ± 155	600 ± 242
4 over 3	225 ± 227	1129 ± 610	−10 ± 260	508 ± 314
(c) Maximum chi square differences with permuted data				
2 over 1	1763.0	773.1	1403.0	687.9
3 over 2	997.3	1580.8	464.3	1170.0
4 over 3	739.6	2871.1	731.7	1230.1

higher likelihood. Under the null hypothesis, 2 times the difference in likelihood difference between a n+1 substate model and a n substate model should be distributed as a chi-square with 410 degrees of freedom to reflect the extra parameters. Table 2a gives the actual chi-square gains associated with the more complex models.[5] The simpler model can be rejected in favor of the more complex model if the chi-square is greater than chance, indicating that the likelihood gain is not just a matter of fitting noise. As points of reference, the critical values for a chi-square with 410 degrees of freedom is 458 for $p = .05$ and 525 for $p = .0001$. Given these thresholds, the conclusions are pretty clear: Steps 2 and 4 are best fit with four substates, Step 3 with one substate, and Step 1 is between the .05 and .0001 threshold for deciding between 3 and 4 substates.

While the 4-substate model provides better fits to the data for all but Step 3, the interpretation of this result is not clear. The interesting interpretation of this result would be that students are going through a sequence of four distinct mental substates. However, other uninteresting factors might be producing these statistically better fits. For instance, it may reflect the fact that the durations in a substate do not perfectly fit a gamma distribution or the fact that the distribution of feature values is not perfectly normal. Estimating more substates might simply reflect using multiple distributions to better fit the true distributions.[6]

The critical feature of the interesting interpretation is that the sequence of scans within a step reflects a sequence of statistically distinguishable substates. To test whether this was true the scans for all the instances of a step were permuted. This assigns a random set of scans to any step instance. The scans assigned to a step are likely to come from various other instances of the same step and are likely occur in a different order than their true order. The same model fitting effort was applied to the permuted data as had been done on the real data, resulting in simulated chi-square differences. If the advantage of the more complex model actually reflects going through an additional substate in a specific order, then the chi-square gain for this random assignment should be less that the gain for the true assignment. However, this should not be the case if the more complex model is simply better at fitting the distribution of step times or creating a mixture that better matches the distribution of feature values.

Any single random permutation might not produce chi-square gains as large as the true gains because of chance. Therefore, each step was permuted 100 times and the model was fit to each simulated data set. Table 2b shows the means and standard deviations of the chi-square gains for the fits to the simulated data. To confidently accept a more complex model over a simpler model, the chi-square gain for the true data should be greater than what is observed with these random permutations. Table 2c shows the largest values obtained from the 100 random permutations. Comparing these simulated gains to the true gains leads to the following conclusions:

Step 1 involves selecting a part of the expression to operate on and identifying the operation. The evidence is not totally clear for discriminating among 1, 2 or 3 substates. We observed four simulated gains for the 2 versus 1 substate contrast that were greater than the true gain and two simulated values greater for the 3 versus 2 contrast. In contrast to these two comparisons, the gain for 4 substates over 3 is just 1.30 standard deviations larger than the mean simulated gain and 11 simulated gains were greater than the true gain. As something of a compromise, we chose the 2-substate model, which is illustrated in Fig. 6.

Step 2 is the most critical step where the student specifies the change that the transformation produces in the algebraic expression. There is strong evidence for a 3-substate model but not for a 4-substate model. We observed no simulated gains greater than

[5] The negative chi-square gain for the 4-substate model for Step 3 reflects the fact that the models are not perfectly nested and it is difficult to combine many discrete distributions given the brief duration of this step.

[6] For sake of brevity, we are skipping over an assessment of these models in terms of their BIC (Kass & Raftery, 1995) scores, which penalizes the more complex models for their added complexity. A straightforward BIC complexity penalty does not guard against the uninteresting interpretations of a better fit just described. The analysis described below that does guard against these uninteresting interpretations.

