COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2013) 1–31

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12068

Discovering the Sequential Structure of Thought

John R. Anderson, Jon M. Fincham

Department of Psychology, Carnegie Mellon University

Received 14 June 2012; received in revised form 26 January 2013; accepted 30 January 2013

Abstract

Multi-voxel pattern recognition techniques combined with Hidden Markov models can be used to discover the mental states that people go through in performing a task. The combined method identifies both the mental states and how their durations vary with experimental conditions. We apply this method to a task where participants solve novel mathematical problems. We identify four states in the solution of these problems: Encoding, Planning, Solving, and Respond. The method allows us to interpret what participants are doing on individual problem-solving trials. The duration of the planning state varies on a trial-to-trial basis with novelty of the problem. The duration of solution stage similarly varies with the amount of computation needed to produce a solution once a plan is devised. The response stage similarly varies with the complexity of the answer produced. In addition, we identified a number of effects that ran counter to a prior model of the task. Thus, we were able to decompose the overall problem-solving time into estimates of its components and in way that serves to guide theory.

Keywords: Multi-voxel pattern recognition; Hidden markov models; Problem solving; Cognitive models

1. Introduction: Using neural imaging to discover problem-solving states

One of the striking features of the human mind is its ability to engage in complex intellectual operations. The challenge is to find ways to shed light on the underlying processes rather than just to be dazzled by our intellectual ability. For many decades, studies of problem solving did not go far beyond demonstrations of successful and failed efforts at problem solving. Newell and Simon (1972), seeking a method to penetrate into the

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

Website http://act-r.psy.cmu.edu/publications/pubinfo.php?id = 1053 contains Matlab code and data files allowing most of the major analyses in this article to be reproduced.

underlying thought processes, relied heavily on verbal protocols as evidence of what happens during a long episode of problem solving and their methodology continues to be used four decades later (e.g., Ericsson, 2006). Despite the use of methods such as protocol analysis, it has been difficult to identify the stages of complex problem solving and how long these stages take. This article will describe how functional magnetic resonance imaging (fMRI) can be used to discover the distinct mental states someone goes through in solving a problem and how the durations of each of these states vary with the conditions of an experiment.

Neural imaging rose to prominence largely after the passing of Newell and Simon.[1] To date, neural imaging techniques have not offered much for understanding the sequential structure of complex thought. Perhaps, this is unsurprising given the long durations and variability of problem-solving behavior. This, coupled with the slow nature of the hemodynamic response, might make it seem unlikely that fMRI could shed any light on complex problem solving. However, the methods we will describe actually take advantage of trial-to-trial variability, making problem solving a natural domain of application for fMRI. Moreover, the long durations of problem solving make the slow hemodynamic response less problematic.

While neural imaging techniques have not had great success in penetrating the sequential structure of complex thought, multi-voxel pattern analysis (MVPA) techniques have shown promising results with respect to penetrating the representational structure of complex thoughts. These techniques attempt to map distributed patterns of brain activity onto various significant categories (for review, Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009). As a recent instance of success, Just, Cherkassky, Aryal, and Mitchell (2010) show how one can map the semantics of different concepts onto patterns of brain activation to enable successful classification of novel words. As another example, Eger et al. (2009) were able to use parietal activation to predict the number of objects that a participant was considering.

We have been trying to extend MVPA to unravel the sequential structure of thought. Our efforts have involved combining MVPA with Hidden Markov Model (HMM) algorithms (Rabiner, 1989) that are used for temporal pattern recognition. We have shown that this combined method is capable of tracking the course of thought over minutes. For instance, Anderson, Betts, Ferris and Fincham (2010, 2012a) tracked students as they interacted with an intelligent tutor, identifying what solution step they were on in a long sequence of steps, and whether they were performing that step correctly. Anderson, Fincham, Yang, and Schneider (2012b) successfully followed participants in the game Concentration, which involved tracking their path through a problem-space that had a high branching factor.

Until this article, these have been "mind-reading" demonstrations that predicted the steps participants were taking during complex problem solving by tracing the patterns of their brain activation. These applications require computer logs of participant behavior that provide a "ground truth," both for training the algorithms and for judging their accuracy. The goal of the research reported in this article is to extend that method to discover what is not in a computer record, which is the sequence of men-

tal states that a participant goes through in solving a problem before they emit an action.

The remainder of this article will consist of four more sections (this introduction being Section 1): Section 2 describes a pair of experiments on a mathematical problem-solving task that requires participants to devise modifications to a learned procedure to solve novel problems. Section 3 describes how our MVPA-HMM method identifies the mental states involved in solving a problem and estimates the duration of these states. Section 4 describes the application of the method in Section 3 to the task in Section 2, showing that it can discover both the mental states and how the durations of these states vary with experimental conditions. Section 5 discusses the general potential of these methods, both limitations and promise.

2. Pyramid experiments

We will consider two experiments on "pyramid problems" with primary reports elsewhere, referred to as Experiment 1 (Anderson, Betts, Ferris, & Fincham, 2011) and Experiment 2 (Wintermute, Betts, Ferris, Fincham, & Anderson, 2012). Table 1 reproduces the instruction given to participants in Experiment 1. Pyramid problems (for instance, 4$3 = *X* from Table 1) involve a base ("4" in this example) that is the first term in an additive sequence and a height ("3" in this example) that determines the number of terms to be added. Each term in the sequence is one less than the previous—so 4$3 = 4 + 3 + 2. As students work with pyramid problems, they quickly master an algorithm for *Regular* problems like those in Table 1. However, they can be presented with *Exception* problems that require they extend their knowledge (and most can do so with at least some success). For instance,

−9$4 = X

X$4 = X.

Table 1
Instructions given to participants on pyramid problems

There is a notation for writing repeated addition where each term added is one less than the previous:
For instance, it is written as 4 + 3 + 2 = 4$3
Since 4 + 3 + 2 = 9, we would evaluate 4$3 as 9 and write 4$3 = 9. The parts of 4$3 are given names:
4 is the base and reflects the number you start with
3 is the height and reflects the total number of items you add, including the base
4$3 is called a pyramid
You will see a variety of problems in which you will need to solve for the variable. Here, three examples are given in the following:
4$3 = *X* you are to type "9" the answer
X$3 = 9 you are to type "4" the base
4$*X* = 9 you are to type "3" the height

To illustrate how students respond to such pyramid problems, below are two protocols from different students as they solved the two problems. For −9$4 = *X*:

"So wait it is one less so it is minus ten…minus nine plus minus ten plus minus eleven is three and then negative… minus 42"

After solving the problem, the student continues to think about it, even though he got it right:

"You need to explain the rule"

The other student on *X*$4 = *X*:

different… Oh, that is interesting…*X*, 3*X*+6 (enters −2)

and after the feedback indicating that the correct answer was 2:

"no, oh 2, 2, 2! Shit!"

These exception problems have been a focus of our attention because they pose challenges to modeling in the ACT-R theory (Anderson et al., 2004), which has been successfully applied to neuroimaging data from more routine mathematical problem solving (e.g., Anderson, 2005; Anderson, Carter, Fincham, Ravizza, & Rosenberg-Lee, 2008; Rosenberg-Lee, Lovett, & Anderson, 2009). Anderson (2007) described an ACT-R model that dealt with the behavioral data from these pyramid problems. However, that model was later disconfirmed by our imaging data. Thus, we have here a class of problems in search of a model.

In both experiments, participants practiced solving regular problems before they were scanned. They did not encounter exception problems until the fMRI sections of the experiments. The most significant differences between the two experiments involved the problems that were used for initial practice and the problems that were used during scanning. In Experiment 1, participants practiced all three forms of problems illustrated in Table 1 (solve for value, solve for base, and solve for height), but in Experiment 2 participants practiced only solve-for-value problems.[2] In the scanner trials of Experiment 1, only 1/5 of the problems were exception problems, whereas 3/4 were exceptions in Experiment 2. Experiment 1 was focused on understanding what students would do when they encountered a problem that posed a surprise challenge. Experiment 2 was focused on understanding how different types of exception problems were handled. Looking at results across the two experiments provides a test of the consistency of our model-discovery method.

The detailed methods are described in the original publications but below we provide a summary.

2.1. Participants

There were 20 participants in Experiment 1 and 36 in Experiment 2. All were right handed between the ages of 18 and 40 (8 females and 12 males in Experiment 1 with a mean age of 23.8; 16 females and 20 males in Experiment 2 with a mean age of 23.1). They were recruited from the general Carnegie Mellon subject population.

2.1.1. Procedure

Prior to the scanner trials, participants practiced regular problems, where the base was in the range 1–9 and the height 1–9. For Experiment 1 the prior practice involved 18 problems, while for Experiment 2 it involved 81 problems. Tables 2 and 3 give the distribution of problem types used in scanner trials of the two experiments. Each exception problem that a participant saw in Experiment 1 was a unique exception, whereas in Experiment 2 participants saw eight different examples of the nine exception types over the course of the experiment.

During the scanner trials, each problem was presented on the screen as shown in Fig. 1 preceded by a 3 s fixation period. Participants had 30 s to input an answer. In Experiment 1, participants entered their answers with a mouse and screen-displayed keypad. In Experiment 2, they used a physical keypad that they had been trained on. After their response or 30 s expired, feedback was presented for 5 s, showing the explanation for the correct answer and indicating whether their response was correct. After feedback in Experiment 2 only, a fixation cross was again presented for 3 s. In both experiments, there was a repetition-detection task for 12 s. During repetition detection, letters appeared on the screen at a rate of 1 per 1.25 s, and participants were instructed to press enter on the keypad when repeated letters occurred. This task served to distract the participants from the main pyramid task and return brain activity to a relatively constant level.

