Running head: TRAVELING THE SECOND BRIDGE

Traveling the Second Bridge: Using fMRI to Assess an ACT-R Model of Geometry Proof

Yvonne S. Kao, Scott A. Douglass, Jon M. Fincham, and John R. Anderson
Carnegie Mellon University

Abstract

We build an ACT-R model of a geometry-proof task based on behavioral, verbal protocol, eye movement, from proficient adult participants and then assess the model using neuroimaging data. Our geometry proof problems differed in terms of the number of steps of inference required to find a proof. Participants' verbal protocols and eye movements indicated that they first encoded the goal statement, then made a number of inferences, and finally came to a conclusion and made their response. In the imaging study we determined the involvement of a number of brain regions in the various stages of the task. The ACT-R model provided reasonable fits to participant accuracy, latency, and the BOLD response in most of the brain regions. However, there was unexpected activity in the motor region early in the problem solution, unexpected activity in the fusiform late in problem solution, and unexpected anterior prefrontal activity after the problem solution. We conclude that the model underestimates the perceptual-motor involvement in geometry learning just as does much of the educational curriculum.

John Bruer (1997) argued that applying neuroscience directly to educational questions was "a bridge too far." However, he also laid out a framework for connecting cognitive neuroscience to instruction via a sequence of two bridges that meet at the field of cognitive science. The first bridge links instruction to cognitive science; the second links cognitive science to cognitive neuroscience. In this paper, we will be traveling the second bridge in the domain of geometry proof. We build a cognitive model of a geometry proof task based on commonly-used behavioral and psychological measures, and then evaluate that model by examining its fit to neuroimaging data. Although we hesitate to make firm educational claims based on our model of this one task, the data do suggest certain processes that might be good targets for instruction.

A number of researchers have argued for the relevance of neuroscience approaches to educational questions in the effort to build this second bridge (e.g., Pettito & Dunbar, in press). Indeed, neuroimaging studies of number sense and arithmetic have made significant contributions to our understanding of the types of mental representations and cognitive processing used in mathematics and suggested targets for intervention (reviewed in Dehaene, Molko, Cohen, & Wilson, 2004; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999). Previous work conducted in our laboratory has combined fMRI with computational modeling to examine the cognitive processes underlying algebra equation-solving (e.g., Qin et al., 2004).

In the first study we will use verbal protocols and eye movements to provide converging evidence on what the major processes might be in a geometry proof problem-solving task and what their order might be. With this information about processes, we will be able to interpret the imaging data from the second, major study. Specifically, by inferring what brain regions are associated with these processes we will be better able to determine the information-processing character of these processes. This overall research effort illustrates how a convergence of

methodologies can lead to progress in inducing the processes behind a complex cognitive task. In particular, it illustrates how behavioral and neuroimaging methodologies can mutually inform each other in the goal of understanding complex mental processes. While our choice of domain is geometry proof problem solving, we think the same techniques could be applied to many domains.

A major factor determining difficulty of a geometry proof is the number steps of inference that are required to make the proof. Therefore, we contrasted problems that required 1 versus 3 steps of inference versus problems that were not provable. To have a simple response that could be used in the scanner we required participants simply to indicate whether the problem required 1 inference, 3 inferences, or if it was not provable. Fig. 1 shows examples of the problems used in these studies. For each problem, the goal statement was listed at the bottom of the diagram and the givens, which were always pairs of congruent segments, congruent angles, or parallel lines, were marked in red directly on the diagram in order to minimize reading requirements. In all cases, the goal was to prove a pair of line segments, angles, or triangles congruent. In half of the problems this goal pair was highlighted on the screen, but this had weak and inconsistent effects and we will report data averaged over this factor.

We performed an initial behavioral study to help us sketch out a model for the performance of the task. Then we performed a larger neuroimaging study to test predictions of that model. While many of the predictions of the model were confirmed in the neuroimaging study, there were surprises. These confirm the value of neuroimaging in coming to a deeper understanding of complex problem solving.

Initial Behavioral Study

Materials and Methods

Participants. Seven members of the Pittsburgh community (5 females) aged 19 to 32 years old (M = 25.29, SD = 4.15) participated in the study. Participants were recruited specifically for above-average ability performing geometry proofs; all had taken a proof-based geometry course in high school.

Procedure. Participants first worked, at their own pace, through a basic geometry tutorial. This tutorial was designed to review basic concepts and theorems from plane geometry as well as familiarize participants with the format of the problems they would be solving during the task. Participants reviewed 13 theorems and properties related to the following concepts: reflexivity, vertical angles, parallel lines, isosceles triangles, and triangle congruence. Participants were asked to use only these theorems in solving the problems and not to perform constructions or apply more advanced knowledge. These theorems are listed in the Appendix.

Next, participants completed 15 practice problems. Proofs were presented on a computer screen and participants entered their responses using the keyboard. Correct/incorrect feedback followed each problem. The first three of these problems had no time limit and were followed by correct/incorrect feedback and complete solutions. The remaining 12 problems were timed and were followed by correct/incorrect feedback only. Participants were allowed a maximum of 30 seconds on each timed problem.

Eye movements and verbal protocol data were collected separately; the order of the tasks was counterbalanced across participants. Prior to the verbal protocol task, participants were instructed to perform a talk-aloud protocol (Ericsson & Simon, 1993) and given a warm-up task. Then participants solved 24 problems while their verbalizations were recorded. An audio cue was inserted at the beginning of each trial to facilitate transcription of the recording. Participants

solved 36 problems during the eye-tracking task while their left eye movements were recorded. Systematic error in the eye movement data was corrected using a 6-point affine transformation (Douglass, in preparation).

Results

An alpha level of .05 was used for all statistical tests. Except for the analyses of accuracy, all analyses were performed on correct trials only.

