Can Neural Imaging Investigate Learning in an Educational Task?

John R. Anderson  
*Carnegie Mellon University*, ja@cmu.edu

Jon Fincham  
*Carnegie Mellon University*, fincham@cmu.edu

Jennifer Ferris  
*Carnegie Mellon University*, jlferris@cmu.edu

Shawn Betts  
*Carnegie Mellon University*

Follow this and additional works at: [http://repository.cmu.edu/psychology](http://repository.cmu.edu/psychology)

Published In

*proceedings of the 36th Annual Carnegie Symposium on Cognition.*

This Conference Proceeding is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Department of Psychology by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Can Neural Imaging Investigate Learning in an Educational Task?

John R. Anderson
Shawn Betts
Jennifer L. Ferris
Jon M. Fincham

Psychology Department
Carnegie Mellon University

Tuesday, July 23, 2013
Abstract

A methodology is described for using fMRI brain imaging to study how students learn with an intelligent tutoring system. Students were imaged while learning an algebra isomorph. A cognitive model, which is capable of learning the material, identified 6 predefined brain regions as reflecting various aspects of the problem-solving and learning process. The model predicted that the learning would be mainly manifested as reduced activation in a lateral inferior prefrontal region reflecting increased fluency in retrieving relevant declarative information. This prediction was confirmed but there was also decreased activation in a predefined fusiform region suggesting increased fluency in parsing the problem representations. Exploratory analysis also found an anterior prefrontal region whose activity predicted individual differences in error rates. The cognitive model was modified to include an increase in visual fluency. A careful assessment was made of the models’ ability to predict the time course of the BOLD response over 1-minute problem-solving episodes.
Neural imaging has contributed to the understanding of many forms of learning. Most of this research has used relatively simple laboratory tasks involving such things as word memory, perceptual priming, or simple skill acquisition. The question of interest in this paper is whether neural imaging can inform an understanding of the kind of learning that takes place in school. Studies of school learning have tended to use data sources such as final test performance that are both very summary in nature and much removed from the time of the learning. Imaging data from the performance of an actual school task might provide information that would be informative about the critical learning events.

We created a learning environment that reproduced the essential aspects of a Cognitive Tutor, an intelligent computer-based instructional system that which has been shown to have success in teaching high school mathematics (Anderson et al., 1995; Koedinger et al., 1997). These are deployed in over 1000 schools nationwide and interact with over 300,000 students in the United States (Koedinger & Corbett, 2006; Ritter et al., 2007, in press). These tutors interact with students as they solve problems providing instruction on an as-needed basis. The simplified system in this paper was designed to enable use in an fMRI scanner and to interact with a simulation of a cognitive model. Figure 1a illustrates its interface. The student selects parts of these equations by pointing and clicking with a mouse. As illustrated in Figure 1a, the selected portion is highlighted in red and the student picks operations to perform on that portion using the menu buttons below. When the student needs to enter information a keypad appears, as in the figure, and the student clicks the keys with a mouse. Figure 1a captures the student at the point where they are about to begin their entry. In this case the student has chosen the two x-terms, selected “Collect” (this option is hidden by keypad), and is ready to enter $3 + 4$ and $*$ into the
green boxes. There are some simple help options should a student get stuck. The curriculum sequence is based on the material in the first 4 chapters of the Foerster’s (1990) classic algebra text. The first one or two problems in a section will contain instruction explaining the transformations in that section. Note that all interaction with this system involves mouse actions and this is the only input device that the student needs in the scanner.

We have been cautious in exposing children to this system. We created an isomorph of algebra, a data-flow representation, for piloting the system with adults. Figures 2a and 2b show data-flow equivalents of a relatively simple equation and a relatively complex equation in this system. Part (a) is the isomorph of the equation $13-4x = 25$ and part (b) is an isomorph of the equation $(2x – 5x) + 13 + 9x = 67$. The student is told that a number comes in the top box, flows through a set of arithmetic operations, and the result is the number that appears in the bottom. They are taught a set of graph transformations isomorphic to the transformations with the linear equations that result in determining what number appeared in the top box. Figure 2b illustrates the data-flow interface for a comparable point to Figure 2a. Students point to boxes in this graph, select operations, and key in results from a keypad. The motor actions are totally isomorphic to the actions for a linear equation and in many cases physically identical. Anderson (2007, Chapter 5) reports a behavioral comparison of adults working with the data-flow tutor with children working with linear equations. Adults are a bit less error prone but show nearly identical learning trajectories as do the children.
The Current Imaging Study

The imaging study reported here was performed with adults using the data-flow representation.

The full experiment involved a 6-day procedure:

- **Day 0:** Participants received some general instruction on the tutor and did the isomorphs of Sections 1-1 and Section 1-2 from the Foerster text. This material involves a familiarization with algebraic expressions and their evaluation.

- **Day 1:** Participants did Sections 1-7 (on 1-step equations), 2-6 (on collection of constants), and 2-7 (on 2-step equations) in the scanner. Figure 2a is an example of a problem from Section 2-7.

- **Days 2-4:** Participants did three sections from Chapter 3 that involved more advanced collection and distribution and three sections from Chapter 4 that involved solving more complex diagrams like Figure 2b. The diagrams included the invert, combine, and evaluate operations in the Day 1 material as steps in their solution, along with other more advanced transformations.

- **Day 5:** They repeated the same sections as Day 1. The problems were different but the operations the same.

Before the repetition on Day 5, participants completed 181 problems from 11 Sections in the Foerster test. Except for the Day 1 and Day 5 problems, the problems were all the odd problems in the Foerster sections. For half of the students, the problems on Day 1 were the odd problems from the Foerster sections and the problems on Day 5 were the even problems and this was reversed for the other half of the students.
The imaging will be focused on the data from Sections 1-7 and 2-6. The problems in Section 2-7 were fewer and more heterogeneous and will be ignored in this report (although performance on these was similar). The solution of problems in both of Sections 1-7 and 2-6 can be divided up into two analogous episodes. Figure 3 illustrates the transitions that define these episodes. During the first episode (transitions between states (a) and (b) or between states (d) and (e) in Figure 3) participants selected one or two boxes to operate up, selected the transformation (“Invert” in section 1-7 or “Combine” in section 2-6), entered information into answer boxes through the keypad, and then performed a 12 second 1-back (Owen et al., 2005) to allow the blood oxygen level dependent (BOLD) responses to return to baseline and prevent them from rehearsing. During the 1-back the screen went blank and participants saw a sequence of letters presented at the rate of one per 1.25 seconds. They were to simply press the mouse if the same letter occurred twice in succession (which happened one third of the time). During the second episode (transitions between states (b) and (c) or between states (e) and (f) in Figure 3) they selected a box with an arithmetic expression that had been created in the first episode, selected “Evaluate”, entered the value of the expression using the keypad, and then went into another 12 second 1-back. Thus, the imaging data can be organized into 2x2x2 design in which the factors are day (1 or 5), problem (invert or combine), and episode (transformation or evaluation).

Method

Participants

Twelve right-handed members of the Pittsburgh community (7 females) aged 18 to 24 years old (M = 21.4 years) completed the study. One participant was excluded because of the appearance of many large sudden shifts in the imaging data.
Procedure

Table 1 details the events that happen in the solution of problems like those illustrated in Figure 3. The table is organized according to a set of 8 interface states that will be critical in defining the data analysis. The 1-back task was used to separate the transformation phase from the evaluation phase. It occupied participants such that they could not plan ahead for the next step but the demand of the 1-back was sufficiently low that the BOLD response would decrease to a constant baseline level\(^1\). The demands of performing either the transformation or the evaluation would be reflected in terms of an increase from this baseline. Except for the 1-back, the steps in solving a problem always had this character of selecting some boxes, selecting an operator, and filling in the result of that operator. Problems in later sections can involve many more than the two cycles (transform and evaluate) that occupied Sections 1-7 and 2-6.