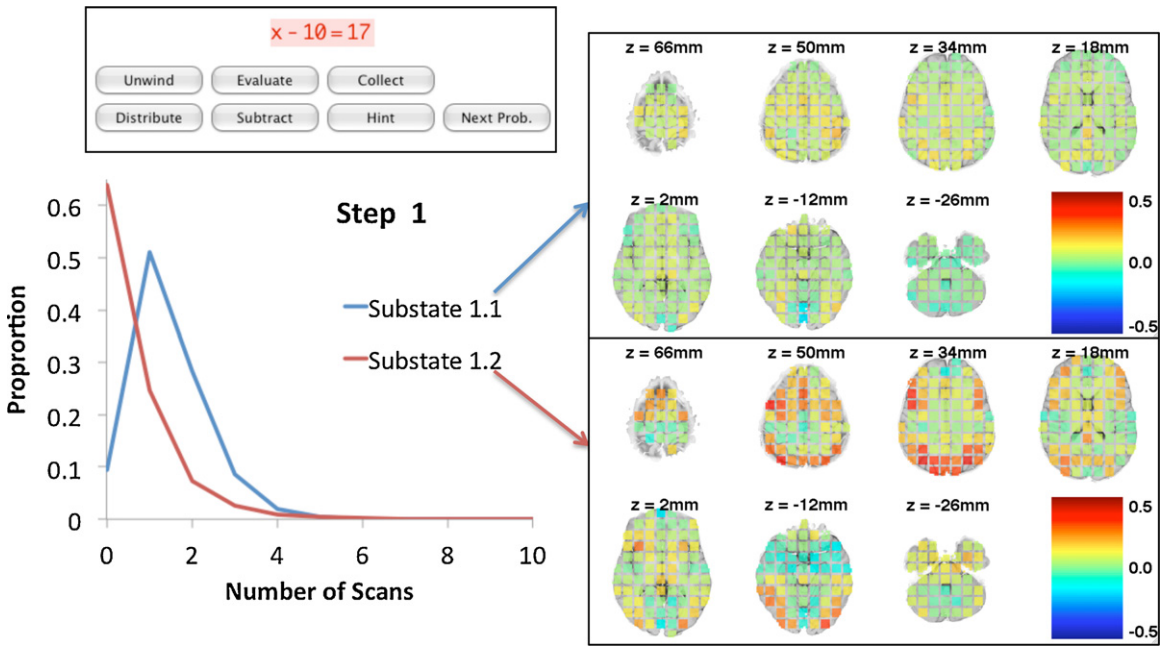

Fig. 6. Step 1: Distribution of lengths of the two substates and their activation patterns. The activation values are *z*-scores for percent deviation from baseline.

the true gains for 2 substates over 1 or for 3 substates over 2. The gain for 4 substates over 3 is just 1.44 standard deviations larger than the mean. Fig. 7 illustrates the parameter values for the 3-substate model.

Step 3 is the simplest step of identifying an arithmetic expression to be evaluated. The evidence clearly points to a 1-substate model since the mean simulated gains are all greater than the true gains. Fig. 8 illustrates the value for the 1-substate model.