2.2. Image analysis

Images were acquired using gradient echo-echo planar image acquisition on a Siemens 3T Allegra (Experiment 1) and on a 3T Verio (Experiment 2). Both experiments acquired 34 axial slices on each TR using a 3.2 mm thick, 64 × 64 matrix. This produces voxels

Table 2
The eight categories of problems in Experiment 1

1. 434 correct regular base problems (e.g., *X*$3 = 9). Mean: 10.4 s
2. 436 correct regular height problems (e.g., 4$*X* = 9). Mean: 9.2 s
3. 431 correct regular value problems (e.g., 4$3 = *X*). Mean: 9.2 s
4. 433 correctly font problems.[17] Mean: 10.0 s
5. 206 correct exception problems. Mean: 15.8 s
6. 171 incorrect regular problems (from 1 to 4 above). Mean: 13.5 s
7. 181 incorrect solved exception problems. Mean: 18.6 s
8. 108 problems timed out (1–5 above). Mean: 30.8 s

Note. There were 480 possible instances of each of Categories 1–5

Table 3
The 13 categories of problems in Experiment 2

Exception problems that use the same addition algorithm but use unusual numbers:
1. 244 correct negative height value problems (e.g., 4$−3 = *X*).[18] Mean: 11.2 s
2. 207 correct negative base value problems (e.g., −2$4 = *X*). Mean: 14.0 s
3. 184 correct large-base value problems (e.g., 208$3 = *X*). Mean: 16.6 s
Exception problems that vary the solution algorithm but use simple numbers:
4. 245 correct unknown height problems (e.g., 5$*X* = 12). Mean: 11.3 s
5. 237 correct unknown base problems (e.g., *X*$4 = 30). Mean: 12.2 s
6. 189 correct double-*X* problems (e.g., *X*$*X* = 15). Mean: 13.5 s
Exception problems that vary algorithm and numbers:
7. 257 correct large-base height problems (e.g., 110$*X* = 534). Mean: 9.6 s
8. 214 correct fractional-height value problems (e.g., 5$2⅓ = *X*).[19] Mean: 12.4 s
9. 179 correct mirror problems (e.g., 200$401 = *X*).[20] Mean: 9.6 s
Other Categories
10. 512 correct regular value problems (e.g., 5$3 = *X*). Mean: 8.7 s
11. 56 incorrect regular value problems (10 above). Mean: 12.2 s
12. 497 incorrect exception problems (1–9 above). Mean: 16.4 s
13. 114 problems timed out (10 above). Mean: 12.2 s

Note. There were 285 possible instances of Categories 1–9 and 570 possible instances of Category 10

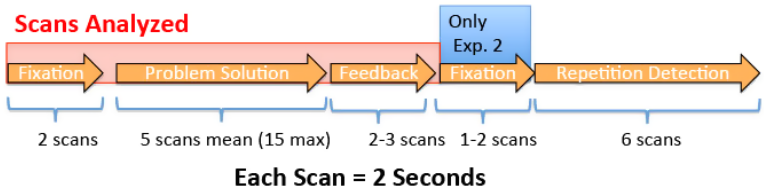


Fig. 1. An illustration of the sequence for Experiment 2: The problem began with a 4-s fixation cross and then was followed by a problem that stayed on the screen until the participant answered or until 30 s were up. Participants responded by entering the answer in a numerical keypad. This was followed by feedback on the correct answer. After 5 s of feedback, there was another 3 s of fixation. Then, participants performed a repetition-detection task for 12 s. In this task, letters appeared on the screen at the rate of 1 per 1.25 s. Participants were instructed to click a key each time they detected a pair of letters that were the same. Experiment 1 has the same structure except it lacked the second fixation period. The scans to be analyzed come from the common procedure from the onset of the first fixation to the end of the feedback.

that are 3.2 mm high and 3.125 × 3.125 mm^2. The anterior commissure-posterior commissure line was on the 11th slice from the bottom scan slice. Acquired images were preprocessed and analyzed using the NIS system and AFNI (Cox, 1996; Cox & Hyde, 1997). Functional images were motion-corrected using 6-parameter 3D registration (AIR; Woods, Grafton, Holmes, Cherry, & Mazziotta, 1998). All images were then co-registered to a common reference structural MRI by means of a 12-parameter 3D registration and smoothed with a 6 mm full-width-half-maximum 3D Gaussian filter to accommodate individual differences in anatomy.

In complex tasks like this, we have found it useful to perform MVPA on whole brain activity. However, as a step of dimension reduction and to accommodate variations in anatomy over participants that may not be dealt with in co-registration, we work with relatively

large regions. A total of 408 regions were created by evenly distributing 4 × 4 × 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. We examined these 408 regions and found that some had many extreme values for some participants, probably reflecting differences in anatomy. These were regions mainly on the top and bottom slices as well as some regions around the edge of the brain. Difficulties in getting consistent signals at edges reflect limitations in the co-registration of different brains to the reference brain and limitations of motion correction. Eliminating these edge regions resulted in a final set of 290 regions (e.g., see Fig. 4) that were used by our combined MVPA-HMM methods.

3. Model identification with combined MVPA-HMM

Our basic approach is to fit HMM models with different numbers of states to the imaging data from an experiment. For each number of states, we estimate a set of parameters that yields the best fit and then we select among the different numbers of states. First, we will describe parameter estimation and then selection of the number of states. The parameter estimation process uses an expectation maximization algorithm (Dempster, Laird, & Rubin, 1977), starting with "neutral" parameters and iteratively re-estimating parameters until convergence. We start with a model with maximal connectivity under the constraint that there are no loops. Fig. 2A illustrates the starting state graph for a three-state model. Three sets of parameters need to be estimated:

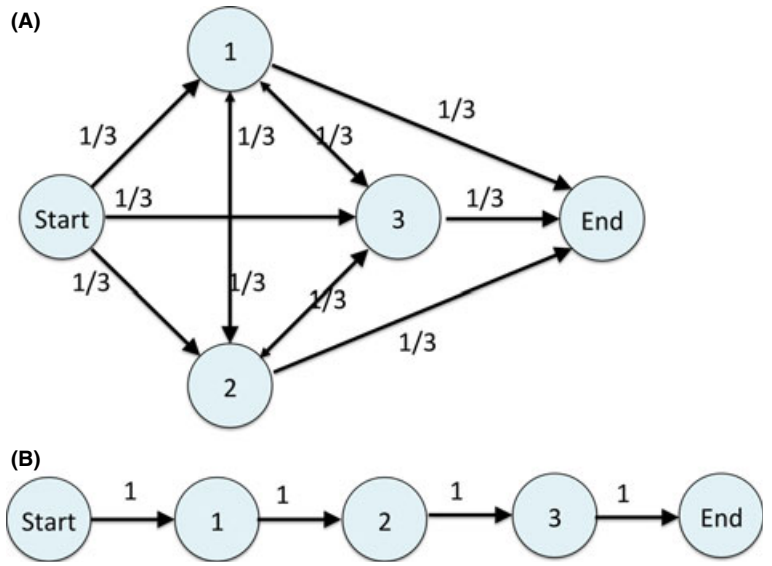


Fig. 2. (A) Illustration of a fully connected three-state model that has no loops with initial probabilities for estimation. (B) Linear model, which is a special case of a three-state model. The values on the links are the probabilities of the transitions.

1. The transition probabilities t_{ij} between states i and j. Fig. 2A shows that all transitions out of a state start out as equiprobable, but the probabilities do not stay equiprobable after the estimation procedure.
2. The distribution of durations in each state. We model state durations as gamma distributions with two parameters. If the mean duration of a trial in the data is t, both the shape and scale parameters for the n states are initialized to be the square root of t/n (and hence have means t/n) and then iteratively re-estimated. We discretize the gamma distribution to the nearest scan. Given that each scan is 2 s in length, this means that the probability of spending exactly m scans in state i is given as follows:

$$G(m|v_i, a_i) = \int_{2m-1}^{2m+1} \text{gamma}\ (t|v_i, a_i)dt,$$

where v_i and a_i are the shape and scale parameters for the gamma distribution for the ith state.[3] Note that this means that the probability of spending 0 scans in the state is the probability of a duration less than 1 s. These are cases where the state is skipped and the model moves on to a successor of that state. Allowing such skipped states (really just very short duration states) is critical in explaining brief trials (1% of trials last only a single scan and 8% last 2 scans). If each state were required to last at least 1 scan, short trials would force a solution structure with paths of few states. By allowing skipped states as a consequence of the distribution of durations, we can account for such short trials without forcing such a structure.
3. The probability of the fMRI activity in the state. Appendix A describes the steps in the signal processing which convert the lagged hemodynamic response into an estimate of the activity during a scan. We obtained estimates for 290 regions (see methods above) and then performed further dimension reduction by a principal component analysis (PCA). The first 20 factors in the PCA seem to result in the best state identification. These 20 factors are normalized to have mean 0 and standard deviations of 1 over all of the scans in the experiment. Each state i will be associated with a set of 20 means for these regions. A set F_j of observed factor values f_{jk} for scan k will have probability:

$$P(F_j|M_i) = \prod_{k=1}^{20} \text{Normal}(f_{jk}, \mu_{ik}, 1),$$

where M_i is the set of means μ_{ik} estimated for state i. The probability is a product on the probabilities of the 20 values assuming each is a normal with its own mean μ_{ik} and standard deviation 1. As described in Appendix A, the factors from the PCA are basically distributed as independent normals. For parameter estimation,

the means for each state are initialized to the overall mean of 0 and then iteratively re-estimated.

Parameters are estimated to maximize the likelihood of the data given a particular state topology. The probability of the data from a particular trial will be the sum of the probabilities of all ways of interpreting that data within that topology. An interpretation of a trial consisting of m scans is a way of breaking those scans into some r periods of residence in the various states of the model. Let $m_1 + m_2 + \ldots + m_r = m$ be one such interpretation, where m_i is the number of scans in state i, 1 is the start state, and r is the end state. The probability of this interpretation is

$$\Pr(m_1 + \cdots + m_r = m) = \prod_{i=1}^{r} \left(t_{i-1,i} G(m_i | v_i a_i) \prod_{j=1}^{m_i} \mathrm{P}(F_j | M_i) \right)$$

The estimation process calculates the summed probability of all such interpretations. This is the probability of the data in that trial. The number of ways of breaking m scans into a sequence on r states grows rapidly with m and r. HMM algorithms can efficiently calculate the summed probability using dynamic programming techniques (see discussion of explicit duration HMMs in Yu, 2010). For both computational and conceptual reasons, we largely use log likelihoods rather than probabilities. The parameter estimation process seeks to minimize the summed log-likelihood of all trials. We have adapted the software of Yu and Kobayashi (2003, 2006) to find maximum likelihood estimates of all the parameters of the model for all the trials in an experiment.