A multivariate repeated-measures ANOVA found no effect of task (verbal protocol vs. eye movement) on accuracy, $F(1, 1) = 1.35$, $p = .29$, MSE = .039. We did find an effect of task on latency, $F(1,1) = 5.98$, $p = .05$, MSE = 22.157. Participants were, on average, 2.51 seconds faster on the eyetracking task (M = 11.60s, SD = 5.57s) than the verbal protocol task (M = 14.11s, SD = 6.17s), most likely because the act of verbalizing slowed participants. There were no significant interactions between task and difficulty and/or highlight, so the remaining behavioral results are collapsed across task.

We did find main effects of difficulty on proportion correct, $F(2,12) = 16.83$, $p < .0005$, MSE = .03, and latency, $F(2,12) = 29.21$, $p < .0005$, MSE = 25.105. Paired, 2-tailed t-tests found that participants correctly answered a higher proportion of 1-inference problems (M = 0.94, SD = 0.06) than 3-inference (M = 0.71, SD = 0.05), $t(6) = 7.62$, $p < .0005$, and not-provable problems (M = 0.75, SD = 0.14), $t(6) = 4.06$, $p = .007$. Participants were also significantly faster on 1-inference problems (M = 7.62s, SD = 0.94s) than on 3-inference problems (M = 13.10s, SD = 2.92s), $t(6) = -6.05$, $p = .001$, and were significantly faster on 3-inference problems than on not-provable problems (M = 17.84s, SD = 4.71s), $t(6) = -3.71$, $p = .10$).

Verbal Protocol Data. The protocols were segmented into statements and these statements were coded as specifying the goal (the statement to be proven), the givens of the

problem, an inference, the conclusion (1, 3, or not-provable), or other extraneous statements. Table 1 reports some statistics from the protocols. Almost 90% of the statements were classified as goal, givens, inference, or conclusion. The number of statements per problem increased by about two statements from 1-inference to 3-inference problems and by another one statement from 3-inference to not-provable problems. Most of this increase was due to an increased number of statements of the givens and the inferences.

These statements tended to be given in a definite order. For purposes of analyzing order of mention, we aggregated adjacent statements of the same type into a single occurrence of that type. Of the 110 instances of the goal under this classification, 104 occurred at the beginning of a problem. There were 116 instances of the conclusion, all of which occurred last. There were 77 instances of givens and 108 instances of inferences after this aggregation. Participants exhibited a weak tendency to articulate the givens before the inferences. The mean serial position of first mention of a given was 2.03 and the mean serial position of an inference was 2.33. There were 51 problems in which both a given and an inference were stated. Of these, 37 problems contained a mention of a given prior to any mentions of inferences. The number of goal statements and conclusions did not vary much with difficulty, only increasing from 2.06 per problem for 1-inference problems to 2.48 for not-provable problems. In contrast, there was a sharp rise in number of givens and inferences as these rose from 0.84 for 1-inference problems to 3.07 for not-provable problems.

Eyetracking Data. Each problem diagram was divided into three regions of interest: the area around the goal statement, the area around each marked given, and the remainder of the diagram. For each region (goal, given, diagram) we calculated whether there was an increase in the mean total amount of time spent fixating the region as a function of difficulty. To determine

this we tested the significance of the linear trend going from 1-inference problems to 3-inference problems, to not-provable problems. The linear trend was significant for the given regions (3.24 seconds difference between not-provable and 1-inference; $t(6) = 2.56$, $p < .05$, two-tailed) and diagram (3.90 sec; $t(6) = 3.19$, $p < .005$) but not for the proof region (0.39 sec; $t(6) = 1.52$, $p > .10$). This is consistent with the results from the verbal protocols that had indicated an increase in time spent on the givens and inferences.

Fig. 2 shows the proportion of time participants were fixating in each region during each of the first seven 2-second time intervals of the experiment. Participants' fixations on the goal statement declined steadily over the course of a trial, while the proportion of time they spent fixating on givens and the rest of the diagram remained high. Individual two-tailed t-tests revealed a significant decrease in the fixations in the goal region, $t(6) = -7.07$, $p < .0005$, and a significant increase in the fixations in the given region, $t(6) = 2.55$, $p = .044$, but no significant change in the fixations in the diagram region, $t(6) = 1.67$, $p = .146$. The negative slope of the goal region (M = -.07, SD = .03), differed significantly from the positive slopes of the given region (M = .03, SD = .04) and of the diagram region (M = .02, SD = .04). These eye movement data are consistent with the protocol data in indicating that participants encode the statement to be proven before the givens and before making inferences.

Discussion: ACT-R Model

It is a somewhat surprising outcome that participants seem to so uniformly attend to the proof statement first. Koedinger and Anderson (1990) found in their study of geometry experts that the givens and diagram are attended to first. In any case, for our task the data strongly suggest the following sequence of stages:

(a) **Goal Stage:** The first step involved encoding of the statement to be proven consistent with the strong tendency to say the goal first. Neither the protocol nor the eye-movement data suggested that this stage changed in length with problem difficulty.

(b) **Inference Stage:** Once the goal statement was encoded, participants engaged in a search for a set of inferences that would prove it. The duration of this stage increased with problem difficulty. This was apparent in the increased number of inferences mentioned in the verbal protocols and the increased fixation on other parts of the diagram. It is reasonable to assume that participants were making approximately 1 and 3 steps of inference for problems that required 1 or 3 steps of inference to make a proof. A number of participants reported that they determined a problem could not be proven by making more than 3 inferences and then giving up. The fact that the difference in latency between not-provable and 3-inference problems (4.74 sec) is approximately equal to the difference between 3-inference and 1-inference problems (5.49 sec) suggests that participants were averaging 2 extra inferences for each increase in difficulty. We will see further evidence for the assumption of 1, 3 and 5 inferences in the latency data for the next study.

(c) **Decision Stage:** There was a final step of making a decision and generating a response. This was consistent with the overwhelming tendency to state the conclusion concurrently or after the response.