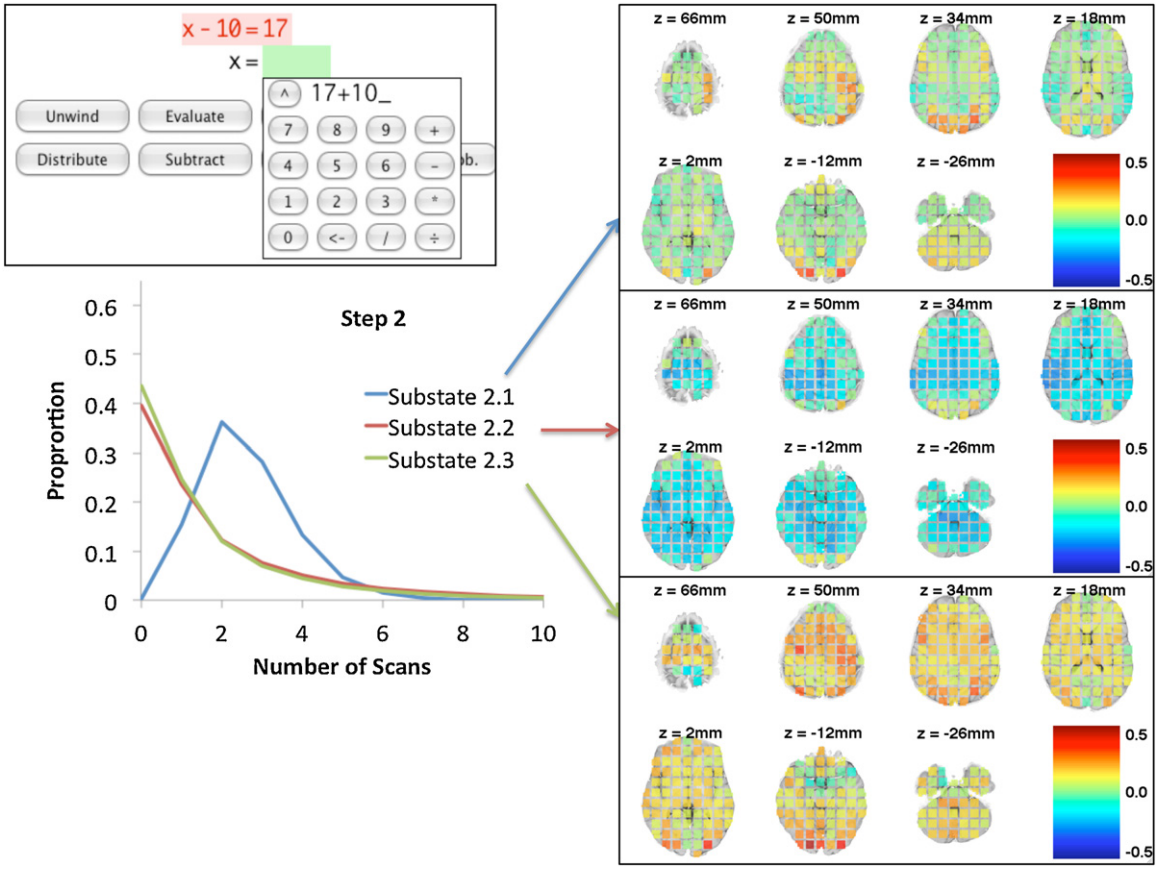

Fig. 7. Step 2: Distribution of lengths of the three substates and their activation patterns. The activation values are *z*-scores for percent deviation from baseline.

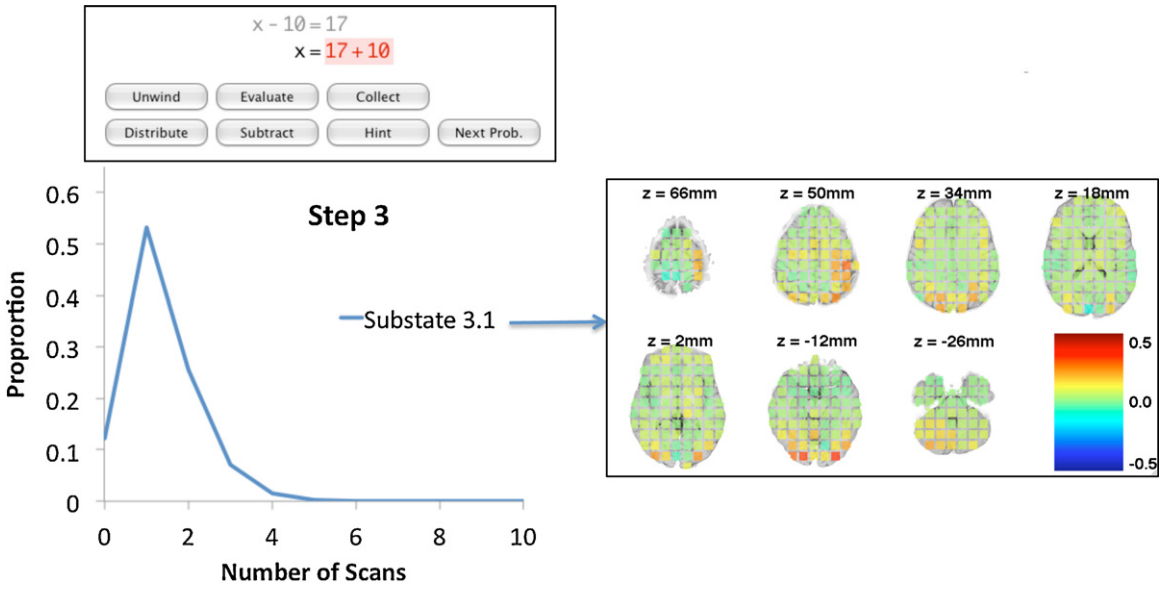

Fig. 8. Step 3: Distribution of lengths of the one substates and its activation pattern. The activation values are *z*-scores for percent deviation from baseline.

Step 4 involves the mental evaluation of the arithmetic expression. The evidence is very strong for a 2-substate model but marginal for a 3-substate model. The true gain for 3 versus 2 substates is only 2.01 standard deviations larger than the simulated mean and we observed two simulated gains greater than the true gain. Fig. 9 illustrates the fit for the 2-substate model.