Appendix B describes proof-of-concept demonstrations of this parameter estimation procedure. There we fit the model to the activation patterns from initial fixation to end of feedback (see Fig. 1) and estimate the best-fitting three-state models for the experiments. The three states we recover correspond closely to actual fixation, problem solving, and feedback periods on a trial-by-trial basis. Given that the participants must be in different mental states during fixation, solution, and feedback, this result indicates that the states identified by this procedure correspond to real mental states. Moreover, the fact that the procedure identifies the trial-by-trial length of these states indicates that the duration estimates are reliable. These duration estimates prove to be the most informative result of the procedure when applied to identifying states within the problem-solving period.

3.1. Model selection

While it is reassuring to know that this method can recover the experimental procedure, the real interest is in what information it can provide about what is happening within the problem-solving phase. What are the mental states that participants go through in solving these problems and how do they vary with condition? As we do not know how many states there are, we need a method for selecting the number of states.

Fitting a model with more states will increase the likelihood of the data because of the extra parameters. To an approximation,[4] an $(n + 1)$-state model is nested within an n-state model—meaning that it will fit the data at least as well. However, this can be just overfitting the data—that is, taking credit for fitting noise that would not replicate. Although there are metrics for penalizing models for their extra parameters like BIC (Kass & Raftery, 1995), they do not extend in simple form to situations where there are so many parameters (Berger, Ghosh, & Mukhopadhyay, 2003) or where observations are not independent as is true of fMRI data (Jones, 2011). In contrast, we have found that cross-validation methods offer an effective way to assess models and identify when the extra model complexity is justified. This article uses simple leave-one-out cross-validation (LOOCV).

This application of LOOCV estimates the maximum likelihood parameters for all but one of the participants and then uses these to calculate the likelihood of the data for the remaining participant. In essence, this is estimating parameters from $k - 1$ participants and predicting for the kth participant. LOOCV rotates this process through all k participants. One model is to be preferred over another if it increases the likelihood of the data for most participants in this LOOCV procedure. We search through a set of models nested by their number of states, increasing the number of states in this search until the additional state does not lead to better performance according to LOOCV. Even if an n-state model is the true model, a $(n + 1)$-state model will fit the $k - 1$ participants better in the estimation phase, but it is at least as likely to fit data of the kth participant worse.

We use a sign test to approximate the probability that x out of k participants would be better fit by a $(n + 1)$-state model than a n-state model if the data were actually generated by a n-state model.[5] According to a sign test, for Experiment 1 with 20 participants, 15 corresponds to 0.05 significance, 17 to 0.005, 19 to 0.0001, and 20 corresponds to 0.00001. For Experiment 2 with 36 participants, these thresholds are 24, 26, 29, and 32.

In applying this LOOCV to the problem-solving phase, the parameter estimation sometimes produced a branching structure (e.g., Fig. 2A or non-linear subsets of the graph) for $k - 1$ participants. However, these branching structures never performed better on average for the kth participant than when we restricted the method to estimating a linear structure (e.g., Fig. 2B gives a three-state linear structure). The poorer performance of the branching structure reflects overfitting: The extra transition probabilities in the graph allowed unconstrained estimation of probabilities of skipping states. In contrast, while the linear models can skip states, the probabilities of doing so are constrained by gamma distributions of durations. The constraint in the linear models led to more stable estimates. In the discussion, we will return to the issue of branching structures, but the next section just considers linear structures with differing numbers of states.

4. Application to the pyramid experiments

Figure 3 shows how the average likelihood of the data for the kth participant varied with number of states in the two experiments and how many participants are better fit by

the more complex model. The figure separately plots the results for fitting just correct trials, just error trials, or all trials. Later sections will discuss in detail different types of correct and incorrect trials. The correct trials appear to have the more stable and interpretable conclusions. The best number of states is not the same between the two experiments: Experiment 1 shows evidence for a three-state model while Experiment 2 shows evidence for a four-state model.

Why did this method find three states in Experiment 1 but four states in Experiment 2? To explore this issue we used the Experiment 2 factor weightings to extract a new set of dimensions for Experiment 1. In contrast to the results obtained with the original Experiment 1 factors, there now is evidence for the four-state solution (17 of 20

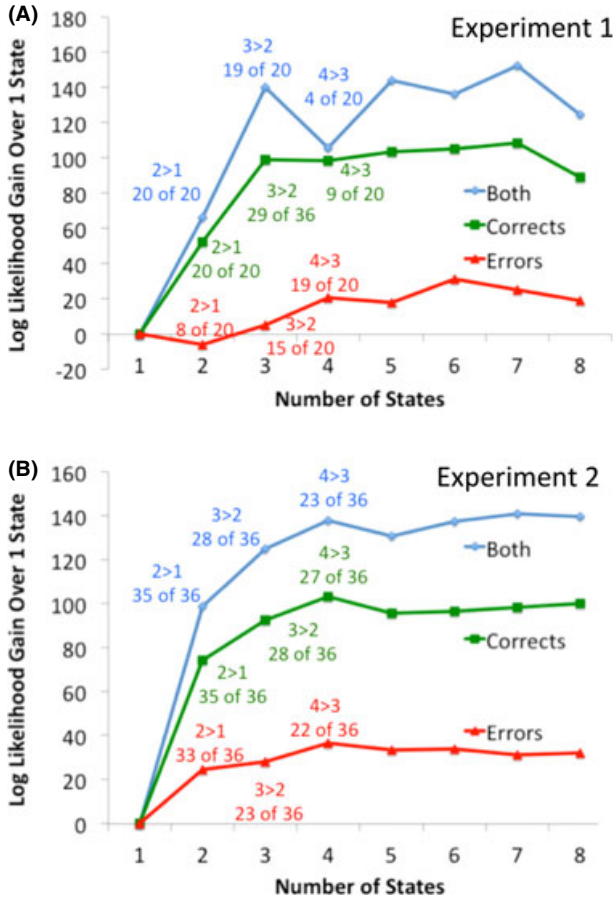


Fig. 3. Likelihood gain as a function of number of states in HMM for Experiment 1 (part A) and Experiment 2 (part B). Also given are the number of participants for whom the two states gave better matches than one state, three states better than two, and four states better than three. While sometimes models with more that four states show greater mean log likelihoods, in no case do they show greater likelihood for a significant majority of the participants than the best-performing model with fewer states.

participants better fit by a four-state solution with a mean improvement in log likelihood of 29.6). Equally, if the factor solution from Experiment 1 is applied to the Experiment 2 data, we fail to find evidence for four states (the four-state model only does better for 14 of the 36 participants). It seems that the first 20 factors in Experiment 2 contained variation that better discriminated the state structure, perhaps because of its much greater use of exception problems. The factor solution for Experiment 2 is more robust given the larger data set and the greater number of observations.

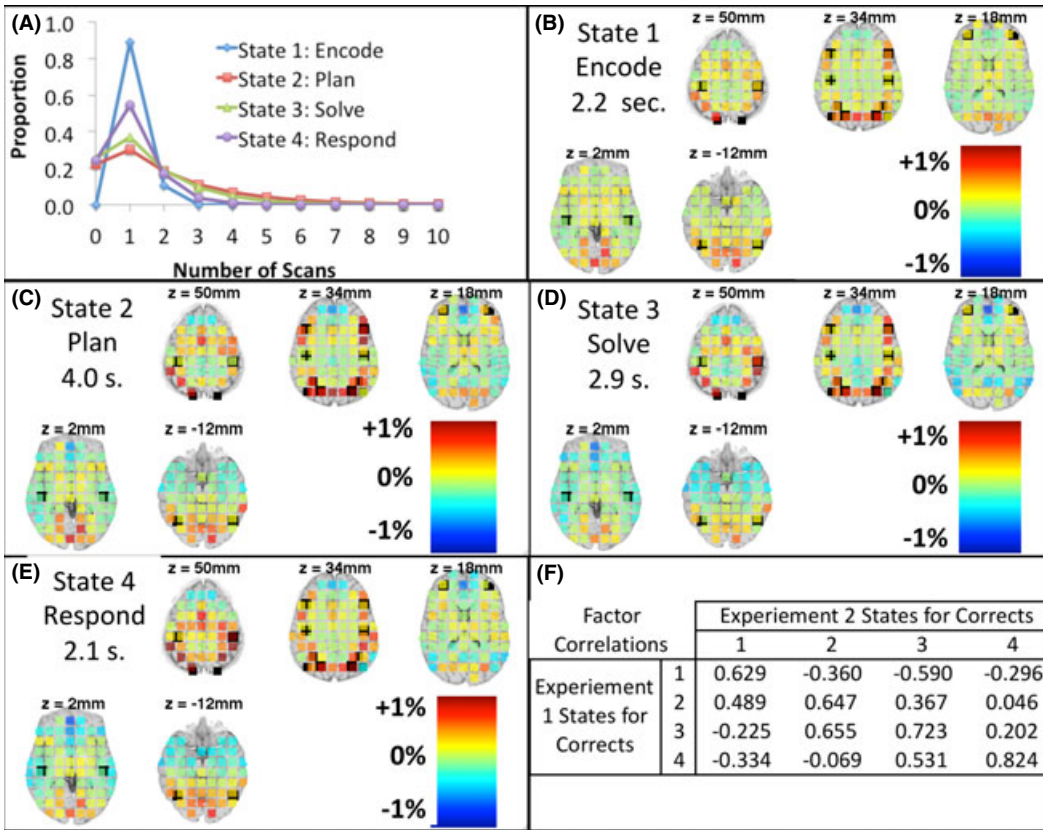
One might think that the problem with Experiment 1 is that we missed a critical dimension of variation in the top 20 factors in its PCA. While there is some truth to this, the evidence indicates that the situation is more a matter of including unsystematic variation. One of the 20 factors found in the PCA for Experiment 1 jumps around from participant to participant in the four-state solution, leading to poor performance in the LOOCV.[6] If we simply delete that one factor, a four-state solution emerges as the best with the remaining 19 dimensions.

This makes the point that the success of this method depends on identifying the critical dimensions of variation. One might wonder about just including all the factors from the PCA. However, the irrelevant dimensions of variation (noise) mask the relevant effects and we fail to identify even three states. The decision to use the first 20 factors of the PCA is only a heuristic solution to the problem of feature selection that haunts all applications of MVPA. We believe the four-state model is the correct one and in further analyses we will use the 19 factors from Experiment 1, deleting its distracting factor.