Based on this analysis we developed an ACT-R model of the geometry-proof task. This is a running model that was capable of solving problems and which gave us time estimates that will be critical to understanding the imaging data. ACT-R (Anderson, 2007) proposes that cognition emerges as an interaction between various perceptual, central, and motor modules. Although the

nature of the task is substantially different from Koedinger and Anderson's (1990) geometry proof task, our model does resemble their geometry proof model in one significant way: both models use diagram configurations to guide retrieval of appropriate problem-solving schemas. A schema is a representation of a geometric pattern from which inferences can be made. The model works with three basic schemas: congruent triangles and their corresponding parts, parallel lines with alternative interior angles and corresponding angles, and isosceles triangles with congruent sides and base angles. The basic solution strategy is to retrieve a schema consistent with the current goal, try to instantiate its pattern for the current diagram, and then to try to retrieve inference rules required to instantiate it.

Fig. 3 shows the trace of this model solving the 3-inference problem illustrated in that figure in terms of the activity of six modules. The four that take the longest times are indicated by boxes – there is a visual module encoding both text-based and diagrammatic information from the screen, a retrieval module recalling geometric knowledge, an imaginal module keeping track of the state of the developing proof, and a manual module for programming hand movements. In addition, the horizontal lines in Fig. 3 reflect the firing of productions that are responsible for selecting cognitive actions and the brackets reflect periods of time when the model is operating under a single goal.

At the beginning of trying to solve this problem the model briefly considers the isosceles-triangles schemas as a basis for proving sides congruent but recognizes that the sides do not fit the isosceles-triangle configuration, and then considers schema involving corresponding parts of the congruent triangles (at about 2 seconds). Once it has identified the corresponding triangles, it goes through the 5 of the 6 pairs of corresponding sides and angles of the triangles trying to find pairs that are congruent until it completes an angle-angle-side configuration. It identifies the

reflexive side between 5 and 6 seconds, one of the given angles about 7 seconds ("Update Proof Angle" in imaginal column), and the other given angle between 8 and 9 seconds ("Update Proof Angle" again). Between 9 and 10 seconds it retrieves the angle-angle-side theorem and concludes it has a proof. It has been keeping a count of the number of inferences and on this basis can respond 3 at about 11 seconds. The model decides a problem cannot be proven if it reaches 5 inferences without a proof.

With respect to the module activity in Fig. 3:

1. The visual module is engaged in two principal activities. The larger boxes in Fig. 3 reflect inspection of the general diagram configurations and the smaller boxes reflect the more focused activities of finding parts in the diagram and checking the givens.
2. The retrieval module is engaged in retrieving different rules for geometry inference and retrieving the completed proof at the end.
3. The imaginal module keeps track of the current state of the proof, noting what rules of inference have been tried, keeping track of the number of inferences, and the various givens used.
4. The manual module is simply called at the end to execute the final response.
5. The procedural module is activated whenever a production rule is selected.
6. The control state in the goal module is reset to control each subtask in the performance of the overall task.

One major characteristic of this model, which is captured in Fig. 3, is that the visual engagement is strongest early as the various relevant parts of the diagram are identified, all the central modules (retrieval, imaginal, procedural, and goal) are activated rather constantly, and the motor module is activated towards the end. This might seem to be the logical progression that any

model for this task would predict. However, we will see that this does not entirely match up with the results of the fMRI study that we will report. We will describe how well this model corresponded to the various data from the task after describing the main imaging study.

Using Imaging Data to Infer Stage Content

The main experiment is an fMRI study of participants performing the same geometry-proof task as in the initial study. In performing this experiment, we are applying an approach previously used to study problem-solving to a new domain. We will briefly describe these earlier studies here.

In a series of simpler tasks (e.g., Anderson, 2005; Anderson, Albert, & Fincham, 2005; Anderson, Qin, Jung, & Carter, 2007; Anderson, Qin, Stenger, & Carter, 2004; Qin et al., 2003; Sohn, Goode, Stenger, Carter, & Anderson, 2003; Sohn, Goode, Stenger, Jung, Carter, & Anderson, 2005), our laboratory has mapped the 6 modules in Fig. 5 onto specific brain regions. We have been able to leverage the existence of relatively well-developed cognitive models for these tasks to identify the functions of these regions. These 6 module/region correspondences are:

1. The manual module maps onto a motor region centered at x = -37, y = -25, z = 47. This includes parts of Brodmann Areas 2 and 4 at the central sulcus.
2. The goal module maps onto an anterior cingulate region centered at x = -5, y = 10, z = 38. This includes parts of Brodmann Areas 24 and 32.
3. The imaginal module mapped onto a parietal region centered at x = - 23, y = -64, z = 34. This includes parts of Brodmann Areas 7, 39, and 40 at the border of the intraparietal sulcus.

4. The retrieval module mapped onto a prefrontal region centered at x = -40, y = 21, z = 21. This includes parts of Brodmann Areas 9, 45, and 46 around the inferior frontal sulcus.
5. The procedural module mapped onto a caudate region centered at Talairach coordinates x = -15, y = 9, z = 2. This is a subcortical structure. This region tracked the amount of procedural effort in performing the task.
6. The visual module mapped onto a fusiform region centered at Talairach coordinates x = -42, y = -60, z = -8. This includes parts of Brodmann Area 37.

The first five of these regions and their right-hemisphere homologues are depicted in Fig. 4, along with the exploratory regions from the main experiment that we will discuss later. These predefined regions are the same regions that we have used in our previous studies.

The functional associations listed above are roughly consistent with the research in other laboratories. Others have associated the parietal region with mental imagery (Dehaene, Piazza, Pinel, & Cohen, 2003; Reichle, Carpenter, & Just, 2000), the prefrontal region with retrieval (Buckner, Kelley & Petersen, 1999; Cabeza, Dolcos, Graham, & Nyberg, 2002; Fletcher & Henson, 2001; Lepage, Ghaffar, Nyberg, & Tulving, 2000; Wagner, Maril, Bjork, & Schacter, 2001; Wagner, Pare-Blagoev, Clark, & Poldrack, 2001), the caudate with procedural execution (Poldrack, Prabakharan, Seger, & Gabrieli, 1999; others), the anterior cingulate with control functions (Carter et al., 2000; D'Esposito, Detre, Alsop, Shin, Atlas, & Grossman, 1995; Posner & Dehaene, 1994), and the motor with manual control (e. g., Roland, Larsen, Lassen, & Skinhoj, 1980).