Figs. 6–9 illustrate the substate models that seemed best. Each step has a first substate with a distribution of number of scans that includes very few 0 scans (i.e., this substate was seldom skipped) and peaks at 1 or 2 scans. The other substates all have peaks at 0 scans (i.e., they are often skipped) and long tails. Thus, the substates appear to divide into nearly obligatory substates (substates 1.1, 2.1, 3.1, and 4.1) and optional substates (substates 1.2, 2.2, 2.3, and 4.2).

The HMM algorithm allows us to estimate the time spent in each substate during each step. Fig. 10 shows the mean times as a function of day and whether an error was made. Generally and not surprisingly, students tend to spend more time in the substates when there is an error and less time on Day 5 when they have learned more. However, the different substates appear to display different patterns of latencies. To explore this we performed a hierarchical clustering of the 8 substates using the Euclidean dis-

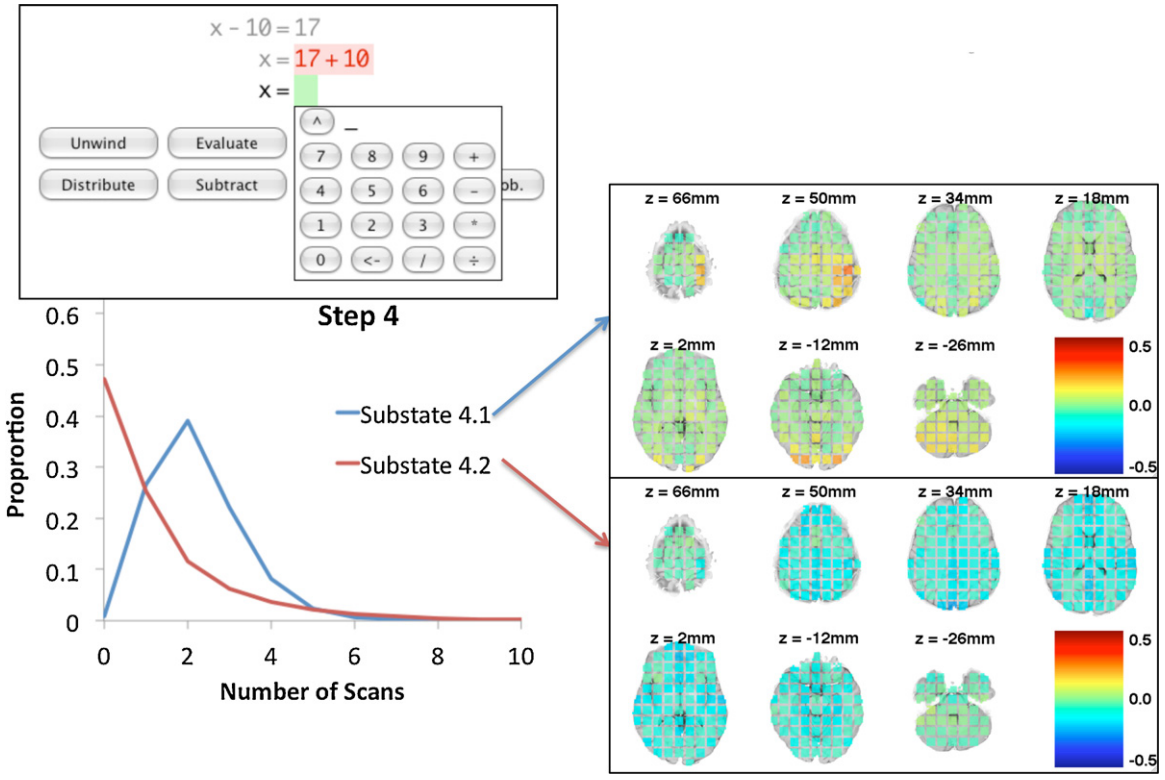

Fig. 9. Step 4: Distribution of lengths of two substates and their activation patterns. The activation values are *z*-scores for percent deviation from baseline.

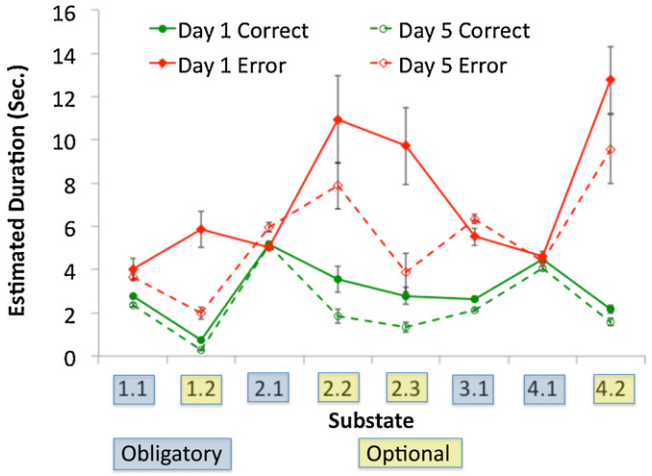

Fig. 10. Mean time in a substate as a function of day and whether an error was made on that step. Means are calculated separately for each student and averaged. The standard errors displayed are estimated from the student means.