Figure 4 illustrates the four-state solution for Experiment 2. The activation patterns in parts (B)–(E) may look similar across states, reflecting the overall brain intercorrelations, but there are differences. Our factor solution projects this 290-dimensional intercorrelated structure down to 20 dimensions of orthogonal variation. The correlations between the factor solutions for the two experiments are given in Fig. 4F. The mean correlation between corresponding states is 0.71, while it is only 0.04 between non-corresponding states.[7]

The states in Fig. 4 are labeled with plausible characterizations of what they involve. These states are given in the following:

1. Encoding the Problem. Areas of high activity for this task are found in the visual areas and the parietal areas.
2. Planning the Solution. The lateral inferior prefrontal cortex (LIPFC) and the angular gyrus (although not high in absolute terms) show greater activation in this state than in the other states. The angular gyrus was a region found to be related to the processing of exceptions in both Anderson et al. (2011) and S. Wintermute et al. (unpublished data) . Given the retrieval functions associated with the LIPFC (e.g., Anderson, 2007) and language comprehension functions associated with the angular gyrus (e.g., Binder, Desai, Graves, & Conant, 2009), one might speculate that this reflects retrieval and processing of instruction and past feedback.
3. Solving the Problem. There is high activation in the parietal and prefrontal regions found active in other studies of routine problem solving (e.g., Anderson, 2005).



(F)

Factor Correlations		Experiement 2 States for Corrects			
		1	2	3	4
Experiement 1 States for Corrects	1	0.629	-0.360	-0.590	-0.296
	2	0.489	0.647	0.367	0.046
	3	-0.225	0.655	0.723	0.202
	4	-0.334	-0.069	0.531	0.824

Fig. 4. The parameters of the linear HMM identified for the problem-solving phase of Experiment 2: (A) The estimated distribution of durations in the three states. (B)–(E) The mean reconstructed activation patterns for the four states. Values displayed here are percent activation above the baseline established by the last scan of the repetition detection. Left is plotted on right. (F) The correlation between the factor solutions for the two experiments.

Also, the motor regions have high activation—many participants report finger counting to keep track of the height for these problems.

4. Responding with the Answer. There is very high activation in the left motor region (right hand). The parietal region also has high activation, reflecting processing the location of the keys on the number pad.

The average correlation between the activation patterns over the 290 regions in Fig. 4B–E is 0.86, reflecting the fact that similar regions tend to be active throughout performance of the task. When we project this activation down to the 20 principal components; however, the four states have a mean intercorrelation of 0.04. It is particularly impressive that we can pull apart the planning and solving states. While these two states have a 0.91 correlation when one considers the 290 regions in Fig. 4, the PCA factors for these two states have only a 0.22 correlation.[8] This clearly indicates that the states are

distinct but also indicates the importance of finding the dimensions in this correlated activation pattern that bring out these differences.

The existence of these four states and their associated activation patterns is sensible and again increases our confidence in the methods, but this very sensibility limits the information we get from this part of the discovery. The new information comes from being able to identify the duration of these states on a trial-by-trial basis. Fig. 5A shows the distributions of durations of these states over trials measured in number of 2-s scans. The Encoding State 1 is the least variable, lasting 1 scan on 89.3% of the trials and 2 scans on 10.4% of the trials. However, the other states vary much more substantially in duration. More than 20% of the trials have a 0 scan duration (i.e., a skipped state, which is the discrete realization of the state lasting less than a second). A measure of the range of durations for these states is how many scans are needed to cover 95% of the trials. For Planning State 2 that range is 0–6 scans, for Solving State 3 it is 0–4 scans, and for Response State 2 it is 0–2 scans.

These trial-to-trials variations in state durations are related to the various types of trials (see Tables 2 and 3), which define different conditions in the experiments. We think this is where we have discovered the most interesting information about the problem solving in these experiments. In these breakdowns, we assume that each state is defined by a

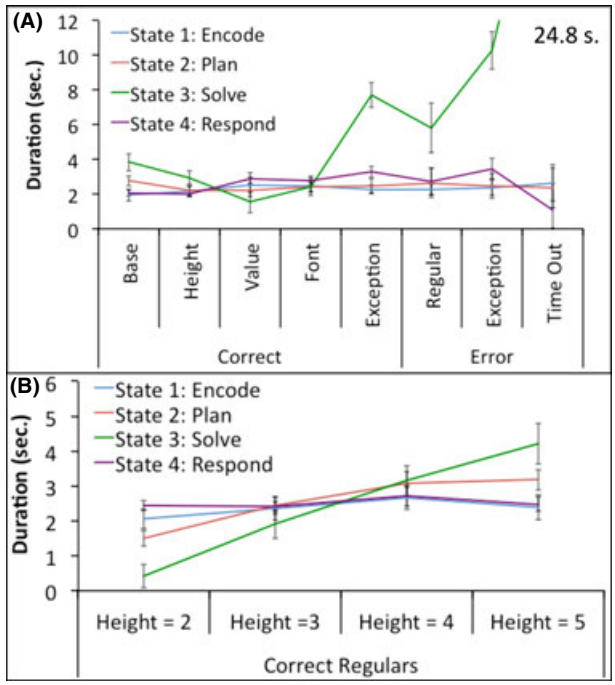


Fig. 5. (A) Variation in duration of states for the four states of Experiment 1 as a function of type of problem. See Table 2 for explanation of types. (B) Variation in duration of states for the four states of Experiment 1 as a function of height for correct regulars.

single activation pattern represented by its mean values for the 19 factors (Experiment 1) or the 20 factors (Experiment 2). This can be regarded as the *Signature* of the state. That signature does not change as a function of condition, but the time participants stay in the state can vary. The next two subsections will discuss how the state durations vary as a function of condition in the two experiments.

4.1. State durations in Experiment 1

We broke Experiment 1 up into the eight categories of trials defined in Table 2. We fit a four-state model to the 19 factors, allowing the duration of a given state to vary as a function of condition, keeping the activation pattern (the signature of the state) to be the same. There are four states and each state requires estimating two parameters to characterize the gamma distribution for that state. Thus, estimaing different distributions for each condition results in $8 \times 4 \times 2 = 64$ timing parameters being estimated rather than 8 that we used in the initial state discovery. Despite all of the extra parameters and the danger of overfitting, this expanded model results in better LOOCV fits for 18 of the 20 participants ($p < .0005$) with a mean log-likelihood improvement of 42.2.

We can extract an estimate of how long each participant spent in each state for each type of problem. Fig. 5A shows these average estimated durations and the standard errors of these means (calculated by a bootstrap procedure[9]). In effect, this analysis has taken the total latency in a condition and decomposed it into four separate dependent measures. The average correlation among the four measures is −0.11—so the information they provide is definitely not redundant with total time.

To consider the variation in each measure separately:

1. Encoding State 1 varies least among conditions. There is a slight tendency for the error conditions to be longer than the correct conditions (2.43 s vs. 2.28 s), but this is well within the uncertainty of the estimates.
2. Planning State 2 also shows little variability in duration as a function of condition. There is an interesting tendency for the solve-for-base problems to be longer than solve-for-height or solve-for-value (2.76 s vs. 2.22 – $t(19) = 2.01$, $p < .1$, two-tailed).[10] One might expect this effect because participants have to guess a base before confirming if it is correct.
3. Solving State 3 shows the greatest variability. It is longest in the time-out condition (which is 24.8 s, off the graph). Among the other conditions, State 3 lasts 2.69 s for correct regulars (average of the first four conditions), 7.71 s for correct exceptions, 5.82 s for incorrect regulars, and 10.25 for incorrect exceptions. The average difference between correct and incorrect is significant ($t(19) = 2.88$, $p < .01$) as is the difference between exceptions and regulars ($t(19) = 4.64$, $p < .0005$). Among the correct regulars, solve-for-base problems take longer than solve-for-height problems (3.83 vs. 2.95 s, $t(19) = 3.48$, $p < .005$) and solve-for-height problems take longer than solve-for-value problems (2.95 vs. 1.56 s, $t(19) = 3.18$, $p < .005$).

4. Response State 4 also shows some significant variation in its duration. This state is briefer in the base and height conditions where participants have to enter a single digit (mean 2.03 s.) than all other conditions (mean 3.03 s.) except the time-out condition where it is the least (1.07 s.). The difference between the single-digit cases and the other conditions (excluding time-outs) is quite significant ($t(35) = 3.04$ $p < .01$), but the high variability in the estimate of the time-out duration prevents any contrast involving the time-out Response State from becoming significant.

This study illustrates the power of the method. In Anderson et al. (2011), where we only looked at total time and average activity throughout the problem-solving interval, we failed to find any difference among the solve-for-base, solve-for-height, and solve-for-value conditions. However, the current analysis shows that each of these three types of problems has a rather different way of dividing up the time among the four states. The solve-for-base and solve-for height problems require changes to the standard computation and take longer in the solving state, but they require a simpler motor response.

The regular problems (categories 1–4) in Fig. 5A involve an equal number of problems with heights 2, 3, 4, and 5. The height determines how many additions have to be performed. Fig. 5B shows how the duration for the states varied for these problems as a function of height. There is very little variation in the duration of the encoding or response states, some increase for the planning State 2, but a much larger increase for Solving State 3. State 3 takes 1.4 s more for each additional term that has to be added. Again, this demonstrates how this method can isolate the location of an effect to a particular stage of the information processing.

While the results for the regular problems are quite informative, the results for the exception problems in Fig. 5A are not particularly satisfying. One would have expected more planning time for exceptions than regulars, but there is no significant evidence of this. We think this is related to the relatively few observations of exceptions in this experiment and the resulting poor identification of the dimensions that distinguish them.

4.2. State durations in Experiment 2

Experiment 2 had 13 categories of trials (see Table 3) with many observations of exceptions. We compared a four-state model whose durations did not vary with condition with a four-state model where the duration of each state could vary with condition. Despite the estimation of 96 additional timing parameters and dangers of overfitting, this expanded model results in better LOOCV fits for 30 of the 36 ($p < .0001$) with a mean log-likelihood improvement of 22.5.

Figure 6A shows the mean estimated durations and the standard errors of these means. The average correlation between the duration for different pairs of states is 0.14; thus, these patterns provide information not contained in the overall mean latencies. Time-outs are almost by definition different than the rest, but they still show an interesting pattern: They are significantly slower in the first three states ($t(35) = 3.21$, $p < .005$, $t(35) = 2.90$, $p < .01$, $t(35) = 3.90$, $p < .005$) but faster for the fourth state ($t(35) = 3.05$, $p < .005$).