Another interesting observation is that the responses in all these predefined regions are stronger in the left hemisphere for algebra and many non-algebra tasks. Given that participants are responding with their right hand, the lateralization of the motor area is to be expected. With

respect to the other regions, we have often found similar responses in the right homologues of these regions but the response was always weaker. Given that the tasks studied typically involved symbolic material, perhaps this asymmetry was to be expected (e.g., Gabrieli, 2001, Robertson & Rafal, 2000). It will be interesting to see if the current geometry task, which has such a large visual component, will continue to produce left-lateralized responses.

Main Study

Method

Participants. Sixteen members of the Pittsburgh community (7 females) aged 18 to 33 years old (M = 24.0, SD = 4.7), including six of the participants from the initial study, completed the study. Again, participants were recruited specifically for above-average ability performing geometry proofs; all had taken a proof-based geometry course in high school. All participants were right-handed.

Procedure. The study spanned two sessions: a training/pre-test session and an fMRI scanning session. The training session was an extended version of the review and practice used in the initial study. Participants worked through the same basic geometry tutorial and reviewed the same 13 theorems and properties. Next, participants were trained on one of four different problem sets. The training set began with 12 practice problems with no time limit. After each problem, participants saw correct/incorrect feedback and a correct solution for the problem. The training sequence ended with 36 timed practice problems in three blocks of 12 problems each. Again, participants had a maximum of 30 seconds to enter a response for each timed problem. After each timed problem, correct/incorrect feedback was displayed on the screen for one

second, but the feedback was not accompanied by solutions. Participants who did not achieve at least 75% accuracy in the last of the timed practice blocks were removed from the study.

fMRI Acquisition. During the scanning session, participants solved a total of 120 timed problems divided into ten blocks. Participants who did not achieve at least 70% accuracy overall on this task were excluded from the analysis. Each block began with a 16-second delay with fixation. Again, participants saw correct/incorrect feedback for one second following each problem. The inter-trial interval was 16 seconds long to allow time for the BOLD response to return to baseline (Aguirre et al., 1998), during which only a fixation cross was present on the screen.

Event-related functional images were acquired using gradient echo-planar image (EPI) acquisition on a Siemens 3-T Allegra Scanner using a standard RF head coil. Imaging parameters were TR = 2000 msec, TE = 30 msec, RF flip angle = 70 deg, FOV = 200 mm, matrix size = 64 x 64 (3.125 x 3.125 mm in-plane resolution per voxel), slice thickness = 3.2 mm, slice gap = 0 mm. Acquisition included 34 oblique axial slices parallel to anterior commissure-posterior commissure (AC-PC) line per volume scan with the AC-PC at slice 24 from the superior. Anatomical scans (34 slices) were acquired using a standard T2-weighted spin-echo pulse sequence, with the middle of slice 24 from the superior through the AC-PC line.

fMRI Analysis. Acquired images were analyzed using the NIS system. All images were co-registered to a common reference structural MRI scan by means of a 12-parameter automatic image registration algorithm (AIR) and smoothed with an 8mm full-width-half-max 3-D Gaussian filter to accommodate individual differences in anatomy. Spatial *F* maps were generated using an analysis of variance (ANOVA).

Because of the high variability in response times for the task, BOLD responses were warped to the median response times for the corresponding level of difficulty. First, for each trial and each region of interest we calculated the percent change of the BOLD (Blood Oxygen Level Dependent) response on each scan with reference to the first scan on a trial. We then broke these change scores into two intervals. One interval was from the first scan until the scan of the response and the other from the first scan after the response until the last scan of the trial. The second period was constant across trials and did not need warping. We then warped the first intervals onto the median interval according to the following procedure for taking a scan sequence of length *n* and deriving a scan sequence of the mean length *m*. It depends on the relative sizes of *m* and *n* such that:

1. If *n* is greater than or equal to *m*, create a sequence of length *m* by taking *m/2* scans from the beginning and *m/2* from the end. If *m* is odd select one more from the beginning. This means just deleting the *n-m* scans in the middle.
2. If *n* is less than *m*, create 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*. Construct the end similarly. If either *n* or *m* is odd, the extra scan is from the beginning.

This creates scan sequences that preserve the temporal structure of the beginning of the trial and around the response but tends to just represent the approximate average activity in the middle of the trial.

Results

One participant was excluded from the analysis for not meeting the 70% accuracy criteria during the scan session. This participant achieved 49% accuracy versus an average of 79% accuracy achieved by the other 15 participants. An alpha level of .05 was used for all statistical

tests.

Behavioral Results. A multivariate repeated-measures ANOVA found a significant effect of difficulty on proportion correct, $F(2, 28) = 14.59$, $p < .0005$, MSE = .011. Participants correctly answered a higher proportion of 1-inference problems (M = 0.87, SD = 0.06) than on either 3-inference problems (M = 0.74, SD = 0.07) or not-provable problems (M = 0.76, SD = 0.07). The majority of these errors were due to participants deciding that 3-inference problems were not provable and vice versa. These data, along with the accuracy and error data for the model we will discuss later, are presented in Fig. 5a-b.

We consider correct trials only for all latency analyses. A multivariate repeated-measures ANOVA on correct latencies also found a significant effect of difficulty, $F(2, 28) = 181.73$, $p < .0005$, MSE = 3.00. Participants were faster on 1-inference problems (M = 6.34 sec., SD = 1.29) than on 3-inference problems (M = 10.93 sec., SD = 2.30), $t(14) = 13.09$, $p < .001$, which were faster than not-provable problems (M = 14.85 sec., SD = 2.87), $t(14) = 9.58$, $p < .001$. As in the initial study, the mean difference in latencies between not-provable problems and 3-inference problems (3.92 sec) is almost the same as the mean difference between 3-inference problems and 1-inference problems (4.59 sec). The mean latencies and the entire latency distribution, along with their model fits, are shown in Fig. 5c-d.