If participants ever got stuck they could ask for the next step by selecting the “Hint” operator. In non-scanner sessions participants could perform incorrect operations. Such errors would require some extra work to correct. In the scanner, if they ever performed an action that was not on the path to the solution, that action was blocked and they were told what the correct action was.

There are 25 problems in Section 1-7 and 20 in 2-6. The first problem in each section was accompanied by instruction that illustrated the transformation. The remaining 24 problems in Section 1-7 were broken into 3 blocks of 8 problems each. The remaining 19 problems in Section 2-6 were broken into two blocks of 7 problems and 1 block of 5 problems. For imaging

\(^1\) Note we are not implying that this baseline reflects no activity just that it reflects an activity level well below that in place during the problem solving. Our modeling will account for the activity during the 1-back as well as during the problem solving.
purposes we ignored the first problem in a block and any problem that involved either a request for a hint or an error.

**FMRI Data Acquisition and Analysis**

Images were acquired using gradient echo-planar image (EPI) acquisition on a Siemens 3T Allegra Scanner using a standard RF head coil (quadrature birdcage), with 2 s repetition time (TR), 30ms echo time (TE), 70° flip angle, and 20cm field of view (FOV). We acquired 34 axial slices on each scan using a 3.2mm-thick, 64×64 matrix. The anterior commissure-posterior commissure (AC-PC) line was on the 11th slice from the bottom.

Acquired images were analyzed using the NIS system. Functional images were motion-corrected using 6-parameter 3D registration (AIR; Woods et al., 1998). All images were then co-registered to a common reference structural MRI by means of a 12-parameter 3D registration (AIR; Woods et al.) and smoothed with an 8mm full-width-half-max 3D Gaussian filter to accommodate individual differences in anatomy. Spatial F-maps were generated using random effects analysis of variance (ANOVA).

**Results**

We performed ANOVAs on the error rates and latencies using day and problem as factors. The average error rates were 15.0% for invert-day-1, 22.1% for combine-day-1, 14.4% for invert-day-2, and 21.8% for combine-day-2. The effect for problem type was significant (F(1,10) = 12.34; p < .01) but not the effects of days (F(1,10) = 0.03) nor the interaction (F(1,10) = 0.002).

We only analyzed the latencies and imaging data for the correct trials. The average latencies
were 53.2 seconds for invert-day-1, 55.5 seconds for combine-day-1, 46.4 seconds for invert-day-2, and 49.8 seconds for combine-day-2. The effect of problem type was significant (F(1,10) = 9.56; p < .05) as was the effect of days (F(1,10) = 45.84; p < .0001) but the interaction was not (F(1,10) = 0.90). With respect to issue of learning, participants maintained constant accuracy but showed a substantial improvement in latency as a function of three days of work with the more advanced material that included the invert, combine, and evaluate operations as steps in their solution.

Figure 4 displays the variability in solution times for the invert problems (Part a) and the combine problems (Part b). This is plotted in 2-second scans. The minimum number of scans is 17 (12 scans for 1-back, 1 scan for the transformation, 1 for the evaluation, 1 scan to indicate done, 2 scans between indicating done and the next problem). The number of scans varied from 19 to 38.

This wide temporal variability creates a considerable challenge with respect to aggregating imaging data from individual trials. To deal with these long latencies and high variability, we used event-locked averaging (Anderson et al., in press). The method takes advantage of the interaction markers that occur as a natural part of a problem-solving episode with these tutors. The event-locked averaging used the 8 intervals defined in Table 2 defined by 8 states (Table 1) in the tutor. Event-locked averaging is a scheme for averaging the BOLD signals from trials consisting of intervals of varying length to obtain a BOLD response for a template that has

\(^2\) Because we wanted to start at a baseline after the 1-back, the first two intervals come from the previous trial – from representation of the problem to the selection of the “Done” operator and then the 4 seconds until the next problem is presented.
intervals of mean length. It assigns the scans from the intervals for particular trials to positions in
the template for purposes of averaging. Figure 5 illustrates how it would choose scans from a
trial (represented above) and assign them to positions in the template (represented below). The
assignment procedure depends on the length \(n\) of the interval in that trial relative to the length \(m\)
of the interval in the template:

1. If \(n\) is greater than or equal to \(m\), the procedure creates a sequence of length \(m\) by taking
   \(m/2\) scans from the beginning and \(m/2\) from the end. If \(m\) is odd, it selects one more from
   the beginning. This means just deleting the \(n-m\) scans in the middle.

2. If \(n\) is less than \(m\), the procedure creates a beginning sequence of length \(m/2\) by taking the
   first \(n/2\) scans and padding with the last scan in this first \(n/2\). It constructs the end
   similarly. If either \(n\) or \(m\) is odd, the extra scan is from the beginning.

The averaging of trials according to this event-locked method creates scan sequences that
preserve the temporal structure of the beginning and end of the sequences but just represent the
approximate average activity in their middle. When there is a rich density of behavioral markers
(as in Table 2) there is little loss of information in the middle of these intervals. This results in
an articulate representation of how the activity in a region changes as problem solving
progresses. For a more detailed discussion of this method see Anderson et al. (in press).

We will use this method of averaging the imaging the data in three separate analyses:

1. An analysis of 6 predefined regions that past research (e.g., Anderson, 2007) has
   indicated to be relevant to the solution of such problems.

2. An exploratory analysis of the whole brain looking for other regions that might be critical
to the learning transitions.
3. A test of whether a cognitive model that performs the task can predict the imaging data.

Analysis of Predefined Regions

Figure 6 illustrates 6 regions that we have repeatedly found to be involved in the learning of mathematical tasks such as algebra (see Anderson, 2007) for a review. They can be grouped into 3 pairs of regions for the current task:

1. There is the motor region where the hand is represented and a region of the fusiform gyrus that has been found to play a critical role in the recognition of visual objects (e.g., Grill-Spector et al., 2004; McCandliss et al., 2003). These regions will track the perceptual-motor aspects of the task. While the data-flow representation is different than standard algebra and the motor actions in the tutor are different than those in typical paper and pencil algebra, regular algebra equation solving does involve significant perceptual-motor components. However, we did not expect any learning to occur in these regions.

2. There are two prefrontal regions that we have found to be critical in the control of the information processing. First, a portion of the lateral inferior prefrontal cortex (LIPFC) has been found to play a critical role in controlling the recall of declarative information (e.g., Buckner et al., 1999; Cabeza et al., 2002; Dobbins & Wagner, 2005; Fletcher & Henson, 2001). Second, we think that a region of the anterior cingulate cortex (ACC) holds the subgoal representations that determine how the information-processing progresses (the end of the paper will discuss other views of the ACC). We predicted decreased activation in the LIPFC with practice because the retrieval of the information gets faster, but no change in the ACC because the control demands remain constant.
3. There are two regions are responsible for the cohesion of the information being operated on and the operations being executed. On the data-representation side the posterior parietal cortex (PPC) appears to hold a representation of information relevant to the problem. Others have found it or nearby regions to be involved spatial processing (e.g., Dehaene et al., 2003; Reichle et al., 2000) and verbal encoding (Clark & Wagner, 2003; Davachi et al., 2001). On the procedural-execution side there is the head of the caudate which is part of the basal ganglia system that some (e.g., Amos, 2000; Frank, et al., 2001; Wise et al., 1996) have speculated serves the function of pattern recognition and selection of cognitive actions. We have sometimes found decreases in activation in these regions if the information processing is re-organized to be more efficient but we have not always found learning-related effects (see Anderson, 2007, for a review). Given the varied outcomes, we did not have strong predictions about these two regions.

Figures 7 – 9 show the results for these pairs of regions organized according to section (1-7 invert versus 2-6 combine) and day. The vertical lines in this graph reflect the occurrence of various states from Table 1 defining the boundaries of the intervals in Table 2. We discuss these figures below:

1. Analysis of activity in the fusiform gyrus and the motor region.

Figure 7 displays the data for the fusiform and motor regions. Part (a) shows the data for the invert problems and part (b) shows the performance for the combine problems. These regions respond both to the execution of a transformation and the evaluation of the result with the BOLD response going down to baseline in the intervals during which the participants are performing the
1-back⁴. The first and last scans for a problem constitute the two baseline scans. To eliminate any linear drift over these relatively long intervals we performed a linear correction so both would have the value of 0.