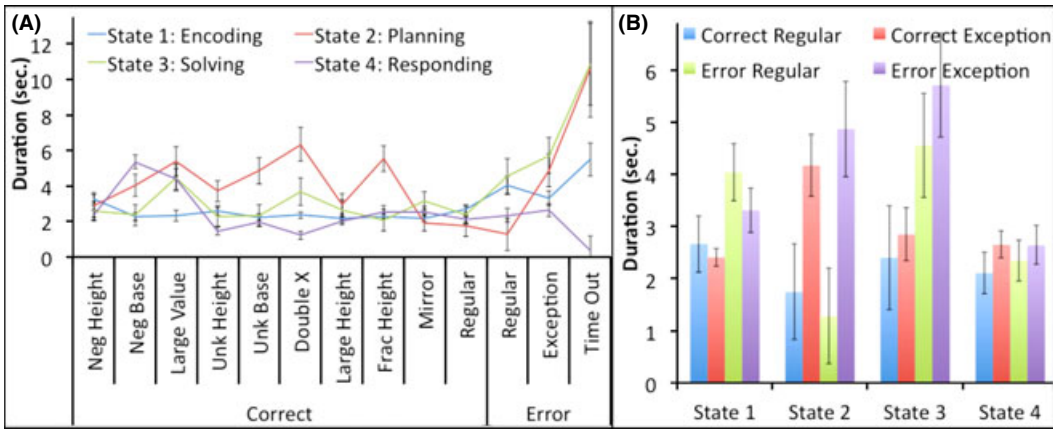


Fig. 6. (A) Variation in duration of states for the four states of Experiment 2 as a function of type of problem. See Table 3 for explanation of types. (B) Variation in duration of states for the four states of Experiment 2 organized by type and correctness.

The faster time in State 4 makes sense because the participants have not responded or at least have not completed their responses.

To present the data from Fig. 6A into a more interpretable pattern, Fig. 6B collapses that data into a 2 × 2 classification of problems according to correctness and type (regular vs. exception). Time-outs are excluded in this aggregation. Looking at the effects in Fig. 6B:

Correctness: Encoding State 1 and Solution State 3 are significantly longer for errors ($t(35) = 3.84$, $p < .0005$; $t(35) = 4.49$, $p < .0001$) while Planning State 2 and Response State 4 are not ($t(35) = 0.26$ and 0.50). This replicates the effects in Experiment 1. Combining the two experiments, it seems reasonable to conclude that participants tend to be a little slower in Encoding State 1 when they are going to make an error, while they are much slower in Solving State 3, where they make the error. The effect in Solving State 3 may reflect their stumbling over calculating the answer (e.g., forgetting a partial result and trying to recover).

Exceptions: Planning State 2 shows much greater time for Exceptions ($t(35) = 5.89$, $p < .0001$). The Response State 4 shows a small but significant effect ($t(35) = 2.10$, $p < .05$), reflecting the fact that exceptions tend to involve longer answers (see below). Encoding State 1 is definitely not slower for exceptions ($t(35) = -1.37$) and the effect for Solving State 3 does not reach significance ($t(35) = 1.53$). Experiment 1 showed a significant effect for Solution State 3. While the experiments differ on which states reach statistical significance, they both agree on the direction on the effects. The only state not slower for Exceptions in both experiments is the initial encoding state.

The above analysis treats all correct exceptions problems as equivalent, but Fig. 6A indicates that these effects do vary among different types of correct exceptions.

Encoding: There is little variation in the duration of Encoding State 1 but negative heights (e.g., 4$−3 = *X*) are slower than the rest ($t(35) = 2.15$, $p < .01$). Correctly solved negative height problems average 3.26 s in this state versus an average of 2.30 s for the other exceptions. This might reflect the facts that participant are uncertain about what is meant by a negative height (the software was programed to permit multiple possible interpretations—see footnote to Table 3).

Planning: There is considerable variation in the duration of the Planning State 2. Of particular interest are the long times for unknown base (e.g., X$4 = 30) and double-X (e.g., X$X = 15) problems, which tend to be solved by a guess-and-check strategy. Their planning states average 1.84 s longer than the other exception types ($t(35) = 4.97$, $p < .001$), perhaps reflecting the guessing process.

Solving: The solving time is 1.83 s longer for the large-base unknown-value problems than the other exception problems ($t(35) = 3.54$; $p < .005$). These problems (e.g., 208 $3 = *X*) require three-digit mental addition.[11]

Responding: Times for the Response State 4 are shorter for single-digit answer problems (unknown height, unknown base, double-X, large-base unknown height: 1.67 s.) than for problems that require two-digit answers (negative height, fractional height: 2.46 s.), and these are longer than conditions that require entering three characters (negative base and large base and unknown value: 4.87 s.).[12] Both pairwise comparisons are quite significant ($t(35) = 3.88$, $p < .0005$; $t(35) = 13.69$, $p < .0001$).

While the effects noted above do not capture all the variability among exception types, they capture the biggest effects. All of these effects are sensible, supporting the ability of this methodology to segment out interesting components of an overall trial latency.

In Experiment 2, each of the exception types was shown with eight different instantiations over eight blocks of the experiment. To understand how the state durations varied over the course of the experiment, we performed a regression analysis on the state times for correct exception problems. For each state, we regressed the estimated mean state duration for each block against the position of block:

Encoding Time = 2.42 – 0.00 * Blocks

Planning Time = 5.83 – 0.32 * Blocks

Solving Time = 3.83 – 0.12 * Blocks

Responding Time = 2.32 – 0.01 * Blocks

The decrease in planning time is the largest and only significant decrease ($t(6) = 3.59$, $p < .05$). This indicates that participants were learning how to solve these problems and most of the speed-up was in these planning times.

Encoding State 1 is 1.75 s longer for errors. This unexpected effect was also found in Experiment 1, but it is stronger and significant in Experiment 2. This suggests that, by

looking at activity in early scans, it should be possible to predict whether a participant will make an error well before the participant issues a response. To determine whether this was true, we performed an analysis restricted to problems on which participants took at least five scans (64.2% of the trials) before responding. We trained a linear classifier to predict whether a trial is correct based on the 20 factor scores on a particular scan. A LOOCV method estimated parameters from 35 of the participants to predict the trials of the 36th participant. Looking at hits (correct trials classified as correct) and false alarms (error trials classified as correct), one can calculate *d*-prime measures for each participant. Fig. 7A shows the results of the classification for each of the first five scans. There is no ability to detect an upcoming error given the activity of the first scan but there is for later scans. The first scan offers no evidence of an error because participants spend one scan encoding the problem in all cases. On the other hand, the activation pattern on the second scan indicates whether the participant is still encoding (symptomatic of an error) or has moved on to the planning state (not symptomatic of an error). Fig. 7B shows the difference in activation patterns for corrects versus errors on scans 2 and 3. It shows greater activation for parietal and prefrontal regions for corrects, suggesting that the participant has advanced to the planning state.

The interpretations that we have placed on these four states have a high degree of internal consistency; but, unlike the three states estimated for the overall experimental procedure (Appendix B), there is no ground truth to validate the results. There is one exception to this: If the last state does reflect the generation of the answer, it should have a high correlation with the actual time taken to key the answer. Experiment 2 collected measures of the amount of time the participant spent keying in the answer for each trial. The trial-by-trial correlation between keying time and estimated State 4 duration for a given trial is quite high (r = .527), very significant, and substantially greater than the next highest correlation (with State 3 duration, r = .221). Fig. 8 shows the relationship between the two measures—the actual time is 1.19 times the estimated time. This provides more evidence for the validity of the state identifications.

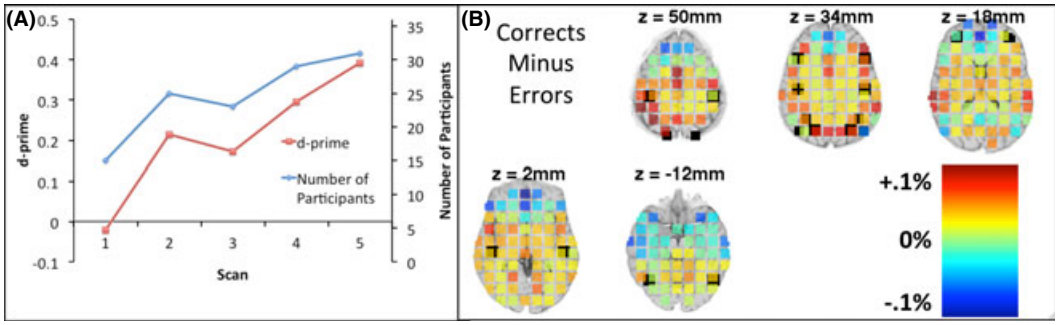


Fig. 7. (A) Ability to detect an upcoming error (on the fifth scan or later) given activity on an early scan. The two measures given are the average *d*-prime and the number of participants with positive *d*-primes. (B) Difference between correct and incorrect problems reconstructed from factors for scans 2 and 3. Left is plotted on right.

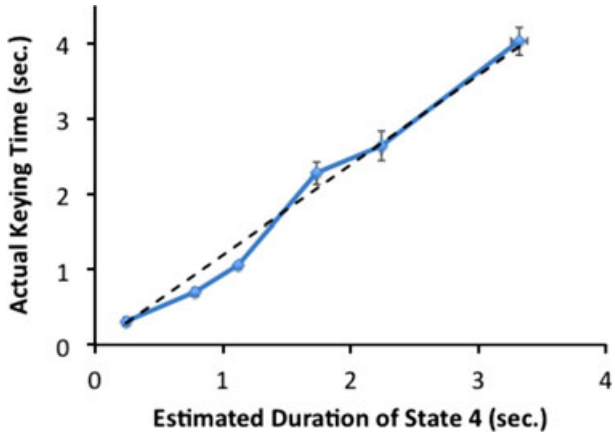


Fig. 8. Actual keying time as a function of the estimated duration of State 4. The dotted line shows the linear trend line $y = 1.19X$.