Modeling the Behavioral Results. Fig. 5 also shows the behavioral data and the predictions from the ACT-R model of the task that was illustrated in Fig. 3[1]. We completed 15 runs of the model through all 120 geometry problems and compiled the data. The model exhibits

---

[1] This model can be downloaded from the models link at the ACT-R website (act-r.psy.cmu.edu) under the title of this paper.

variability in the steps of inference it takes and the mean time it takes to perform various actions. If the model took more than 30 seconds to enter a response, it was considered a "No Response" error, as our participants faced a 30-second time limit. Our analysis of the behavioral data involves comparing the 15 model runs against the 15 subjects, essentially treating the source of the data (model versus human) as a factor in the design.

As Fig. 5a-d shows, the model fits the accuracy and latency data from our fifteen participants reasonably well. Chi-square tests of deviation found that the average model is more accurate than the average participant on 1-inference, $t(28) = 2.79$, $p < .01$, and 3-inference problems, $t(28) = 3.43$, $p < .005$, but the accuracies are still quite similar (Fig. 5a). The model's error distribution (Fig. 5b) captures the overall pattern of participants' errors, with a correlation between the two ($r = 0.916$). Still there are some significant differences. The model makes a significantly smaller proportion of errors calling 1-inference problems 3-inference problems ($t(28) = 2.77$), and calling 3-inference problems 1-inference ($t(28) = 3.08$) problems. Correspondingly, it more often fails to respond to 1-inference problems ($t(28) = 2.20$) and 3 inference problems ($t(28) = 4.15$). The model seems to lose track of where it is in a proof much more frequently than the participants, which often leads the model to not attempt a response at all, rather than merely taking too long to respond.

The model fits the mean participant latencies for each difficulty condition with no significant deviations (Fig. 5c). Fig. 5d presents the distribution of responses at various latencies. There is a good match up here ($r = .909$) but there are points of discrepancy. In particular, we find significant deviations from the data for 1-inference problems ($\chi^2(14) = 119.83$, $p < .001$) but not for 3-inference or not-provable problems. The most significant of these deviations occur between 4 and 10 seconds from problem onset.

As a further test of the model's behavioral outcomes, we analyzed the 3-inference problems in terms of the number of schemas the participant needed to retrieve in order to solve each problem correctly. Of the 3-inference problems in the test set, 65% were problems that could be solved using a single schema, e.g., corresponding parts of congruent triangles. The remaining 35% of the 3-inference problems required two schemas to solve, e.g., first using a parallel lines schema to prove a pair of angles congruent and then using the corresponding parts of congruent triangles schema to finish the proof. These fits are shown in Fig. 5e-h.

As before, the average model is more accurate than the average participant for both 1-schema ($t(28) = 2.90$, $p < .01$) and 2-schema ($t(28) = 3.24$, $p < .005$) problems (Fig. 5e). The model mirrors the error distribution for 1-schema problems, but differs significantly for 2-schema problems (Fig. 5f). While participants often decide that 2-schema, 3-inference problems can be solved in one inference, the model never does this. And while participants never fail to make a response, the model often does this. This is similar to what we see when we examine the error distribution across all problem types.

The model latencies for 1-schema and 2-schema problems still capture the major patterns in the data, but the fits are not as good as when we considered all 3-inference problems together. The model fits the mean latency (Fig. 5g) for 1-schema problems well, but is a bit slower than the participants for 2-schema problems ($t(28) = 1,86$, $p < .10$). The latency distributions in Fig. 5h closely fit the data in most places ($r = .897$), but does show some significant deviations for both 1-schema ($\chi^2(14) = 32.59$, $p = .003$) and 2-schema problems ($\chi^2(14) = 40.08$, $p < .001$). These deviations occur between 10 and 12 seconds for the 1-schema problems and at 16 and 24 seconds for 2-schema problems.

By way of summary, while the model does not fit the behavioral data perfectly, it does match up pretty well. The most notable points of deviation may reflect both the fact that its geometry knowledge is rather limited and that it deploys the same knowledge from run to run. Our participants were undoubtedly more knowledgeable and could have solved problems beyond the capacity of the model. The actual knowledge they had also probably varied from participant to participant. Nonetheless, we judged the behavioral data that we have obtained as basically supporting the ACT-R instantiation of the 3-stage model that we developed from our pilot studies.

Modeling the Imaging Results. In order to use this model to predict the BOLD response in each of our predefined regions of interest, we first determined a demand function *d*(x) for each region, which has a value of one when the module associated with that region is active and a value of zero when it is inactive. In terms of Fig. 3, the demand function would be one for the time length of the boxes.[2]

Whenever there is demand for a module we assume that it will drive a hemodynamic response described by *b*(t), which is a standard gamma function used in previous studies to represent the hemodynamic response (Boyton, Engel, Glover, & Heeger, 1996; Cohen, 1997; Dale & Buckner, 1997; Glover, 1999):

$$b(t) = m\left(\frac{t}{s}\right)^{a} e^{-(t/s)} .$$

---

[2] The procedural component is active for 50 msec whenever a production fires (line in Fig. 3). The goal component is only active at the transitions between goals and for this we use a point version of the demand function that is active just at these points of time.

In this function, *m* is the magnitude of the response, *s* is a time scaling parameter, and *a* determines the steepness of the BOLD response—the greater the value of *a*, the steeper the function. We used values of $a = 5$ and $s = 1$ second for all regions and just estimated different magnitude parameters on a per region basis.

We then convolved functions *d*(x) and *b*(t) to produce the complete BOLD response function:

$$B(t) = \int_0^t d(x)b(t-x)dx .$$

We fit the BOLD responses in Fig. 6 by estimating values for the magnitude parameters for each region. We estimated the significance of the deviations by calculating the following chi-square statistic:

$$\chi^2(df) = \frac{\sum_i (\hat{Y}_i - \overline{Y}_i)^2}{s^2_{\overline{Y}}}$$

in which the summation in the numerator is of the deviations between predicted and mean BOLD response and the denominator is the variance of the means determined by the participant-by-condition interaction. Since the first and last data points in each condition are constrained to be 0 and there are 11, 13, and 16 data points for each of the 3 conditions respectively, there are 33 degrees of freedom. Subtracting 1 for the estimation of the magnitude parameter means that our chi-square measures have 32 degrees of freedom

Imaging Data: Predefined Regions. Fig. 6a-f presents the data for the six left predefined regions that have been used in our past research and their fits to the model described above.