Qualitatively, these data show some other trends beyond the overall response to the task structure of transformations followed by evaluations. First, there is a consistent small bump in the BOLD response over the first few scans reflecting the processing of redisplay and selecting the done operation (Steps 7 & 8 in Table 1). Second, the fusiform shows a marked decrease from Day 1 to Day 5. Third the response in the motor area rises more rapidly than does the response in the fusiform. These patterns are strikingly similar for the invert section (Figure 7a) and the combine section (Figure 7b).

To determine the significance of such trends we performed two analyses of variance (ANOVA) on the areas under the curves – one for the transform part of the problem and the other for the evaluate part of the problem. We analyzed these two portions of the curve separately because the transformation phase involves more actions than the evaluation phase and is not meaningful to compare them. The scans to the end of the first 1-back (0 – 17 for invert on day 1; 0 – 15 for invert on day 5; 0 – 19 for combine on day 1; 0 – 17 for combine on day 5) were used in the transformation ANOVA and the remaining scans were used in the evaluation ANOVA. To get some idea of differential time course, both the transformation scans and evaluation scans were broken in two halves. Thus, the factors in the ANOVA were region (fusiform versus motor), operation (invert versus combine), day (1 versus 5), and half (first versus second). The

³ Note that the number of scans in this figure is one more than the number of intervals listed in Table 1 because the first and last scans are the last scan of the second n-back.
dependent measure was the area under the curve during each half of the transformation phase or each half of the evaluation phase. All the transformation intervals had an even number of scans and so could be broken right in the middle. In the case of the evaluation interval, some cases were odd in which case the activation for middle scan was divided between the two halves.

With respect to the transformation phase, the two significant main effects were region (F(1,10) = 8.61; p < .05, reflecting the greater activity in the motor region) and half (F(1,10) = 39.92; p < .0001, reflecting greater activity in the second half). These two factors participate in a marginally significant interaction (F(1,10) = 4.85, p = .052) such that the difference between halves is greater in the case of the fusiform (areas under the curve are 0.6% versus 3.3%) than in the case of the motor region (2.5% versus 4.3%). This is consistent with the observation of a more rapid rise in the motor region. The final significant effect is a day-by-region interaction (F(1,10) = 5.53; p < .05) such that the fusiform shows a decrease in activation as function of days (2.5% on day 1 versus 1.4% on day 5) while the motor does not (3.3% on day 1 versus 3.5% on day 5). The effect of practice in the fusiform is significant by a simple t-test (t(10) = 2.25, p < .05)\(^4\) confirming the apparent decrease in activity. Noteworthy, there are no interactions with operations (invert versus combine) implying that the two topics replicate each other. The area under the curve is slightly greater for combine versus invert (2.9% versus 2.5%) perhaps reflecting the extra entry actions (see State 2 in Table 1) but this difference did not reach significance (F(1,10) = 1.55).

\(^4\) All t-tests are two tailed.
With respect to the evaluation phase, the only significant main effect was region ($F(1,10) = 10.33; p < .01$), again reflecting the greater activity in the motor region. The interaction between region and half is highly significant ($F(1,10) = 55.88; p < .0001$), in the same direction as the more marginal interaction for the transformations. The activity actually decreased over halves in the motor (2.1% to 1.8%) but increased in the fusiform (0.8% to 1.3%). The day-by-region interaction, which had been significant in the transformation phase, is now marginal ($F(1,10) = 3.79; p < .15$) but in the same direction as the fusiform showing a decrease as function of days (1.3% on day one versus 0.8% on day five) while the motor does not (2.0% on day one versus 1.9% on day five). The effect of practice in the fusiform is significant by a simple t-test ($t(10) = 2.41, p < .05$). Again, there are no interactions with operations (simple versus combine) implying that the two topics replicate each other. One would not expect a difference, as evaluation is the same for both invert and combine.

Absolute differences in response magnitude between the two regions are not particularly interesting but the interactions with regions are. The interaction with practice confirms the apparent trend in Figure 7 that the fusiform is decreasing in activation – an unpredicted outcome. The other interaction with halves confirms that the fusiform is responding later relatively more than the motor region. Later we will describe a model that attributes the interaction with halves to the fact that visual processing is more involved in the 1-back than the motor. The visual system has to encode each element while the motor system has only to occasionally generate a simple repetitive press.

2. Analysis of activity in the LIPFC and ACC
Figure 8 compares the activity of the LIPFC and ACC for the invert operation (Part a) and the combine operation (Part b). They show a similar response to task structure as Figure 7. To assess the various trends separate ANOVAs were performed on the transformation and evaluation phases.

With respect to the transformation phase, the only significant main effect was half (F(1,10) = 21.46; p < .001), reflecting greater activity in the second half. This is modulated by a highly significant interaction with region (F(1,10) = 28.88; p < .001), such that the difference is much greater in the LIPFC (0.9% in first half and 3.1% in second half) than in the ACC (1.0% in first half and 1.5% in second half). The other significant interaction is between day and region (F(1,10) = 10.24; p < .01), such that the activity decreases in the LIPFC over days (2.4% to 1.5%) but increases in the ACC (1.0% versus 1.5%). The effect of practice in the LIPFC is significant by a simple t-test (t(10) = 2.44, p < .05) while the reverse effect in the ACC is marginal (t(10) = -1.91, p < .10). Again there are no interactions with operations (invert versus combine) implying that the two topics replicate each other.

With respect to the evaluation phase, the significant main effects were operation (F(1,10) = 22.75; p < .001) and day (F(1,10) = 8.02; p < .05). The area under the curve was greater for invert (1.1%) than combine (0.5%), which is somewhat surprising since evaluation is essentially the same process after either transformation. However, the combine section does follow the invert section and so evaluation in the combine section may benefit by the practice in the invert section. The effect of day is modulated by a marginally significant day-by-region interaction (F(1,10) = 3.48; p < .10), such that the reduction with practice is greater for the LIPFC (from
1.0% to 0.5%) than the ACC (from 0.9% to 0.7%). The effect in LIPFC is significant (t(10) = 2.98, p < .05) while the effect in the ACC is not significant (t(10) = 0.96). While operation has a large main effect, once again it participates in no significant interactions implying the two topics again replicate the effects.

Probably the most significant result is that the LIPFC shows strong learning effects while the ACC does not. As noted earlier this was predicted on the basis of reduced retrieval demands on the LIPFC and constant subgoaling demands in the ACC. The other interaction with region was that the ACC became active earlier than the LIPFC.

3. Analysis of activity in the PPC and the caudate.

Figure 9 compares the activity for the parietal and caudate regions for the invert operation (Part a) and the combine operation (Part b). These two regions have often behaved similarly in past experiments but appear to behave quite differently in this experiment.

With respect to the transformation phase, the two significant main effects were region (F(1,10) = 29.49; p < .001, reflecting the low activity in the caudate) and half (F(1,10) = 51.48; p < .0001, reflecting greater activity in the second half). These two factors participate in a highly significant interaction (F(1,10) = 78.82; p < .0001) reflecting the fact that activity increased greatly in the second half in the parietal (1.6% to 4.6%) while it decreased in the caudate (0.5% to 0.4%). The only other significant interaction was between region and operation (F(1,10) = 10.57; p < .01) reflecting the fact that the activity was greater in the caudate for the invert operation than the combine (0.7% versus 0.2%) but this effect was reversed in the parietal (2.8%
versus 3.4%). With respect to the evaluation phase there is only a significant main effect of region (F(1,10) = 24.84; p < .001). The factor of days did not participate in any significant effects in either analysis.

The two noteworthy results in these analyses were the lack of effects of practice and the interaction between region and half. As in the case of the ACC, the caudate is distinguished by relatively high activation early. Its activation is particularly low throughout the evaluation phase. The average area under the curves in Figure 9 during the evaluation phase in the caudate is 0.