5. Discussion

This MVPA-HMM methodology for discovering states can yield interesting conclusions about how people solve problems. The method was successful in discovering both the experimenter-defined structure (Appendix B) and the participant-determined solution structure (Fig. 4). With respect to the participant-determined structure, there were four states that reflected encoding, planning, problem solving, and response execution. A particular promise of this method is its ability to identify how the duration of these states vary with condition (e.g., Figs. 5 and 6). These state-duration measures are not strongly intercorrelated and provide a much more articulate analysis of the internal structure of the task than just overall latency.

The discovery of encoding, planning, problem solving, and responding states is not surprising and was anticipated in a previous cognitive model for the task (Anderson, 2007). However, there were a number of surprising aspects of these states not anticipated:

1. The original model had a longer encoding state for exception problems because the model responded as soon as an unusual symbol (such as a negative number) was encountered. In contrast to that model, this was the one state that was not longer for exception problems in either experiment.
2. The encoding state was longer when an error would be made—which was also not expected. Perhaps, long encoding times are associated with confusion about the interpretation of the problem, which in turn can lead to an error.
3. We had not anticipated that highly practiced regular value problems would still have a planning stage. Although the Planning state is not as long as for exception problems, but it is still present for regular problems.[13] This suggests that students usually consider an approach before acting on it.

4. We had not anticipated the different profiles of state durations for solve-for-base, solve-for-height, and solve-for-value problems in Experiment 1 (where they had all been practiced in advance of the scanner trials). In fact, we had concluded earlier (Anderson, Betts, Ferris, & Fincham, 2011) that these three types of problems had similar activation and latency patterns.
5. We also had not anticipated that the problem-solving phase (and not the planning phase) would be associated with errors (Fig. 6B) for non-time-out trials. The only error condition associated with elevated planning times are time-outs. Thus, it seems that either participants fail to come up with a plan and time out or they come up with a plan and have trouble in executing it.

While we have shown that we can reliably recover a state structure that replicates across experiments and which reflects meaningful steps of processing, one might still wonder just what is reflected in the brain signatures (e.g., Fig. 4) associated with these states. In a cognitive architecture like ACT-R, information processing is happening at a much finer grain size than these multi-second states. We think these states reflect periods when the current goal demands a rather constant pattern of resource deployment (e.g., module activity in ACT-R). As such, they can serve as outlines for developing more detailed information-processing models. It is only when we have specified the information-processing steps that we will have given precise meaning to our labels for these states (Encoding, Planning, Solving, and Responding). Such a model could also inform us about how consistent we should expect the activation patterns to be within states.

If we were to develop an ACT-R model from this sketch, the current data would encourage us to base the encoding, solving, and responding states on the processes in the current model of algebra equation solving (e.g., Anderson, 2005). The activation patterns in these states are approximately what the activation patterns would be during these stages in that model—the encoding state involving perceptual and representational regions (visual and parietal); the solving state involving representational and retrieval regions (parietal and prefrontal); and the responding stage involving motor activity with visual monitoring. The one addition would be to add finger counting as part of the solving state. These models did not involve a planning state. To provide insight into what is happening in the planning state, we can look at its associated activation pattern. The increased angular gyrus and LIPFC activation in this stage suggests that participants are retrieving instructions and past feedback and reflecting on such information. There are a number of ACT-R models of instruction following (e.g., Anderson, 2007; Taatgen, Huss, Dickison, & Anderson, 2008) that could serve as a basis for modeling the processes in this state.

5.1. The structure of individual trials

These analyses rely on the ability of the method to diagnose individual trials and we have given evidence that it does so accurately (e.g., Fig. 8). The success of our earlier mind-reading efforts (e.g., Anderson et al., 2010, 2012a,b) also depended on diagnosing the unique way each trial unfolded. This individual trial analysis can be brought to bear

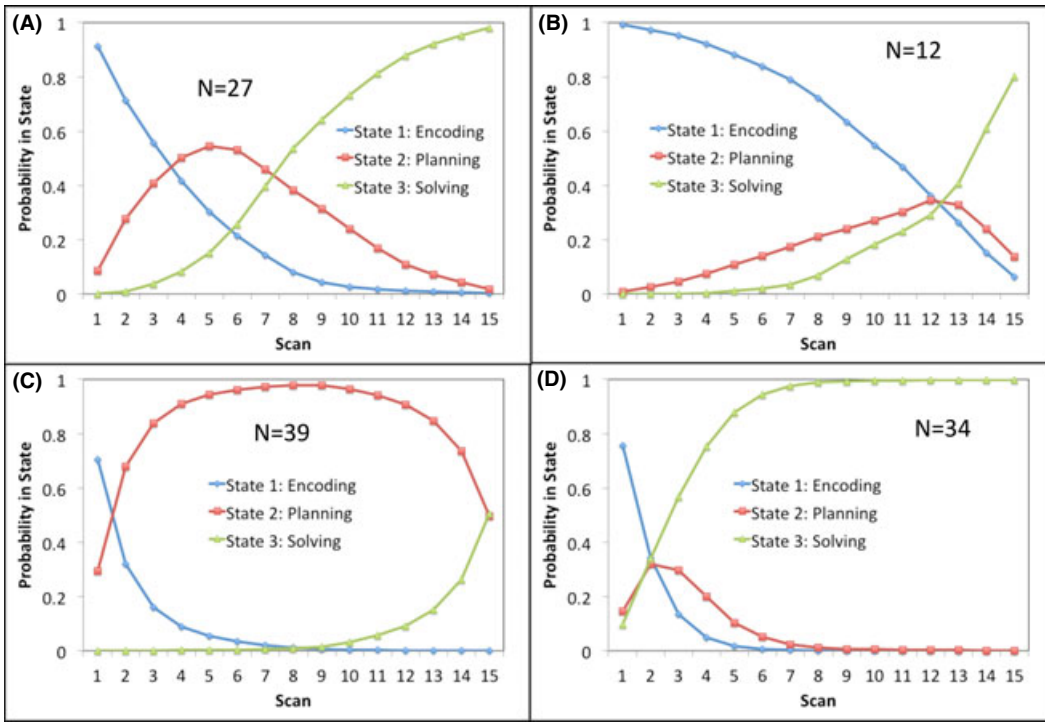


Fig. 9. Different patterns of state engagement displayed on different time-out trials. Each panels gives the number of trials observed to display that pattern.

in the current task and we can see what is happening on each trial. While we cannot present here the thousands of individual state trajectories identified by the MVPA-HMM method, Fig. 9 provides an illustration of the variability that does exist. It shows four distinct patterns of engagement that appeared on those trials where participants timed out and so took 15 scans.[14] The four panels of the figure show the estimated probability that a participant was in a particular state on a trial (ignoring the fourth response state which was seldom diagnosed as occurring at all). These four patterns summarize all of the time-out trials and the panels give the frequency of each pattern.

- On 27 of the trials (panel A), participants spend substantial time in all of the Encoding, Planning, and Solving states (estimated means of 3.5, 4.2, and 7.4 scans, respectively).
- On 12 of the trials (panel B), participants seem stuck in the encoding state and show low probability of getting on to later states (estimated means of 9.5, 2.6, and 2.8 scans).
- On 39 of the trials (panel C), participants seem stuck in the planning state (estimated means of 1.4, 12.5, and 1.2 scans).
- On 34 of the trials (panel D), participants seem stuck in the solving state (estimated means of 1.3, 1.2, and 12.5 scans).

The first pattern seems to reflect a participant making progress on the problem but not having enough time to complete it, the second pattern a participant stuck in interpreting the problem, the third a participant unable to come up with a plan of action, and the fourth a participant trying to apply an algorithm that will not work. This ability to do such individual-trial diagnosis opens up the potential for instructional applications.

5.2. Branching structures?

Why were linear structures always better than branching structures in this experiment? It is worth discussing a couple of reasons to have expected a branching structure but why such a structure was not found:

1. Participants might have planned only for some problems (principally exceptions) and not others, and this should produce a branch where the planning state can be skipped. However, the current method already allows a zero scan residence in a state. This is constrained by the overall gamma distribution and is not an additional free parameter. As noted earlier, sometimes the estimation process converged on branching structures that involved such skipping, but these structures failed to outperform linear structures in the LOOCV. Apparently, the constrained procedure for estimating state skipping that comes with fitting a linear model is more robust.
2. Some participants might have been taking non-conventional solution methods and one might expect to find different paths reflecting the different methods. For instance, in verbal protocols, we have observed a minority of participants to discover a variant of Gauss's apocryphal solution[15] where, rather than adding a sequence of terms, they multiply the height of the problem by the average term. For instance, 5$3 is calculated as $4 \times 3 = 12$ rather than as $5 + 4 + 3 = 12$. Also, different exception problems require different algorithms and one might have expected to find branching to different problem-solving states. However, these different states may have similar brain signatures. The method will find branching structures only if the signatures of the branching states are distinct. If just the state durations are different, the method will tend to treat this as a single state with a distribution of durations that merges the distributions for the different algorithms.

However, our results do not prove a linear structure is best. With enough training data, the method might have been able to estimate the parameters for a branching structure with enough reliability to outperform a linear structure in LOOCV. Also, parameter estimation uses a hill-climbing algorithm and there is no guarantee that it will find an optimum.

5.3. Generalization to other types of problem solving

To provide focus to this article, we have applied this method exclusively to pyramid problems. However, it is widely applicable. In our own laboratory, we have applied it to a task that involved addition of fractions and a transfer task in a tutoring system. While

these are tasks of interest to our laboratory, there is no reason why the methods need be limited to mathematical problem solving. However, for these methods to be useful, it is necessary that they take place over a long enough period that they can be profitably decomposed into multi-second states. For instance, the mean problem time was 14 s in the fraction task and 31 s in the transfer task. These two tasks contained information not in the pyramid task, but which would be part of many problem-solving tasks—participants entered intermediate calculations as they thought about the problem. Interpretation of these keying actions is ambiguous; nevertheless, adding in these intermediate actions improves state identification.

More generally, there is no reason why these methods need be limited to imaging data. Indeed, this work can be seen as an extension of earlier work using HMMs to parse eye movements (Salvucci & Anderson, 2001). The essence of the approach is to take a number of time-varying dimensions, all of which are somewhat correlated with the underlying problem solving, and using the convergence of information to identify states. The special advantage of imaging data is the number of dimensions of variation that it offers.