Table 2 presents results for difficulty and difficulty × scan interactions for all six predefined regions of interest in both hemispheres. The effects are quite comparable in the two hemispheres.

All predefined regions showed significant effects of the difficulty variable. It is noteworthy that the motor region shows an effect of problem difficulty. This is quite unlike the algebra results (e.g., Anderson, 2005) in which the only effect of problem difficulty was to delay the BOLD response in the motor area. In this experiment we are seeing a significantly more sustained BOLD response in the motor area in the more difficult conditions. A number of participants reported that they were aware of covert finger movements as they pointed to various parts of the diagram. We assume that the effect of difficulty reflects these covert movements that would occur more often in the more difficult conditions. Table 3 summarizes the model's fit to each of the predefined regions. It gives the magnitude scaling parameter that minimized the sum of squared error between model and data, the chi-square measure of deviation, and the correlation measure of correspondence. Table 3 summarizes the fits to the right predefined regions as well. The same models fit the left and right regions with one exception: the right motor region is modeled as having no activity because participants respond only with their right hand. With respect to the left regions, for which the model's predictions are most relevant, the correlations were generally greater than 0.9. Some of the deviations for the 0.9 correlations were significant but we judge these as relatively minor. The left motor and left fusiform are the two regions that gave much lower correlations and for which the fits are qualitatively at variance with the data. They clearly indicate something is happening that is not captured by our model; we will comment on these more in the general discussion.

Imaging Data: Exploratory Regions. An exploratory analysis found eight regions of 50 or more contiguous voxels that were significant at $p < .001$ for the difficulty × scan interaction. One

of these regions was a large area of 5881 voxels that contained our predefined parietal, left motor, and left prefrontal regions, the left homologue to the right prefrontal exploratory region (a) in Fig. 4, and a large portion of the visual cortex. Because this region overlaps so many of our predefined regions, we will not comment on this region further. The remaining seven exploratory regions are pictured in Fig. 4. Details regarding the size, location, and the magnitude of the response to each condition are presented in Table 4. The magnitude of the response is defined as the sum of the scans from problem onset.

Two of the exploratory regions correspond well with our predefined regions. The right frontal exploratory region (region a in Fig. 4) is superior, but includes much of our predefined prefrontal region. The anterior cingulate exploratory region (region b in Fig. 4) occupies much of the same area as our predefined left and right anterior cingulate regions. Thus it is no surprise that the patterns of activity in these exploratory regions are similar to the activity in their predefined counterparts.

We found two pairs of exploratory regions that appeared to be particularly noteworthy and did not overlap with any of our predefined regions. First there were two very anterior prefrontal regions found by the exploratory analysis (regions c and d in Fig. 4). These have patterns of activity (Figs. 7a and 7b) that resemble the activity in the other frontal regions, but peak approximately four seconds after the mean response time for each difficulty condition. This would indicate that these regions are heavily involved in post-processing of the task and may reflect various metacognitive activities. The association of this region with metacognitive activity is consistent with the gateway hypothesis (Burgess, Gilbert, Okuda, & Simons, 2006; Gilbert, Spengler, Simons, Frith, & Burgess, in press) that sees the medial anterior prefrontal region as serving to direct processing in a stimulus independent manner.

We also found two insula regions (regions e and f in Fig. 4). The insula has been implicated in a wide variety of functions, including processing and integration of autonomic and visceral information, olfactory and gustatory functions, affect, and motivation. Of particular relevance to the current study, the caudal insula appears to be necessary for spatial learning (Flynn, Benson, & Ardila, 1999).

Discussion

We would like to begin with a discussion of the model fits. In most cases the model gave at least a good approximation to the behavioral and imaging data. A number of the tests of deviation were significant, but the patterns were quite similar as indicated by measures of correlation. Most of the residual differences probably reflected the fact that the knowledge embedded in the model was only an approximation to the knowledge that participants had.

The two major exceptions to our positive assessment were the results in the motor area and the fusiform region, both of which were quite unexpected. Fig. 8 provides a contrast between the results in the motor, prefrontal, and fusiform regions and the model predictions. The figure provides the comparison for the not-provable problems, which took the longest and offers the clearest contrast. This graph has been scaled so the maximum of each function is 1. As can be seen, the predictions are that activation would be maximal first in the fusiform, reflecting the early visual processing, then in the prefrontal, reflecting the relatively uniform distribution of retrieval, and last in the motor region, reflecting the programming of the response at the end. While the prefrontal went as expected, the results are almost reversed for the motor and fusiform regions. There is early, unpredicted activity in the motor region. The activation in the fusiform reaches maximum at the very end, substantially after the model predicts the major visual engagement.

In the motor region, the model appears to fit the final rise in activity that corresponds to the actual button press, but the heavy motor involvement from problem onset remains a bit of a puzzle. Interestingly, the right motor shows equal early engagement but not the final rise. Although we have suggested that this might be due to participants' covert finger movements during the task, it is unclear to us whether many covert finger movements would produce a level of task involvement roughly equivalent to a single overt button press. Perhaps the motor activity is also being driven by mental rotation, as some studies have found motor cortex activation during mental rotation tasks (e.g., Richter et al., 2000; Tomasino, Borroni, Isaja, & Rumiati, 2005; Windischberger, Lamm, Bauer, & Moser, 2003). However, these studies have also found that primary motor cortex is primarily responsible for the button-press response, with premotor and supplementary motor areas involved in the actual mental rotation (Richter et al., 2000; Windischberger, et al., 2003), unless the mental rotation is of hands (Tomasino et al., 2005).