tances between the mean times of the 4 conditions in each substate. Fig. 11a shows the resulting tree. The four obligatory substates cluster together and separate from the optional substates. Separate analyses of variance show that none of the obligatory substates show a significant effect of practice (Day 1 versus Day 5 – see Fig. 10) whereas all of the optional substates do.[7] Optional substates 1.2 and 2.3 cluster separately from 2.2 and 4.2 because they are the only substates to show a significant interaction ($F(1,15)$'s of 13.73 and 8.41, p's < .01 and .005, respectively) between day and correctness. The interaction occurs because more than twice as much time is spent in these substates on Day 1 when there is an error than any other circumstance.

Fig. 11b shows the results of a clustering analysis using distance between the activation patterns of these substates (i.e. distances between the 408 values – see Figs. 6–9). It reveals a similar pattern–optional substates 2.2 and 4.2 cluster together because they show least activation, the obligatory substates cluster together with moderate activation, and optional substates 1.2 and 2.3 cluster together with the most activation. In summary, three different measures suggest the same organization of the substates – distribution of number of scans, effect of condition on mean time, and the voxel activation patterns.

While variation in overall activation seemed to be the basis for the organization in Fig. 11b, it would be interesting to know what regions contribute most to this variation. Fig. 12 shows the standard deviation in activation values for the 408 regions. Areas that tend to vary substantially across substates include the right dorsolateral prefrontal cortex, left motor area, bilateral parietal cortex, and visual areas. The visual and motor areas reflect the fact that algebraic symbol manipulation has fundamental perceptual-motor components (Goldstone, Landy, & Son, 2010) even in this interface. The parietal areas overlap with the regions that activate in other studies of basic arithmetic (Dehaene, Piazza, Pinel, & Cohen, 2003) and algebra (Anderson, 2005). Right dorsolateral prefrontal regions have been reported to be more active in children during the performance of arithmetic tasks (e.g., Ansari & Dhital, 2006; Rivera, Reiss, Eckert, & Menon, 2005).

To summarize this example, the combination of MVPA and HMM model evaluation can serve to select among models for a task. In this case the model selection indicates how many substates to associate with particular steps. Further, these different substates appear to have different relationships to performance of the task. The overall mean duration in the obligatory substates is 4.2 s and this varies relatively little with step or condition. These would appear to reflect the constant aspects of performing the algebraic operations. The optional substates are much more engaged when an error is made. Time in these optional substates shows a significant decrease with practice unlike the obligatory substates. The four optional substates split into two subgroups – two (1.2 and 2.3) that involve heighted engagement of algebraic regions and two (2.2 and 4.2) that show decreased engagement of these regions. It remains to be determined what relationship these substates might have to the actual mastery of algebra.

4. Applications to tutor development

This methodology is appropriate to a wide range of mental tracking tasks and model evaluation. Nonetheless, this paper will continue its focus on applications to tutoring. Perhaps the first question in the mind of the reader is how could this methodology actually be applied to a Cognitive Tutor, given that regular instruction in an fMRI scanner is impractical. The potential application is not to delivery of instruction (which will still take place in classrooms and at home on a computer), but rather to the development of the instruction. One could evaluate different tutoring designs by seeing their consequences for the mental states of students, as illustrated in the first application. Additionally, one could improve upon the cognitive models in these tutors as illustrated in the second application.

With respect to the first goal of tracking student thought, the first application, while perhaps impressive, was rather useless. We already had computer logs that indicated the segmentation of performance into steps and whether these steps were being performed correctly. This provided a solid definition of ground truth. On the other hand, the methodology might be quite useful in discriminating among states within a period of time when there are not behavioral markers. For instance, one might consider diagnosing when an optional substate occurs. It may turn out that the highly engaged optional substates (1.2 and 2.3) reflect an effort to understand problematic material whereas the low engaged substates (2.2 and 4.2) may reflect a zoning out. Progress is being made on developing behavioral assessments of such mental states within the context of intelligent tutoring systems (e.g., Graesser et al., 2008). One could investigate whether these behavioral assessments correlated with the substates. If one could confidently assign such an interpretation to these substates, one could then use them to assess different tutoring systems by whether they produce engagement versus zoning out.