**Summary**

Four of the 6 predefined regions showed the predicted relationship to the task structure and the learning manipulation (Day 1 versus Day 5). As predicted, the Motor, ACC, and PPC all showed a response to task structure but did not show an effect of learning. Also as predicted the LIPFC showed a response to task structure and a decrease with experience. The fusiform showed the predicted response to task structure but unexpectedly also showed learning-related decreases. Besides their differential response to the learning manipulation the regions differed in their degree of involvement early in an operation. Some regions (motor, ACC, and caudate) showed about as much activation late as early whereas others (fusiform, LIPFC, PPC) showed more than twice as much activation late than early. It will be challenge to the model to produce these differential timings of involvement.

We were interested if any of these regions would predict individual differences in performance. For each region for each participant we collected both a measure of total activity (area under the
curve) averaged over days and a difference between activity on the first day and the fifth day. The performance measures were similarly calculated for mean accuracy across days, mean latency, and changes in these two measures from Day 1 to Day 5. Table 3 presents the correlation between these performance measures and the activity measures for the 6 predefined regions plus 2 of the exploratory regions to be discussed. Under “Mean Performance” it reports the correlation of mean activity in each region with the performance measures. Under “Performance Change” it reports the correlation of change in activation with the change in performance measures. There were two significant correlations for the predefined regions. The first is the correlation between mean ACC activity and error rate such that activity was greater when error rate was higher. Note the activity is measured on error-free trials and thus does not reflect response to errors per se as some theories of the ACC would expect (e.g., Falkenstein et al., 1995; Gehring et al., 1993). Rather, this would seem to support the error-likelihood theory (Brown & Braver, 2005) that relates ACC activity to a general measure of error rate. Second is the correlation between the decrease in latency from Day 1 to Day 5 and the corresponding decreased activity in the LIPFC, supporting the prediction that declarative learning is critical to the latency gains.

**Exploratory Analysis**

We defined a statistical contrast (illustrated in Figure 7 for Invert problems) to find regions that showed reduced activation from Day 1 to Day 5. The analysis only used the same scans that went

---

5 Note that, because each participants’ activity is warped to the same number of scans, this correlation is not confounded with number of scans.
into the predefined analysis and so only included correct trials for Sections 1-7 and 2-6. For Day 1 the contrast weighted positively all scans but the one before the first and last by the number of times they contributed to the data in the event-locked procedure illustrated in Figure 5. The first and last scans where weighted negatively by \(-n/2\) where \(n\) are the number of middle scans. This yields a measure of the amount that scans for the trial rise above the baseline defined by the beginning and last scans. To get a measure of learning, the contrast subtracted this measure for Day 5 from the measure for Day 1. As illustrated in Figure 7 this comes down to inverting the signs of the weights for Day 5. The weights defined by these numbers were regressed against the sequence of BOLD values for each voxel. The threshold for including a region was that at least 25 contiguous voxels were significant at the .01 level.

Although this analysis was capable of finding regions whose activity increased with days it only found regions whose activity decreased. The 10 areas reaching threshold are given in Table 4 and illustrated in Figure 10 along with the predefined ROIs. As can be seen, many of the regions overlap with the predefined regions. Regions b and c overlap with the right and left LIPFC. Regions d and e overlap with the left PPC. Regions h and i overlap with the left and right fusiform. With respect to the LIPFC and fusiform, this confirms learning effects that were observed in the predefined regions. In addition there were four new areas – a left premotor, a right anterior prefrontal, and left and right occipital areas. The average intercorrelation among the premotor and occipital regions was .956 while their average correlation with the anterior prefrontal was .604. Therefore, we decided to look, in more detail, at the premotor (BA6) and anterior prefrontal (BA10) as representative of the two patterns of activation. Figure 11 compares the activity of these two regions for the invert operation (Part a) and the combine operation (Part
b). As in the analyses of predefined regions we performed separate ANOVAs on the transformation phase and the evaluation phase.

With respect to the transformation phase, there were three significant effects: day (F(1,10) = 25.81; p < .001, reflecting the factor that selected these regions), region (F(1,10) = 23.35; p < .001, reflecting the much stronger activation in the premotor region), and half (F(1,10) = 36.46; p < .0001, reflecting greater activity in the second half). There is a significant interaction between these regions and half (F(1,10) = 11.01; p < .01) but it is a rather technical interaction. The difference between first and second half is greater for the premotor region (2.3% versus 4.7%) than the difference for the anterior prefrontal region (0.4% versus 1.2%) but this just reflects the greater magnitude of the response in the premotor region. Proportionately, the response is twice as large in the second half than the first half for the premotor but three times as large in the anterior prefrontal region.

With respect to the evaluation phase, the significant main effects were day (F(1,10) = 18.58; p < .005) and region (F(1,10) = 36.31; p < .0001). There is a significant interaction between half and region (F(1,10) = 20.98; p < .001) again but this time it is a more interesting interaction. The response decreases from first half to second half in the premotor (1.9% versus 1.6%) while it increases in the anterior prefrontal (-0.3% to 0.4%). Also, there is a significant 3-way interaction between day, half, and region (F(1,10) = 6.32; p < .05). This interaction is driven by the fact that the region-by-half interaction becomes more extreme on Day 5. Note that, as tended to be the case with the predefined regions, these regions are not sensitive to the operation (invert versus
combine). Operation showed no significant effects nor was it involved in any interactions. Thus, the two sections of the curriculum produced the same pattern of results.

The anterior prefrontal region shows a pattern that not seen in any other region. Averaging over transformation and evaluation and over invert and combine, the rise from baseline in the first half is not significant (0.16% effect; t(10) = 0.05) while it is in the second half (0.81% effect; t(10) = 2.54, p < .05). In fact, except for the first day transformation phase the first half response averages a negative response. The evaluation step has already been practiced in earlier sections while the transformation step is new. It seems that once an operation is familiar, this region only responds after the algebra is done and the participant is into the 1-back. In episodic memory experiments this region also gives a late positive response that can peak after the motor area controlling response generation (e.g., Reynolds et al., 2005; Rugg et al., 2003). In a geometry proof task this region rises upon completion of the proof (Kao et al, in revision). Perhaps it is a metacognitive region whose activity signals reflection on a past activity.

Table 3 indicates a quite significant correlation between error rate and activity of this anterior prefrontal region. In this regard it might seem similar to the ACC, which also showed a significant correlation with error rate. However, the overall temporal profile in this region correlates more weakly with the ACC (r = .37) than it does with any other region (these r’s are from .41 to .70). This low correlation reflects both the fact that this region’s activity is late while the ACC tends to be active early and the fact that this region shows learning-related changes while the ACC does not. One might speculate that BA10 activity is greater in participants who are making more errors because they need to reflect more on what is correct versus what is not.
Curiously, while this region has a strong correlation with error rates it shows a strong decrease across days the error rate does not decrease across days. The correlation ($r = .40$) is only marginal between individual differences in changes in error rates and individual differences in learning-related changes in the activity of this region.

**Testing a Cognitive Model**

Figure 4 showed the behavioral predictions of a cognitive model\(^6\) developed in the ACT-R architecture (Anderson, 2007; Anderson et al., 2004). We modified that model to deal with the new interface and the 1-back. ACT-R assumes that cognition emerges through the interaction of a set of modules. Figure 12 illustrates the activity of the six modules relevant to the performance of this task in the solution of a problem in Section 1-7 on day 1. The figure is broken into two panels, each spanning about 25 seconds going down the panel. The first panel begins at the end of 1-back for the previous panel and the second panel stops in the middle of the second 1-back; so the two panels reflect slightly less than a full problem. Each panel has marked the events that are used to break the imaging data into intervals (see Table 2). The 6 columns in each panel represent the activity of the 6 modules and the length of the boxes in the columns reflect the durations of the module engagements. Below we review these modules in left-to-right order as they appear in Figure 12:

\(^6\) The model which runs with the ACT-R software in Macintosh Common Lisp (MCL) can be downloaded at the ACT-R models web-site (http://act-r.psy.cmu.edu/models/) under the title of this article. Running instructions are in the file read&start.lisp. This also contains the experimental algebra tutor.
1. The visual module is active throughout the performance of the task. It takes a time uniformly distributed between 100 and 300 milliseconds to encode a box or button and a time uniformly distributed between 25 and 75 milliseconds to encode the letters. This accounts for the difference in the size of the boxes when working on the problem and when working on the 1-back.