Acknowledgments

This study was supported by the National Science Foundation grant DRL-1007945 and a James S. McDonnell Scholar Award. We thank Julie Fiez for her advice on this research and Jelmer Borst, Anna Manelis, and Josh Rule for their comments on the article and Rob Kass for discussion of the statistical issues.

Notes

1. However, Simon did express optimism about what these data would reveal.
2. Solve-for-height and solve-for-base were treated as exception problems in Experiment 2 because the scanner trials were the first time participants saw these forms.
3. To optimize the estimation process, we approximate these integrals using their midpoints.
4. The discretization of the gamma distribution can cause exceptions.
5. We ran simulations in which data were generated from a n-state model and examined how likely LOOCV was to favor the $(n + 1)$-state model for differing numbers of participants (e.g., if one runs 20 simulated n-state participants, how often will LOOCV fit 15 participants better assuming a $(n + 1)$-state model). The probability of X or more out of k participants favored by the $(n + 1)$-state model is actually less than the probability of getting X or more out of k from a binomial distribution with $p = .5$. This is because of non-independence in the estimations from the various samples. Note also that the sign test provides an upper bound on the probability of data generated by an n-state model being better fit in LOOCV by an $(n + 1)$-

state model. It does not provide a bound on the probability that data generated by a (n + 1)-state model will be better fit by a n-state model in LOOCV.

6. We discovered this factor in a comparison of which factors contributed significantly to the variance among states in the three-factor and the four-factor solutions. The offending factor jumped from accounting for 1% of the variance in the three-state solution to 29% in the four-state solution. We then noticed that different participants had very different values for this factor.
7. The 19 Experiment 1 state means for each were converted back to 290 voxel activations using the factor coefficients for Experiment 1. These activations were converted into 20 Experiment 2 factor values for each state using the Experiment 2 factor coefficients. These converted means for Experiment 1 were correlated with the Experiment 2 means. Likewise, the Experiment 2 state means were converted into Experiment 1 factor values and correlated with the Experiment 1 values. The results are close and Fig. 4F reports the average of the two correlations.
8. This is a correlation within Experiment 2 factors, rather than between Experiment 1 and 2 factors, as in Fig. 4F.
9. Data sets are generated by randomly selecting 20 subjects with replacement from the experiment's 20 subjects (thus a subject can be represented 0, 1, or more times in a generated data set). Constraining all data sets to the same state signatures, maximum likelihood parameters are calculated for the 8 × 3 state durations. The standard deviation of the resulting time estimates provides an estimate of the standard error of the estimates from the actual data set.
10. These t contrasts are calculated using a standard deviation estimated from the variability of the contrast over iterations in the bootstrap procedure. Note that these standard errors will often be smaller than the standard errors of the means plotted in Figs. 5 and 6 (just as within-subject contrasts often have smaller standard deviations than the means). All significance levels reported are for two-tailed t's.
11. Answers to the large-base unknown-height problems can be estimated without dealing with all three digits.
12. Depending on the specific problem, mirror problems could require one, three, or four characters and so are excluded in this contrast.
13. The model does allow a state to be skipped (a 0-scan duration). In Experiment 2, the estimated probability of skipping the planning state is .40 for correct regulars and .20 for correct exceptions.
14. These four patterns were obtained by doing k-means clustering of the state occupancy for the time-out trials—see explanation of the code at http://act-r.psy.cmu.edu/publications/pubinfo.php?id=1053.
15. It is claimed that Gauss faced with a teacher's request to sum the digits 1–100 found the formula for triangular numbers.
16. We can find a branching structure if one exists. For instance, if we generate an artificial data set by splicing the fixation and feedback scans into their own "trials" and using the problem-solving scans as separate trials by themselves, the best three-state solutions involve branching structures where one branch has a single

state corresponding to the solution and the other branch has two states corresponding to the fixation and feedback periods.

17. To investigate simple surprise effects of exceptions, an equal mixture of the above three categories were presented in odd colors and fonts.
18. The answer to negative height problems is not obvious. The software accepted a variety of solutions. Almost all solutions to this problem were either 4 $-3 = 4 + 5 + 6 = 15 or 4$−3 = 5 + 6 + 7 = 18, both of which were accepted.
19. This is solved as 5$2⅓ = 5 + 4 + ⅓(3) = 10.
20. This is solved as 200$401 = 200 + 199 + … −199 + −200 = 0.

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. (2012). Tracking problem solving by multivariate pattern analysis and hidden Markov model algorithms. *Neuropsychologia*, *50*, 487–498.

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, 23, 983–3997.

Anderson, J. R., Betts, S. A., Ferris, J. L., & Fincham, J. M. (2012a). Tracking children's mental states while solving algebra equations. *Human Brain mapping*, *33*, 2650–2665.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*, 1036–1060.

Anderson, J. R., Carter, C. S., Fincham, J. M., Ravizza, S. M., & Rosenberg-Lee, M. (2008). Using fMRI to test models of complex cognition. *Cognitive Science*, *32*, 1323–1348.

Anderson, J. R., Fincham, J. M., Yang, J., & Schneider, D. W. (2012b). Using brain imaging to track problem solving in a complex state space. *NeuroImage*, *60*, 633–643.

Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference*, *112*, 241–258.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Cox, R. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Cox, R., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*, 171–178.

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

Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidth, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology*, *19*, 1608–1615.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. Hoffman (Eds.), *Handbook of expertise and expert performance* (pp. 223–241). New York: Cambridge University Press.

Fair, D. A., Cohen, A. L., Dosenbach, N. U., Church, J. A., Miezin, F. M., Barch, D. M., Raichle, M. E., Petersen, S. E., & Schlaggar, B. L. (2008). The maturing architecture of the brain's default network. *Proceedings of the National Academy of Sciences* USA, *105*, 4028–4032.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A. P., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: Characterising differential responses. *Neuroimage*, *7*, 30–40.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, *9*, 416–429.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, *30*, 3050–3056, doi:10.1002/sim.4323. Epub 2011 Jul 29.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, *5*, e8622.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

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.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, *45*, S199–S209.

Rabiner, R. E. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286.

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *Neuroimage*, *37*, 1083–1090.

Rosenberg-Lee, M., Lovett, M., & Anderson, J. R. (2009). Neural correlates of arithmetic calculation strategies. *Cognitive, Affective, and Behavioral Neuroscience*, *9*, 270–285.

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

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*(3), 548–565.

Wintermute, S., Betts, S., Ferris, J. L., Fincham, J. M., & Anderson, J. R. (2012). Networks supporting execution of mathematical skills versus acquisition of new mathematical competence. *PLoS ONE*, *7*, e50154, 1–16.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., & Mazziotta, J. C. (1998). Automated image registration: I. General methods and intrastudent intramodality validation. *Journal of Computer Assisted Tomography*, *22*, 139–152.

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, *174*, 215–243.

Yu, S.-Z., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit duration hidden Markov model. *IEEE Signal Processing Letters*, *10*, 11–14.

Yu, S.-Z., & Kobayashi, H. (2006). Practical implementation of an efficient forward-backward algorithm for an explicit duration Hidden Markov model. *IEEE Transactions on Signal Processing*, *54*, 1947–195.

Appendix A: Signal processing

We use regions of analysis that are 12.5 × 12.5 × 12.8 mm. The hemodynamic activity is correlated across regions and appears to be distributed as a multivariate normal, except that there are more extreme values than would be expected. Starting with 408 regions, we eliminated regions with too many extreme values (more than 10 scans with values five standard deviations beyond the mean, resulting in 290 for these experiments)

and bound the percent signal changes in the remaining regions to be within ±5% of baseline (truncating about 0.001 of the observations).

Figure 10 illustrates the processing of the BOLD (blood-oxygen-level-dependent) signal for a region on a single trial. We calculate the BOLD response as the percent change from a linear baseline defined from first scan (beginning of fixation before problem onset) to last scan (beginning of fixation before next problem). The example in Fig. 10 comes from a real trial and is representative of the noise in a trial. For instance, the last smaller ups and down at the end are well into the baseline task and probably not related to processing.

We assume BOLD response is produced by the convolution of an underlying activity signal with a hemodynamic response. The hemodynamic function is the SPM difference of gammas (Friston et al., 1998; —g = gamma(6,1)-gamma(16, 1)/6). A Wiener filter (Glover, 1999) with a noise parameter of 0.1 was used to deconvolve the BOLD response into an inferred activity signal (Matlab: deconvwnr(bold,g,.1)). To an approximation, this results in shifting the BOLD signal to the left by 2 scans (4 s). We have used this simple scan shift in our mind-reading studies (e.g., Anderson et al., 2010, 2012a; Anderson et al. in press a, b) where there is not a constant baseline activity interspersed at regular intervals.

The analysis in the Appendix B is focused on the scans from fixation to feedback and the analysis in the main body is focused on the subset of those scans that come from the problem-solving phase. While the scans in the *n*-back that follow the feedback contribute critically to the estimation of engagement in the earlier period, they are not used in any analysis. We also exclude the scans from the first problem in a block (always a warm up problem). Experiment 1 involved 20 participants solving 120 problems each, yielding 41,635 scans of which 24,170 were in the analyzed segments of trials (see Fig. 1). Experiment 2 involved 36 participants solving 88 problems each, yielding 65,959 scans of which 32,503 were analyzed.

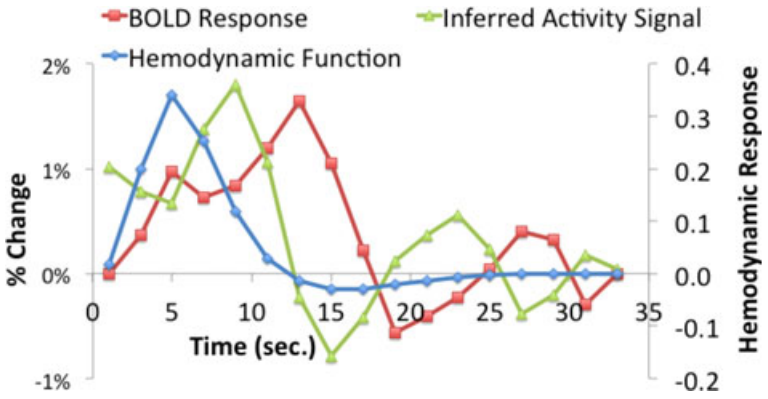


Fig. 10. Elements of the signal processing: The BOLD signal is percent change from baseline. Using a Weiner filter to deconvolve the hemodynamic response function results in the inferred activity at a time point.