In the fusiform region, the model approximately captures the initial rise in the response, but predicts that activity would drop off much more quickly than it actually does. We believe that this is because the model does not adequately capture the visual-processing demands of the task. During the tasks, participants are continually fixating on different parts of the diagram as long as it is present on the screen, often returning to fixate on critical parts of the diagram or the goal statement multiple times. The model, however, only looks at the goal statement once, and then does not display the same visual scanning behavior—the visual module goes idle during much of the later portion of the task. We should note that in some recent studies, where we have looked at participants solving longer algebra and arithmetic problems (Anderson et al., submitted), we have found a decrease in fusiform activation after an initial heavy engagement as

the ACT-R models for those tasks would imply. Therefore, the growing fusiform activation in this task may reflect the unique visual characteristics of geometry.

There were a couple of other features of this geometry task that are different from results we have obtained in algebra and arithmetic. First, as can be seen from Table 3, the right prefrontal gave as good a fit to the predictions from the retrieval module as the left prefrontal and the magnitude of the response is greater. This is the only study of mathematical problem solving we have done that has found such strong involvement in the right prefrontal. Usually, there is much weaker activation in the right prefrontal region. Perhaps this reflects the visual nature of geometry.

Second, this is the only experiment we have run where we have found significant activation in the anterior prefrontal area. This is also the only study to find a positively responding area that increases its activation after the task completes. Perhaps the reasoning components of geometry evoke more meta-cognitive activity than the more algorithmic activity of arithmetic and algebra.

The imaging data thus lead us to draw the following conclusions about our model and about the cognitive processing in our geometry proof task:

1. There is greater motor and sensory involvement in the task. It seems that tasks that could in principle be done by central cognition alone are recruiting the visual and manual systems for support.
2. Second, there is greater involvement of right hemisphere regions particularly in the retrieval of geometric knowledge.
3. There is greater evidence for metacognitive involvement in a task where the steps are not very routine. Although this in itself is not surprising, it is a bit of a mystery

to us why this metacognitive involvement was greatest after the completion of the task.

General Discussion

There seem to be two major messages about geometry proof problem-solving that emerge from our model and its fit to the neuroimaging data, messages that suggest possible targets for geometry instruction. The first message is about the primacy of visual processing throughout the task. Although it is not surprising that geometry would require a significant amount of visual processing, it seems clear that our model, with its relatively straightforward visual encoding, has seriously underestimated the visual processing requirements. The model also did not anticipate that other areas, such motor areas, would be involved with the visuospatial demands. The recruitment of motor areas to help with a geometry task seems consistent with work by Alibali (2005), who has suggested that gesturing may assist in problem solving by helping to keep spatial representations active.

It appears that educators often make the same assumption that we did, that visually encoding and interpreting diagrammatic information is not cognitively taxing. Geometry instruction tends to focus on teaching formulas and theorems; there is very little explicit instruction on how to interpret diagrams. This is in spite of much prior work in geometry that suggests visualization skills support the acquisition of geometric reasoning skills (e.g., Bishop, 1983; Bishop, 1986), that spatial processing demands form the bulk of the cognitive load for students solving geometry problems (Lean & Clements, 1981), and that common errors can result from insufficient or inflexible processing of diagrammatic information (Hoz, 1981;

Koedinger & Cross, 2000). The amount of fusiform activity we found in our study certainly supports these ideas.

The second message of our study is the evidence for the use of a schema-based problem-solving strategy by our participants. This problem-solving strategy has implications for both latency and activity in the model's retrieval module, as the model will always retrieve a schema before any individual theorems. The model's predictions held up well to the scrutiny of a schema-based analysis of the 3-inference problems. In addition, the fit of the retrieval module predictions to the BOLD data from our predefined prefrontal region was quite good. This finding is also not surprising, as effective schema organization is a feature of proficient and expert problem-solving in all domains (e.g., Chi, Glaser, & Rees, 1981), including geometry (Koedinger & Anderson, 1990). What is interesting here is the way our model, as well as Koedinger and Anderson's model, uses particular diagram configurations to guide the selection of particular schemas while solving a problem. Thus it seems that not only is it important to teach visual processing skills to geometry students, but it is important to teach students about the significance of certain diagram configurations and to link them to problem-solving schemas.

Finally, we would like to conclude with a methodological point about the value of brain imaging. The model we came up with was based on a rich array of behavioral evidence – accuracy, error patterns, latency distributions, verbal protocols, and eye movements. While the match of the model to this array of data was not perfect, nothing suggested the results we would see in the motor region or the fusiform. It was only by going to imaging data that we became aware of these complications and the need to enrich our consideration of the problem solving.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M.D. (1998). The variability of human, BOLD hemodynamic responses. *NeuroImage, 8*, 360-369

Alibali, M.W. (2005). Gesture in spatial cognition: Expressing, communicating and thinking about spatial information. *Spatial Cognition & Computation, 5*, 307-331.

Anderson, J.R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science, 29*(3), 313-341.

Anderson, J.R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Anderson, J.R., Albert, M.V., & Fincham, J.M. (2005). Tracing problem solving in real time: fMRI analysis of the subject-paced Tower of Hanoi. *Journal of Cognitive Neuroscience, 17*, 1261-1274.

Anderson, J. R., Carter, C.S., Fincham, J.M., Qin, Y., Ravizza, S.M., & Rosenberg-Lee, M. (submitted). The image of complexity.

Anderson, J. R., Qin, Y., Jung, K.J., & Carter, C.S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology, 54,* 185-217.

Anderson, J.R., Qin, Y., Stenger, V.A., & Carter, C.S. (2004). The relationship of three cortical regions to an information-processing model. *Journal of Cognitive Neuroscience, 16*(4), 637-653.

Bishop, A.J. (1983). Space and geometry. In R. Lesh & M. Landau (Eds.), *Acquisition of Mathematics Concepts and Processes* (pp. 176-203). New York: Academic Press.

Bishop, A.J. (1986). What are some obstacles to learning geometry? *Studies in Mathematics Education, 5*, 141-159.