2. The procedural module represents the firing of productions that unpack the logic involved in solving an equation of this form. Each of the small boxes in the procedural column represents 50 milliseconds to fire a production (a fixed parameter of the ACT-R theory). These productions fire throughout the task, but vary in their frequency of firing in different parts of the task reflecting the different amounts of cognitive engagement at those points.

3. The goal module holds the various subgoals that control the solution. The changes between subgoals are instantaneous in ACT-R and are reflected in Figure 12 by lines. As a general rule a subgoal is associated with executing each action (i.e., click a box, an operator, enter a result) and monitoring that the correct change has taken place with the interface. In addition, subgoals are associated with bridging the delays and 1-back.

4. The retrieval or declarative module is occupied retrieving task instructions about what to do next and retrieving various algebraic and arithmetic facts. The smaller boxes reflect retrieval of information about what to do next while the larger two boxes reflect retrieval of general facts. The retrieval of task information is represented as faster than the retrieval of algebraic and arithmetic facts because the task information repeats every problem and so should be highly active. The box labeled 1 reflects retrieval of an algebraic fact (in this case, that / is the inverse of *) while the box labeled 2 reflects
retrieval of the arithmetic fact (in this case that \(40 / 5 = 8\)) that will be required in the evaluation step. Note that retrieval of the arithmetic fact happens as soon as the arithmetic expression has been constructed which is at the beginning of the 1-back period. The time for the retrievals of instructions varied uniformly for 0 to 1.2 seconds while the time for the retrieval of facts varied uniformly from 0 to 3.25 seconds.

5. The imaginal module builds up representations of the contents of the various boxes attended. The time for these imaginal updates was .2 seconds, which is a frequently used value in ACT-R.

6. The manual module is activated to program control of mouse movement. Based on earlier research we used the following estimates for motor time: 100 milliseconds for a simple click during the 1-back, 400 milliseconds to move to and click a box, 800 milliseconds to move and click an operator (because of the longer distance movements), and 400 milliseconds per character entered into the pop-up menu. These times vary uniformly trial to trial from half to one-and-a-half times these mean values.

A number of values were estimated in fitting these latencies. These include the mean visual, retrieval, and manual times and their variability. These were estimated to fit the distribution of latencies in Figure 4. In addition, to fit the speed up from Day 1 to Day 5 we assumed that practice had decreased the retrieval times by a factor of .5 and the visual encoding times by a factor of .667. ACT-R predicts the decrease in retrieval times because of the increase in strength of declarative facts with practice. However, the decrease in visual encoding times was a post hoc assumption based on the decrease in the fusiform. The ACT-R architecture does not have any mechanism yet to produce such a perceptual learning effect. Given that this many parameters
were estimated, the match between theory and latency data in Figure 4 is not very compelling evidence for the model. The more demanding tests concern the ability to predict the BOLD responses in the associated regions given these latency estimates.

**Predicting the BOLD Response**

The patterns of activity of the modules in Figure 12 lead to predictions for the patterns of activity that should be seen in associated brain regions as in Figures 7-9 (Anderson, 2007; Anderson et al., in press). These predictions are derived by treating the activity in Figure 12 like as a 0-1 boxcar function (much like a design matrix in typical a typical SPM analysis –Friston et al., 2006) and convolving this with a standard hemodynamic response represented as a gamma function (e.g., Glover, 1999):

$$H(t) = m(t/s)^a e^{-t/(s)}$$

The parameter $m$ is the magnitude parameter and determines the height of the function, the parameter $s$ is the scale parameter and determines the temporal spread, and the parameter $a$ is the shape parameter and determines the narrowness of the function. In this paper $a$ is fixed to be 6, $s$ to be 0.75 seconds, and estimate the magnitude to fit the magnitude $m$ of response in a brain region.

---

7 The setting of the parameters $a$ and $s$ come from another experiment (Kao et al., submitted) where we just had participants press a finger to a visual signal and tried to model the response in the motor area. Handwerker et al. (2004) present data indicating that the variability in the BOLD response across regions is small relative to the variability across participants.
Because there were few significant differences between the BOLD responses for the invert problems of section 1-7 and the combine problems of section 2-6 we combined them into a single profile. The aggregate intervals used for this event locking are displayed in Table 2. The trials from the model were averaged according the same event-locked procedure as the data. The fits of the model predictions to the 6 predefined regions are displayed in Figure 13. The figure gives the correlations between the pattern predicted by the module and the observed data. These correlations do not depend on the estimation of any parameter.

The magnitude of the predicted BOLD response depends on estimating a magnitude parameter for each region. This parameter was set to minimize the following measure of deviation:

$$\sum_i \left( \frac{\hat{B}_i - \bar{B}_i}{s_{\hat{B}_i}} \right)^2$$

**Sum of Squared Standard Errors of Prediction (SSSEP)**

which sums the squares of the deviations between the predicted and the mean BOLD responses for scan $i$ divided by the standard errors of the means. Under the null hypothesis that these are just random deviations this statistic has an expected value equal to the number of data points fit minus the number of parameters estimated. Anderson et al. (in press) advocate estimating the magnitude parameter to fit one curve and then testing parameter-free its predictions for another curve. For this experiment, we estimated the magnitude parameter to fit the Day 1 data and predicted the Day 5 data. Given 27 non-zero BOLD values for Day 1 and 23 for Day 5, the expected values for the SSSEP measures are 26 (since one parameter is measure to fit that data) for Day 1 and 23 for Day 5. Values much larger than this would indicate significant deviations. Anderson et al. show that under the null hypothesis the SSSEP statistic has a gamma distributions and gives a formula for calculating the parameters of the gamma distribution that
deals with the non-independence of successive points on the BOLD response. A gamma distribution can be characterized by an index \( \alpha \) and a scale \( \beta \) and these can be calculated as:

\[
\alpha = \frac{n}{2} S(n, r^2)
\]

\[
\beta = 2 \cdot S(n, r^2)
\]

\[
S(n, r) = 1 + \frac{2r^2}{1 - r^2} \left( 1 - \frac{1 - r^{2n}}{n(1 - r^2)} \right)
\]

where \( r \) is the correlation of the error of measurement between adjacent points on the BOLD response and \( n \) is the expected value of the SSSEP statistic. Given an estimate of .837 from our data for \( r \), the gamma distribution for 26 points has index 2.56 and scale 10.14 while for 23 points the index is 2.30 and the scale is 9.98. With these parameters the .05 critical value for significant deviations is 57.2 for the Day 1 deviations and 53.2 for the Day 5 deviations.

Anderson et al. note that testing for critical values has the undesirable consequence of rewarding noise in the data. If the measurement of the data is noisy it is unlikely that models that have serious problems will reach the critical value. On the other hand, with enough precision in measurement any model will produce a critical value because there will be always some small and unimportant deviation. Therefore, it is also useful to compare the relative goodness of fit of a model with that of alternative models. We will do this by trying to fit each of the 6 modules to each of the 6 predefined regions.

Table 5 provides the sums of standardized deviations obtained by fitting each module to each region for Day 1 in part (a) and for Day 5 in part (b). As can be seen, most of the sums exceed their critical values, but different modules show different degrees of fit to different regions. Part (c) adds the log likelihood for the two sums given the assumed gamma distributions to get an
overall measure of how well a module fits a region. Below we will discuss the fits in Figure 13 to each of the regions (columns in Table 5). In a comparison of two modules (rows) as fits to a region, for each 2.3 units that one of the log likelihoods is less negative it is 10 times more likely.