With respect to model evaluation, the sluggish nature of the hemodynamic response probably places a limit on how refined of a model one can identify. In this regard, note that Step 2, which produced the most substates (three), was the longest while Step 3, which produced just one substate, was the shortest. This reinforces the point made in the introduction that these HMM methods applied to fMRI are probably most appropriate for the long episodes that are typical of complex problem solving.

There have been some successful efforts to evaluate more detailed models of complex sequential behavior (Borst, Taatgen, Stocco, & Van Rijn, 2010) including student performance with this tutor (Anderson, Betts, Ferris, & Fincham, in press-b). These efforts have been based in the ACT-R architecture (Anderson, 2007) which makes predictions about mental events as brief as 50 ms. These cognitive modeling approaches have been more traditional in their use of fMRI data by identifying specific regions of interest and comparing model predictions against these regions, or, alternatively,

[7] All but substates 2.1 and 4.1 show significantly longer times when an error is made.

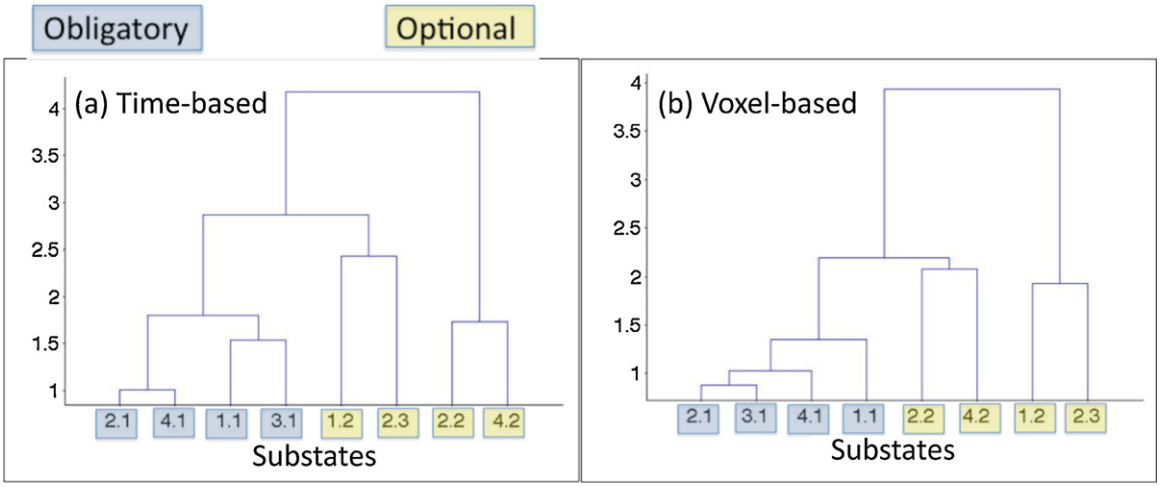

Fig. 11. Hierarchical clustering of substates: (a) Based on Euclidean distance between the 4 mean times associated with a substate (see Fig. 10); (b) Based on Euclidean distance between the 408 voxel values m associated with a substate (see Figs. 5–9).

creating traditional SPM design matrices (Friston, 2006) in order to detect brain regions associated with aspects of the model. These models make predictions about the timing of specific mental steps such as retrieval of a relevant arithmetic facts and much of the model fitting is concerned with how such times vary as a function of condition. For instance, Anderson et al. (in press-b) modeled the improvement from Day 1 to Day 5 in terms of less time spent retrieving algebraic knowledge.

There are two reasons why we have chosen instead this MVPA approach for an application like tutor development. First, the detailed factors that affect these model fits, such as retrieval time, are not particularly critical to instructional decisions. Rather the critical information for tutoring concerns mental states that last for longer periods. The methods described in this paper are suited for this temporal grain size. Second, one loses discriminative power by looking at only a subset of brain activity in specific regions of interest (although perhaps not theoretical power). For instance, in the first mind-reading application we tried only using specific regions with known relevance to algebraic problem solving, but the level of success was substantially reduced. While there are theoretical reasons for wanting to understand specific regions and focusing just on their activity, this can be at cross-purposes with the goal of achieving the most accurate diagnosis of mental state.