We perform a PCA of this activity to find the 20 dimensions of greatest variation and divide the resulting eignenvectors by their eigenvalues to get dimensions that have zero correlation, mean zero, and standard deviation 1. These are a close approximation to independent normals except, like the BOLD signals from which they arise, they have too many extreme values. These are truncated to be within ±5 standard deviations (truncating about 0.0005 of the observations).

All the software used in the analysis and instructions for reproducing critical figures in this article are available at http://act-r.psy.cmu.edu/publications/pubinfo.php?id=1053

Appendix B: Segmenting the experimental structure

As Fig. 1 illustrates, the actual problem solving is just one phase of a larger experimental sequence. The analyzed scans consist of an initial 4 s of fixation and waiting, the participant-controlled problem solving, and then 5 s of feedback. Plausibly, the best-fitting three-state structure for these analyzed scans should be a linear structure like Fig. 2B and the durations of the three states should match up with the durations of the initial fixation period, the problem-solving period, and the feedback. To see if this was the case, we fit a three-state model to the trials.

Even though it starts with a fully connected three-state model like the one in Fig. 2A, the method identified the linear model in Fig. 2B as the best-fitting structure for both experiments.[16] Fig. 11A illustrates the durations of the three states in Experiment 1.

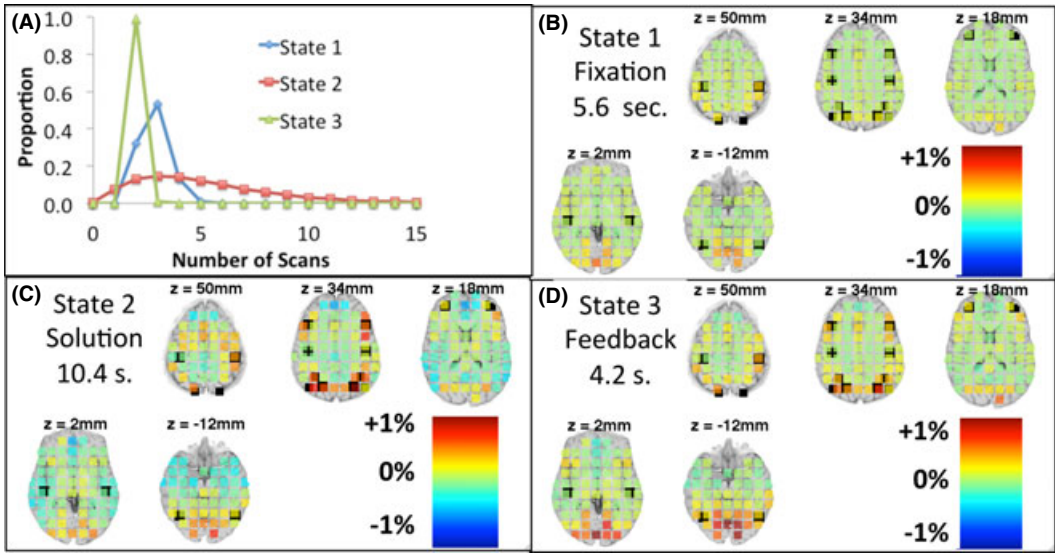
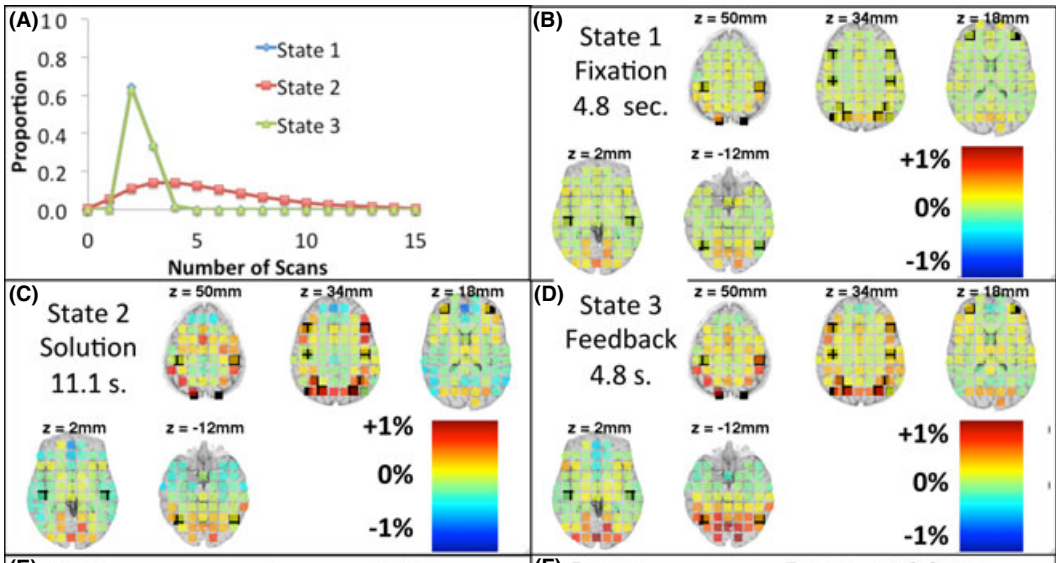


Fig. 11. The parameters of the linear HMM identified for Experiment 1: (A) The estimated distribution of durations in the three states. (B)–(D) The mean reconstructed activation patterns for the three states. Values displayed here are percent activation above the baseline established by the last scan of the repetition detection. Left is plotted on right.

There are relatively tight distributions of estimated durations for the first and third state and a wide distribution for the second state. The mean duration for the first state is a little longer than the 4-s fixation period, and mean estimated duration for the third state is a little shorter than the 5-s feedback period. However, the algorithm has basically discovered the three phases of the experiment.

Figures 11B–D illustrate activation patterns for the three states, reconstructed from the 20 factor means for each state. State 1 shows little variation in activation but more activation in default mode network areas (e.g., Fair et al., 2008; Raichle & Snyder, 2007) than the other states. State 2 shows the highest variation with strong activation in parietal, prefrontal, and motor areas that are engaged by algebra problem solving (e.g., Anderson, 2005). State 3 shows relatively high activation in visual and parietal areas that might be associated with attending to feedback.

Figure 12 shows the results of applying the algorithm to the scans of Experiment 2 (compare with Fig. 11). Fig. 12E gives the mean intercorrelations between state activation patterns in the two experiments. The activation patterns for corresponding



(E) BOLD Correlations

Experment 1 States	Experment 2 States 1	2	3
1	0.911	0.494	0.710
2	0.502	0.943	0.839
3	0.766	0.746	0.922

(F) Factor Correlations

Experment 1 States	Experment 2 States 1	2	3
1	0.943	-0.758	-0.142
2	-0.806	0.928	-0.268
3	0.036	-0.582	0.769

Fig. 12. The parameters of the linear HMM identified for Experiment 2. See Fig. 10 for explanation of (A)–(D). With respect to (A), the estimated distributed for States 1 and 3 are very similar and overlap. Part (E) gives the correlation between the mean percent BOLD change for Experiment 1 and 2 states. Part (F) gives the mean correlation between factor means for the two experiments.

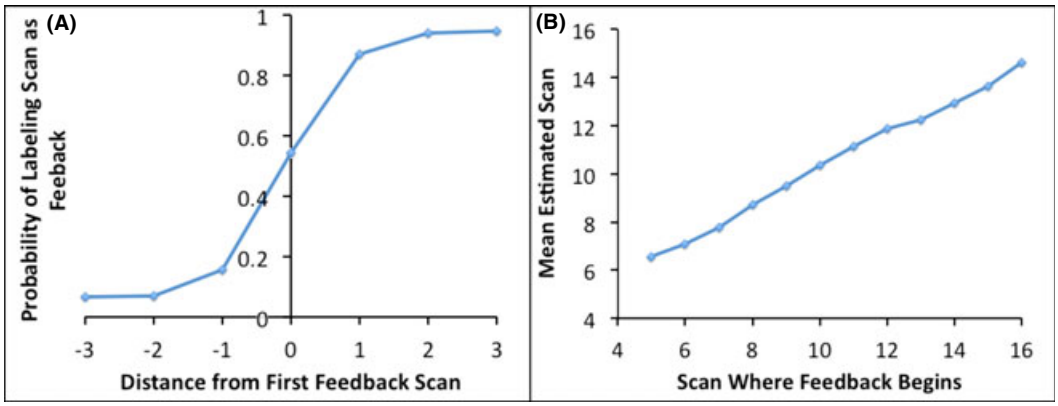


Fig. 13. Ability for the self-trained HMM to identify when feedback begins as the scans come in. (A) Probability of labeling a scan as feedback as function of distance from initiation of feedback. (B) Estimated first trial of feedback as a function of true first trial of feedback.

states are quite similar, having an average intercorrelation of 0.93. However, non-corresponding states have somewhat high correlations (mean of 0.68), reflecting the intercorrelation that exists in the raw BOLD values. In contrast, the factor scores from which these are constructed are not correlated across the total data set. Fig. 12F shows the intercorrelations of the factor means. The correlation between corresponding states is still high (mean 0.88) but the mean correlation between non-corresponding states is now negative (mean −0.42).

We examined how well the self-trained HMM could do at detecting the onset of feedback on a trial-by-trial basis. We used the Forward Algorithm (see Anderson et al., 2012a) that interprets each scan as it comes in without waiting until the end of the trial. We looked at Experiment 2 because there is an extra fixation period after the feedback, resulting in a full four scans after feedback onset. This gives more room for overshoot errors. Estimating a three-state solution for the expanded Experiment 2 trials produced similar results except that the third state is longer—5.2, 11.6, and 8.0 s for the three states (compare with times in Fig. 12).

The Forward algorithm gives the probability that a scan comes from each of the three states. Fig. 13A shows the probability of assigning a scan to State 3 as a function of when the feedback begins (0 on the *X* axis). Fig. 13B looks at the estimated first scan of feedback (defined as when State 3 becomes the most probable interpretation) as a function of the true first scan of feedback. Both halves of Fig. 13 show that the HMM is quite sensitive to the onset of feedback. Previously, we had trained HMMs with ground truth data as to where such boundaries are. In this case, the method is deciding where the boundaries are without the benefit of knowing the ground truth.