**Fusiform.** The visual module produces the best fit to the fusiform, as hypothesized, but with significant deviations. The deviations, however, are rather technical in character – the observed BOLD response returns to baseline during the 1-back one scan later than predicted. Removing these next-to-last scans in the 1-back from the SSEP statistic removes the significant deviations.

**Motor.** The motor region is best fit by the manual module, as hypothesized. The closest competing module is the imaginal which is one sixth as likely. The significant deviations in the fit of the manual module to the Day 1 motor profile reflect a failure to predict the dip below baseline during the first 1-back and the slow return to baseline during the second 1-back. Again, these seem rather technical difficulties that do not challenge the fundamental correspondence between module and region.

**LIPFC.** There are no significant deviations in the fit of the retrieval module to the LIPFC and it fits 10 times better than the closest competitor (the visual module). While not significant, the module does somewhat overpredict the reduction in activation from Day 1 to Day5 – empirically, the area under the curve on Day 5 is 52% of the area on Day 1 whereas 45% is predicted.
ACC. There are no significant deviations in the fit of the goal module to the ACC and it fits 15 times better than the closest competitor (the imaginal module). While not significant, there is a qualitative mismatch in the predictions of the goal module. The goal module predicts a double peak in the ACC on both Days 1 and Day 5 during the transformation phase. The source of this can be seen in bottom half of the left side of Figure 12 – a single subgoal controls the keying of the expression (e.g., 90/5) which takes a long time allowing the BOLD response to drop off. That predicted dip occurs on Day 1 but not Day 5.

PCC. The PPC is best fit by the hypothesized imaginal module but there are significant deviations on Day 5. The model again predicts the BOLD response goes down to baseline one scan sooner that it does. Without the last non-zero point the SSSEP statistics would be 18.17 for Day 1 and 34.54 for Day 5. Thus, again the model predicts a more rapid drop to zero at the end than is empirically observed.

Caudate. In contrast to the relatively good fit for the other modules, the procedural module gives the worst fit of the 6 modules to the activity in the caudate (see Table 5c). As Figure 12 illustrates, relatively few productions fire early because of the 4 seconds of dead time inserted between the keying of the next problem operation for the previous problem and the presentation of the next problem. However, the beginning is the period of greatest caudate activity. Unfortunately, as is apparent in Figure 13c, the noise is high relative to the magnitude of the caudate response making it hard to ascertain the actual pattern of activity in this region. However, it clearly does not match the predictions of the procedural module.
We were surprised, given the strong influence of the task structure, how discriminative these results were. The 6 module-to-regions mapping proposed by the ACT-R theory is more than 170 times more likely than the nearest 1-1 mapping (which involves assigning the imaginal module to the motor region and the manual module to the parietal region). If we consider the assignments of the visual to the fusiform, and the manual to the motor established, and only consider various assignments of retrieval, goal, and imaginal, procedural to LIPFC, ACC, PPC, and caudate the ACT-R assignment is over 90,000 times more likely than the next best 1-1 mapping (which involves assigning goal to PPC and imaginal to ACC).

With the exception of the caudate-procedural mapping the mappings lead to good predictions, usually differing only by an undershoot or the speed of the final drop-off in the BOLD response (retrieval to LIPFC, imaginal to PPC). The one qualitative failure for these regions was that the ACC did not show the mid-transformation drop on Day 5.

**Discussion**

This research has indicated that fMRI can be used to analyze learning in an intelligent tutoring system. The richness of interaction with a tutor creates many behavioral markers for segmenting the interactions into short sequences of scans. Event-locked averaging allows extraction of a meaningful time course of processing.

Much of the analysis was focused on the response obtained from 6 predefined regions whose functional significance was identified in previous research on simpler tasks. While all regions
showed a rise during the algebraic task (transformation or evaluation) and a drop off during the 1-back, they differed in the pattern of activation during the periods of engagement. All regions except the caudate showed more activation in the second half of these intervals than the first half but that ratio varied from 3.3 times greater in the fusiform to 1.2 in the ACC. The reason for the heavy late activity in the fusiform was that this region was engaged in encoding the results of the transformations and the elements of the 1-back. The reason for the relatively low, late activity in the ACC was that it was heavily engaged in early strategy selection.

In addition to these effects of task structure we did find the predicted effects of learning in the LIPFC region, reflecting the reduced retrieval effort with practice. However, we also found an unexpected reduction in the activation of fusiform region that was of equivalent proportional magnitude. This surprising result suggests that an important ingredient in the learning was an increased facility to parse the data-flow representations. It remains to be seen whether children will show a comparable decrease when they practice with regular algebraic expressions. The learning-related decreases in the fusiform are consistent with other reports of priming (e.g., McCandless et al., 2003) and reports of training-induced decreases (e.g., Xue et al., 2006).

The exploratory analysis found a number of regions showing learning effects. Of those that did not overlap with predefined regions the anterior prefrontal region seemed the most interesting. Its activity correlated quite strongly with individual differences in error rate. Its activity was also unusual in that it tended to peak after completion of an operation. We suggested that this region might be involved some sort of post-task reflection. Numerous theories of the role of this region have a certain metacognitive flavor – such as monitoring of memory retrieval (e.g., Buckner,
2003; Rugg et al., 2004), branching (Koechlin & Hyafil, 2007), shift between external and internal processes (Burgess et al., 2007), and relational integration (Christoff et al., 2001).

The results are also relevant to interpretations of activity in the anterior cingulate cortex. The results from this experiment are not consistent with the error-based interpretations (e.g., Falkenstein et al., 1995; Gehring et al., 1993) of the ACC. A BOLD response occurs in the ACC on trials without errors and so it cannot be a response to errors. The fact that its activity is correlated with overall error rate is compatible with the error-likelihood theory (Brown & Braver, 2005) of the anterior cingulate. However, since its activity varies over the course of a trial it cannot be responding to an estimate of the overall error rate for that trial. The ACT-R theory relates its activity to the need to set different control states. The highly interactive nature of the task means the predictions of ACT-R’s goal module are highly correlated with its manual module (r = .925) since different motor actions require different goal states. Therefore, the manual module is a close competitor to the goal module for predictions of ACC activity. Thus, the results of this experiment can be viewed as largely consistent with response conflict theories of ACC activity (Botvinick et al., 2001; Carter et al., 2000; Yeung et al., 2004).

While the fits of most modules were relatively good, we never have done an experiment that so definitely indicated that the procedural module does not match up with the caudate. We have found satisfactory fits to the caudate in some experiments that are of short duration (e.g., Anderson, 2005). In experiments with more extended tasks (e.g., Anderson et al., submitted; Anderson & Qin, in press) we have tended to find the same early rise in caudate activity. The
procedural module provided such a bad fit in this case because this early caudate rise occurred during a period where the module predicted no activity at all.

To conclude by addressing the title of this article, this paper has shown that imaging can help identify the components that are undergoing change in an educational task. The learning-related changes in the imaging data suggested increased fluency in declarative retrieval (LIPFC), increased efficiency in perceptual processing (fusiform), and decreased metacognitive monitoring (anterior prefrontal cortex). The artificial nature of the problems (data-flow diagrams) and the advanced level of the students (Carnegie Mellon undergraduates) prevent using these learning results directly for educational recommendations. However, were they to replicate with children learning algebra and relate to out-of-scanner achievement gains they could be used to suggest where to place emphasis in educational programs. For instance, consider the different possible implications should one of these regions and its associated function prove critical:

- **LIPFC**: If the critical factor were strengthening of specific declarative information then simple practice of the relevant facts might be the recommendation. This might come down to just studying the formulas and rules.

- **Fusiform**: If the critical factor were the fluency with which students could parse algebraic expressions a program of practice in reading algebra might be recommended. This might come down to a mathematical reading program.

- **Anterior Prefrontal**: If the critical factor were reasoning about the domain them a program that stressed algebraic reasoning might be recommended. For instance, students might be encouraged to justify their transformations by algebraic constraints.
Of course, any such recommendations await more research.
ACKNOWLEDGMENTS

This research was supported by NSF award REC-0087396 and DARPA grant AFOSR - FA9550-07-1-0359.
References


Table 1

8 States of the Tutor for Sections 1-7 and 2-6 and the Events that Intervene

1. **Problem.** A problem like Figure 3a or 3d is presented. The participant clicks two boxes to operate on and then the transformation operator (“Invert” or “Combine”).