However, these two approaches should be complementary and could feed the development of each other. The patterns of brain activity displayed in this paper are sensible – for instance, Fig. 3 shows that algebraic errors are associated with strong weights in the region of the anterior cingulate, which would be expected in many theories of the function of this region (e.g. Botvinick, Braver, Carter, Barch, & Cohen, 2001; Falkenstein, Hohnbein, & Hoorman, 1995; Sohn et al., 2007; Yeung, Botvinick, & Cohen, 2004). Similarly, Fig. 12 revealed that a considerable portion of the variance between substates is associated with the posterior parietal cortex, which we have found critically engaged in algebra problem solving. Having strong theoretical constraints might guide appropriate feature selection and help deal with the problem of overfitting that is ever present in such MVPA applications.

Understanding the full pattern of activity engaged by different mental substates might guide the development of cognitive architectures like ACT-R. For instance, while ACT-R has had considerable success in modeling routine tasks including routine algebraic manipulation, it cannot really explain the much of the higher-order reasoning and metacognition that is critical to mastery of algebra (e.g. Anderson et al., 2011, in press-a). By identifying the sequential structure of the brain activation that is engaged by complex

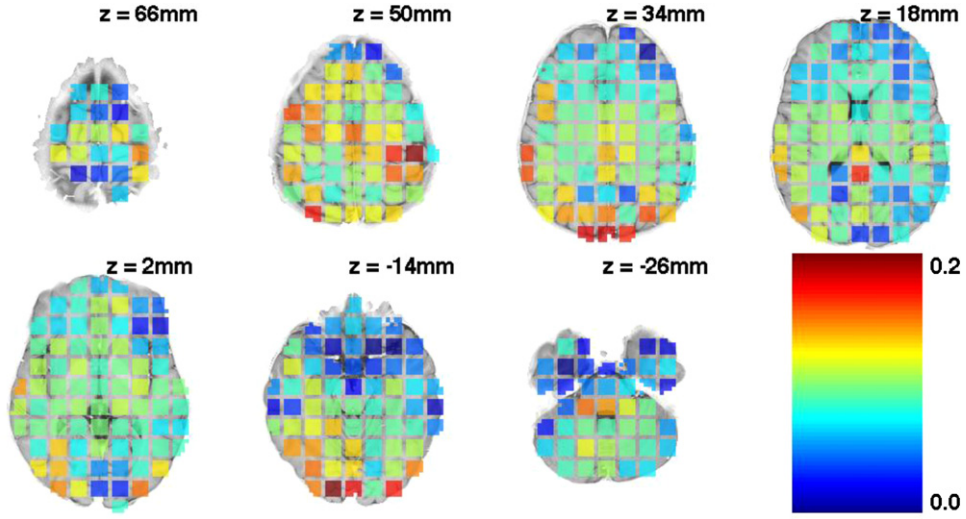

Fig. 12. Standard deviations of voxel values associated with the 8 substates.

problems we hope to be able to get clues to additional components that need to be added to the architecture.

Acknowledgements

This research was supported by a James S. McDonnell Scholar Award. I would like to thank Sam Wintermute for his comments on the paper.

Appendix A. Modeling details in Application 2

The probability $G_i(k)$ that substate i takes k scans is:

$$G(k, v_i, a_i) = \int_{2k-1}^{2k+1} g(t, v_i, a_i)dt$$

where v_i and a_i are the parameters of the gamma function ($2k$ in integral because each scan is 2 s). An interpretation I of a step is a sequence of n substates i whose lengths k_i sum to the number of scans in that step. The probability of such an interpretation I is:

$$P(I) = \prod_{i=1}^{n} \left(G(k_i, v_i, a_i) \prod_{j=1}^{k_i} p_i(fMRI_j) \right)$$

where the probability of k_i scans in substate i is multiplied by the probability of the images associated with those scans (which, as in the previous application, are at lag 2 scans). As in our use of LDA in example 1, assume that the distribution of the 408 voxels values in a substate is multivariate normal. To avoid dealing with the correlation structure these values, a principal component analysis was performed of the values in a substate to get 408 orthogonal values f_{jm} for each scan j. The probability of these values for substate i is given by a product of normal densities for these principal components[8]:

$$P_i(fMRI_j) = \prod_{m=1}^{408} N(f_{jm}, \mu_m, \sigma_m)$$

This requires estimating 408 means μ_{im} for each substate i. The standard deviations σ_m are calculated as the square roots of the eigenvalues of the principal component analysis and do not have to be estimated for the substates.

The parameters were estimated using expectation maximization as described in Rabiner (1989). There are many possible ways to break up k scans in a step into substates of k_i scans. HMMs can combine these to determine the probability that any scan is in any substate, which is critical to the re-estimation step in expectation maximization. This allows iterative calculation of a set of parameters that maximizes the overall probability of the data. This probability is what is critical in discriminating among models of different complexity. This overall probability can be calculated from the total probability of each step, which is summed over all possible interpretations I_{Step} of the steps:

$$Pr(Step) = \sum_{I \in I_{step}} P(I)$$

$$\text{where} \quad I_{step} = \{(k_1, \ldots, k_n) | k_1 + \ldots + k_n = \text{scans(step)}\}$$

This specifies the probability of a single step. Over all the students there are almost 2000 observations for each of the four steps. The probabilities of all N tokens of Step j are combined to determine the log likelihood of the data for Step j given a model with its set of estimated parameters:

$$L(N\ tokens) = \sum_{x=1}^{N} \ln(Pr(token_x)).$$

[8] While the principal component analysis guarantees zero correlation, the assumption of independent normal distributions is only approximately correct. We do observe more extreme values than would be expected under a normal and extreme values on different dimensions tend to co-occur.

References

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*, 313–342.
Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.
Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive Tutors: Lessons learned. *The Journal of Learning Sciences*, *4*, 167–207.
Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. (2010). Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Science, USA*, *107*, 7018–7023.
Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. (2011). Cognitive and metacognitive activity in mathematical problem solving: Prefrontal and parietal patterns. *Cognitive, Affective, and Behavioral Neuroscience*, *11*, 52–67.
Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. Tracking children's mental states while solving algebra equations. *Human Brain Mapping*, in press-a.
Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. Can neural imaging be used to investigate learning in an educational task? In: J. Staszewski (Ed.), *Expertise and skill acquisition: The impact of William G. Chase*, in press-b.
Ansari, D., & Dhital, B. (2006). Age-related changes in the activation of the intraparietal sulcus during nonsymbolic magnitude processing: An event-related functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, *18*, 1820–1828.
Borst, J. P., Taatgen, N. A., Stocco, A., & Van Rijn, H. (2010). The neural correlates of problem states: Testing fMRI predictions of a computational model of multitasking. *PLoS One*, *5*(9)
Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.
Brunstein, A., Betts, S., & Anderson, J. R. (2009). Practice enables successful learning under minimal guidance. *Journal of Educational Psychology*, *101*, 790–802.
Davatzikos, C., Ruparel, Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, *28*, 663–668.
Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*, 487–506.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
Falkenstein, M., Hohnbein, J., & Hoorman, J. (1995). Event related potential correlates of errors in reaction tasks. In G. Karmos, M. Molnar, V. Csepe, I. Czigler, & J. E. Desmedt (Eds.), *Perspectives of event-related potentials research* (pp. 287–296). Amsterdam: Elevier Science B.V.
Foerster, P. A. (1990). *Algebra I* (2nd ed.). Menlo Park, CA: Addison-Wesley Publishing.
Friston, K. J. (2006). *Statistical parametric mapping: The analysis of functional brain*. Academic Press: London.
Goldstone, R. L., Landy, D., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, *2*, 265–284.
Graesser, A. C., D'Mello, S. K., Craig, S. D., Witherspoon, A., Sullins, J., McDaniel, B., et al. (2008). The relationship between affect states and dialogue patterns during interactions with AutoTutor. *Journal of Interactive Learning Research*, *19*, 293–312.
Haynes, J. D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Current Trends in Biology*, *15*, 1301–1307.
Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Trends in Biology*, *17*, 323–328.
Hsu, C. W., Chang, C. C., & Lin, C. J. (2009). *A practical guide to support vector classification*. Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University.
Hutchinson, R., Niculescu, R. S., Keller, T. A., Rustandi, I., & Mitchell, T. M. (2009). Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models. *NeuroImage*, *46*, 87–104.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–1195.
Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.
Rabiner, R. E. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

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

Rivera, S. M., Reiss, A. L., Eckert, M. A., & Menon, V. (2005). Developmental changes in mental arithmetic: Evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex*, *15*, 1779–1790.

Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human–Computer Interaction*, *16*, 39–86.

Sohn, M.-H., Albert, M. V., Stenger, V. A., Jung, K.-J., Carter, C. S., & Anderson, J. R. (2007). Anticipation of conflict monitoring in the anterior cingulate cortex and the prefrontal cortex. *Proceedings of National Academy of Science*, *104*, 10330–10334.

Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: A statistical view informed by ACT-R. *Cognitive Science*, *32*, 1349–1375.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959.

Yu, S. Z. (2010). Hidde Semi-Markov Models. *Artificial Intelligence*, *174*, 215–243.