2. **Transform.** The display changes to a state not shown in Figure 3 but is illustrated in Figure 1b for a combine operation.
   
a. In the case of an invert transformation, a green box is brought up, the participant selects it by clicking, a keypad is brought up and the participant enters an expression that undoes the arithmetic operation (e.g., “90/5” in the transition between Figures 3a and 3b).

b. In the case of a combine operation two boxes are highlighted in green for entry -- a large box for the expression and a small box for an operator (as Figure 1b). The participant can click on these in either order. When a box is clicked a keypad will be brought up for entering the expression or operator.

3. **Transformation Filled In.** A display like Figure 3b or 3e would appear on the screen for 1 second. The problem would be removed from the screen and crosshairs would be presented on the screen for 1 second followed by a sequence of 8 letters each for a second separated by .25 seconds. The participant was to press the mouse if a letter repeated – which happened on a random 1/3 of the presentations. After the 8th letter the crosshairs would appear again for a second.

4. **Problem Again.** The display from step 3 reappears. The participant selects the box with the expression to evaluate and selects the “Evaluate” operator.
5. **Evaluate.** A green square appears where the evaluation will go, the participant selects it by clicking, the keypad appears, and the participant enters the value in a keypad.

6. **Evaluation Filled In.** A display like Figure 3c or 3f would appear on the screen for 1 second and the participant would go through the same 1-back procedure as described in 3 above.

7. **Problem Last Time.** The display from step 6 reappears and the participant selects the “Next Problem” operator.

8. **Done.** Crosshairs appear on the screen for 4 seconds and then the next problem is presented.
Table 2

Mean Number of Scans in the 8 Behaviorally Defined Intervals

<table>
<thead>
<tr>
<th>Interval to</th>
<th>Invert Day 1</th>
<th>Invert Day 5</th>
<th>Combine Day 1</th>
<th>Combine Day 5</th>
<th>Aggregate Day 1</th>
<th>Aggregate Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Done</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1. Problem</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2. Transform</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3. Fill-in</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>4. Problem</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5. Evaluate</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6. Fill-in</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7. End of 1back</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>23</td>
<td>28</td>
<td>25</td>
<td>28</td>
<td>24</td>
</tr>
</tbody>
</table>
Table 3

Correlations between Regional Activation and Performance
(*p < .05; **p < .01 – one tailed)

<table>
<thead>
<tr>
<th></th>
<th>Mean Performance</th>
<th>Performance Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Latency</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Fusiform</td>
<td>0.195</td>
<td>0.184</td>
</tr>
<tr>
<td>Motor</td>
<td>-0.282</td>
<td>0.314</td>
</tr>
<tr>
<td>LIPFC</td>
<td>-0.207</td>
<td>0.403</td>
</tr>
<tr>
<td>ACC</td>
<td>0.032</td>
<td>0.535*</td>
</tr>
<tr>
<td>PPC</td>
<td>0.294</td>
<td>0.509</td>
</tr>
<tr>
<td>Caudate</td>
<td>-0.289</td>
<td>0.274</td>
</tr>
<tr>
<td>BA6</td>
<td>-0.285</td>
<td>0.395</td>
</tr>
<tr>
<td>BA10</td>
<td>0.228</td>
<td>0.689**</td>
</tr>
</tbody>
</table>
Table 4

Regions showing a decrease from Day 1 to Day 5

<table>
<thead>
<tr>
<th>Region of interest</th>
<th>Brodmann area(s)</th>
<th>Voxel count</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>Invert Day 1</th>
<th>Invert Day 5</th>
<th>Comb Day 1</th>
<th>Comb Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Left Premotor</td>
<td>6</td>
<td>35</td>
<td>-27</td>
<td>-4</td>
<td>59</td>
<td>13.5%</td>
<td>7.9%</td>
<td>11.1%</td>
<td>8.4%</td>
</tr>
<tr>
<td>b. Right Prefrontal</td>
<td>6, 9,46</td>
<td>123</td>
<td>46</td>
<td>19</td>
<td>39</td>
<td>5.9%</td>
<td>3.2%</td>
<td>6.2%</td>
<td>0.4%</td>
</tr>
<tr>
<td>c. Left Prefrontal</td>
<td>6, 9,46</td>
<td>178</td>
<td>-44</td>
<td>11</td>
<td>34</td>
<td>9.4%</td>
<td>5.7%</td>
<td>8.4%</td>
<td>4.6%</td>
</tr>
<tr>
<td>d. Left Precuneus</td>
<td>7</td>
<td>34</td>
<td>-12</td>
<td>-73</td>
<td>43</td>
<td>15.0%</td>
<td>10.1%</td>
<td>15.1%</td>
<td>12.4%</td>
</tr>
<tr>
<td>e. Left Precuneus/Cuneus</td>
<td>19, 8</td>
<td>91</td>
<td>-26</td>
<td>-76</td>
<td>34</td>
<td>13.3%</td>
<td>9.5%</td>
<td>13.0%</td>
<td>9.1%</td>
</tr>
<tr>
<td>f. Right Anterior Prefrontal</td>
<td>10</td>
<td>30</td>
<td>41</td>
<td>47</td>
<td>3</td>
<td>1.9%</td>
<td>-0.1%</td>
<td>2.4%</td>
<td>0.1%</td>
</tr>
<tr>
<td>g. Right Occipital</td>
<td>18, 19</td>
<td>253</td>
<td>24</td>
<td>-72</td>
<td>-5</td>
<td>12.8%</td>
<td>7.4%</td>
<td>13.4%</td>
<td>9.1%</td>
</tr>
<tr>
<td>h. Left Fusiform</td>
<td>37, 20</td>
<td>32</td>
<td>-53</td>
<td>-52</td>
<td>-4</td>
<td>6.5%</td>
<td>3.7%</td>
<td>6.3%</td>
<td>2.8%</td>
</tr>
<tr>
<td>i. Right Fusiform</td>
<td>37, 20</td>
<td>28</td>
<td>57</td>
<td>-49</td>
<td>-5</td>
<td>5.5%</td>
<td>2.2%</td>
<td>3.1%</td>
<td>0.3%</td>
</tr>
<tr>
<td>j. Left Occipital</td>
<td>18,19</td>
<td>203</td>
<td>-25</td>
<td>-73</td>
<td>4</td>
<td>17.1%</td>
<td>10.6%</td>
<td>16.7%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>
Table 5
Results of Fitting Different Modules to Different Predefined Regions.

<table>
<thead>
<tr>
<th>(a) Day 1</th>
<th>Fusiform</th>
<th>Motor</th>
<th>LIPFC</th>
<th>ACC</th>
<th>PPC</th>
<th>Caudate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>104.91</td>
<td>104.20</td>
<td>78.76</td>
<td>82.48</td>
<td>42.67</td>
<td>35.53</td>
</tr>
<tr>
<td>Manual</td>
<td>143.58</td>
<td>56.42</td>
<td>48.39</td>
<td>76.34</td>
<td>49.53</td>
<td>29.52</td>
</tr>
<tr>
<td>Retrieval</td>
<td>288.06</td>
<td>177.39</td>
<td>29.44</td>
<td>79.68</td>
<td>123.74</td>
<td>24.01</td>
</tr>
<tr>
<td>Goal</td>
<td>198.65</td>
<td>157.40</td>
<td>66.66</td>
<td>44.03</td>
<td>126.22</td>
<td>17.18</td>
</tr>
<tr>
<td>Imaginal</td>
<td>122.58</td>
<td>75.90</td>
<td>40.83</td>
<td>78.47</td>
<td>37.25</td>
<td>23.11</td>
</tr>
<tr>
<td>Procedural</td>
<td>998.80</td>
<td>362.11</td>
<td>335.88</td>
<td>266.31</td>
<td>486.24</td>
<td>85.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) Day 5</th>
<th>Fusiform</th>
<th>Motor</th>
<th>LIPFC</th>
<th>ACC</th>
<th>PPC</th>
<th>Caudate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>51.48</td>
<td>103.85</td>
<td>25.67</td>
<td>80.23</td>
<td>119.75</td>
<td>62.32</td>
</tr>
<tr>
<td>Manual</td>
<td>122.57</td>
<td>42.60</td>
<td>133.96</td>
<td>40.08</td>
<td>92.20</td>
<td>82.95</td>
</tr>
<tr>
<td>Retrieval</td>
<td>143.90</td>
<td>205.04</td>
<td>41.60</td>
<td>114.90</td>
<td>241.12</td>
<td>67.61</td>
</tr>
<tr>
<td>Goal</td>
<td>128.01</td>
<td>95.38</td>
<td>129.87</td>
<td>28.30</td>
<td>83.13</td>
<td>48.03</td>
</tr>
<tr>
<td>Imaginal</td>
<td>135.17</td>
<td>47.32</td>
<td>102.20</td>
<td>32.78</td>
<td>60.99</td>
<td>66.83</td>
</tr>
<tr>
<td>Procedural</td>
<td>167.83</td>
<td>111.96</td>
<td>126.26</td>
<td>65.81</td>
<td>133.85</td>
<td>95.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) Log Likelihood</th>
<th>Fusiform</th>
<th>Motor</th>
<th>LIPFC</th>
<th>ACC</th>
<th>PPC</th>
<th>Caudate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>-14.80</td>
<td>-19.07</td>
<td>-11.00</td>
<td>-15.27</td>
<td>-15.82</td>
<td>-10.50</td>
</tr>
<tr>
<td>Retrieval</td>
<td>-39.14</td>
<td>-34.69</td>
<td>-8.65</td>
<td>-18.06</td>
<td>-33.38</td>
<td>-10.40</td>
</tr>
<tr>
<td>Imaginal</td>
<td>-23.42</td>
<td>-12.15</td>
<td>-14.16</td>
<td>-11.37</td>
<td>-10.49</td>
<td>-10.31</td>
</tr>
<tr>
<td>Procedural</td>
<td>-109.22</td>
<td>-43.19</td>
<td>-42.00</td>
<td>-30.32</td>
<td>-56.88</td>
<td>-16.79</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. Illustration of the interface for the algebra tutor during a transformation fill in: (a) The linear representation; (b) The data-flow representation.

Figure 2. The data-flow equivalents of (a) 13 - 13-4x = 25 and (b) (2x − 5x) + 13 + 9x = 67.

Figure 3. The invert problems in Section 1.7 go through two episodes: first State (a), is inverted to achieve State (b) and then this is evaluated to achieve state (c). The combine problems in Section 2.6 similarly go through two episodes: first State (d) is combined to achieve State (e) and then this evaluated to achieve state (f). The actions to achieve these transformations are specified in Table 1.

Figure 4. Distribution of number of scans to solve problems – Part (a) shows the distribution for invert problems and part (b) shows the distribution for combine problems. Dotted lines connect the observed proportions of solutions of various lengths and solid lines show the predictions of the ACT-R model.

Figure 5. An illustration of how the scans from the 8 intervals in particular trials might be assigned to the template for invert problems on Day 1 and Day 5 for the purpose of averaging. The numbers associated with the scans on the trial are used in the statistical contrast that looks for an effect of learning in the exploratory analyses.
Figure 6. An illustration of the locations of the 6 brain regions associated with ACT-R modules. In part (a) are the regions close to the surface of the cortex and in part (b) are the regions deeper in the brain. The Talairach coordinates are for the left side, which is where the data will be reported from. Most of the regions are cubes 5 voxels long, 5 voxels wide, and 4 voxels high. The exceptions are the procedural (caudate), which is 4 x 4 x 4 the goal (ACC), which is 5 x 3 x 4, and the fusiform which is 5x5x3. A voxel in our research is 3.125 mm long and wide and 3.2 mm high.

Figure 7. A comparison of the BOLD response in the fusiform and motor region for (a) the invert Section 1.7 and (b) the combine Section 2.6. The vertical lines are the boundaries of the intervals defined in Table 2.

Figure 8. A comparison of the BOLD response in the LIPFC and the ACC for (a) the invert Section 1.7 and (b) the combine Section 2.6. The vertical lines are the boundaries of the intervals defined in Table 2.

Figure 9. A comparison of the BOLD response in the PPC and the caudate for (a) the invert Section 1.7 and (b) the combine Section 2.6. The vertical lines are the boundaries of the intervals defined in Table 2.

Figure 10. A representation of the 10 exploratory ROIs (colored and lettered – see Table 3) and their relationship to the 6 predefined ROIs (in black compare to Figure 6).
Figure 11. A comparison of the BOLD response in the BA 6 (Region a in Figure 10, Table 3) and the BA 10 (Region f in Figure 10, Table 3) for (a) the invert Section 1.7 and (b) the combine Section 2.6.

Figure 12. Representation of (left-to-right) the visual, procedural, goal, retrieval, imaginal, and manual modules in the solution of an invert (Section 1.7) problem (see Figure 3a) by ACT-R on Day 1. Each of the 2 columns represents 25 seconds of problem solving. The box labeled 1 reflects retrieval of an algebraic fact (in this case that / is the inverse of *) while the box labeled 2 reflects retrieval of the arithmetic fact (in this case that 40 / 5 = 8) that will be required in the evaluation step.

Figure 13. Comparison of predictions and data for the 6 predefined regions of the experiment for which ACT-R makes predictions. The results are averaged over Section 1.7 (invert) and Section 2.6 (combine). The solid curve gives the predictions of the ACT-R model. The dotted lines display the data and the standard errors of the mean for each data point. Part (a) shows the fits corresponding to the fusiform and motor (see also Figure 7), part (b) shows the fits corresponding to the LIPFC and ACC (see also Figure 8), and part (c) shows the fits corresponding to the PPC and caudate (see also Figure 9).
Figure 1

(a) \[ ((3 \times x) + (4 \times x)) + 8 = 22 \]

(b) \[ \text{[Diagram]} \]
Figure 2

(a)  
\[ 5 \times \quad + \quad 4 \quad \rightarrow \quad 39 \]

(b)  
\[ 2 \times \quad 5 \times \quad - \quad + \quad 13 \quad 9 \times \quad + \quad 67 \]
Figure 3

(a) Sect. 1-7
Invert

\[ 5 \times 90 \]

(b)

\[ 90 \div 5 \]

\[ 18 \]

(c)

(d) Sect. 2-6
Combine

\[ 5 + (-3) \]

(e)

\[ 5 - 3 \]

(f)

\[ + 2 \]
Figure 4

(a) Sect. 1-7 Invert

(b) Sect. 2-6 Combine
Figure 5

(a) Day 1 Invert

(b) Day 5 Invert
Figure 6

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Visual</td>
<td>-42</td>
<td>-61</td>
<td>-9</td>
</tr>
<tr>
<td>2. Manual</td>
<td>-41</td>
<td>-20</td>
<td>50</td>
</tr>
<tr>
<td>3. Imaginal</td>
<td>-23</td>
<td>-63</td>
<td>40</td>
</tr>
<tr>
<td>4. Declarative</td>
<td>-43</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>5. Goal</td>
<td>-7</td>
<td>10</td>
<td>39</td>
</tr>
<tr>
<td>6. Procedural</td>
<td>-14</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 7

(a) Section 1-7, Invert

(b) Section 2.6, Combine
Figure 8

(a) Section 1-7,
Invert

(b) Section 2.6,
Combine
Figure 9

(a) Section 1.7,
Invert

(b) Section 2.6,
Combine
Figure 10
Figure 11

(a) Section 1.7, Invert

(b) Section 2.6, Combine
Figure 12

1. Problem
2. Transform
3. Fill-in
4. Problem
5. Evaluate
6. Fill-in
7. End 1back
8. Done
Figure 13

(a)

(b)
Figure 13 continued

(c)