Boyton, G.M., Engel, S.A., Glover, G.H., & Heeger, D.J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience, 16*, 4207-4221.

Bruer, J.T. (1997). Education and the brain: A bridge too far. *Educational Researcher*, *26*(8), 4-16.

Buckner, R.L., Kelley, W.M., Petersen, S.E. (1999). Frontal cortex contributes to human memory formation. *Nature Neuroscience, 2*, 311-314.

Burgess, P.W., Gilbert, S.J., Okuda, J., & Simons, J.S. (2006). Rostral prefrontal brain regions (area 10): A gateway between inner thought and the external world? In Prinz,W. & Sebanz, N. (Eds.), *Disorders of Volition*. Cambridge (pp. 373-396), MA: MIT Press.

Cabeza, R., Dolcos, F., Graham, R., & Nyberg, L. (2002). Similarities and differences in the neural correlates of episodic memory retrieval and working memory. *NeuroImage, 16*(2), 317-330.

Carter, C.S., MacDonald, A.M., Botvinick, M., Ross, L.L., Stenger, V.A., Noll, D., & Cohen, J.D., (2000). Parsing executive processes: Strategic versus evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the U.S.A., 97,* 1944-1948.

Chi, M.T.H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 1, pp. 7-76). Hillsdale, NJ: Erlbaum.

Cohen, M.S. (1997). Parametric analysis of fMRI data using linear systems methods. *NeuroImage, 6*, 93-103.

Dale, A.M. & Buckner, R.L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping, 5*, 329-340.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A.J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology, 14*, 218-224.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*(3-6), 487-506.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284,* 970-974.

D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S., Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature, 378*, 279-281.

Douglass, S. (in preparation). Compensating for systematic error in eye movement data.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis* (revised ed.). Cambridge, Massachusetts: MIT Press.

Fangmeier, T., Knauff, M., Ruff, C. C., & Sloutsky, V. (2006). FMRI evidence for a three-stage model of deductive reasoning. *Journal of Cognitive Neuroscience, 18*(3), 320-334.

Fletcher, P.C. & Henson, R.N. (2001). Frontal lobes and human memory: Insights from functional neuroimaging. *Brain: A Journal of Neurology, 124*(5), 849-881.

Flynn, F.G., Benson, D.F., & Ardila, A. (1999). Anatomy of the insula – functional and clinical correlates. *Aphasiology, 13*(1), 55-78.

Gabrieli, J.D.E. (2001). Functional neuroimaging of episodic memory. In Cabeza, R., & Kingstone, A., (Eds.) *Handbook of Functional Neuroimaging of Cognition*. MIT Press: Cambridge, MA, 253-291.

Gilbert, S.J., Spengler, S., Simons, J.S., Frith, C.D., & Burgess, P.W. (in press). Differential functions of lateral and medial rostral prefrontal cortex (area 10) revealed by brain-behaviour correlations. *Cerebral Cortex*.

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

Hoz, R. (1981). The effects of rigidity on school geometry learning. *Educational Studies in Mathematics, 12*, 171-190.

Koedinger, K.R. & Anderson, J.R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*, 511-550.

Koedinger, K.R., & Cross, K. (2000). Making informed decisions in educational technology design: Toward meta-cognitive support in a cognitive tutor for geometry. In *Proceedings of the Annual Meeting of the American Educational Research Association (AERA)*, New Orleans, LA.

Lean, G. & Clements, M.A. (1981). Spatial ability, visual imagery, and mathematical performance. *Educational Studies in Mathematics, 12*, 267-299.

LePage, M., Ghaffar, O., Nybert, L., & Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Sciences of the U.S.A., 97*(1), 506-11.

Petitto, L.A. and Dunbar, K.N. (in press). New findings from educational neuroscience on bilingual brains, scientific brains, and the educated mind. In K. Fischer & T. Katzir (Eds.), *Building Usable Knowledge in Mind, Brain, & Education*. Cambridge University Press.

Poldrack, R.A., Prabakharan, V., Seger, C., & Gabrieli, J.D.E. (1999). Striatal activation during cognitive skill learning. *Neuropsychology, 13*, 564-574.

Posner, M.I. & Dehaene, S. (1994). Attentional networks. *Trends in Neurosciences, 17*(2), 75-79.

Qin, Y., Carter, C.S., Silk, E.M., Stenger, V.A., Fissell, K., Goode, A., & Anderson, J.R. (2004). The change of the brain activation patterns as children learn algebra equation solving. *Proceedings of the National Academy of Sciences of the U. S. A., 101*(15), 5686-5691.

Qin, Y., Sohn, M.H., Anderson, J.R., Stenger, V.A., Fissell, K., Goode, A. Carter, C.S. (2003). Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences of the U. S. A., 100*(8), 4951-4956.

Reichle, E.D., Carpenter, P.A., Just, M.A. (2000). The neural bases of strategy and skill in sentence-picture verification. *Cognitive Psychology, 40*(4), 261-295.

Richter, W., Somorjai, R., Summers, R., Jarmasz, M., Menon, R.S., Gati, J.S., Georgopoulos, A. P., Tegeler, C., Ugurbil, K., & Kim, S.G. (2000). Motor area activity during mental rotation studied by time-resolved single-trial fMRI. *Journal of Cognitive Neuroscience, 12*(2), 310-20.

Robertson, L.C. & Rafal, R. (2000). Disorders of visual attention. In M. Gazzaniga (Ed.), *The New Cognitive Neuroscience* (2nd ed.). Cambridge: MIT Press.

Roland, P.E., Larsen, B., Lassen, N.A., & Skinhoj, E. (1980). Supplementary motor area and other cortical areas inorganization of voluntary movements in man. *Journal of Neurophysiology, 43*, 118–136.

Sohn, M.H., Goode, A., Stenger, V.A., Carter, C.S., & Anderson, J.R. (2003). Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior

parietal cortex, *Proceedings of National Academy of Sciences, 100*, 7412-7417.

Sohn, M.H., Goode, A., Stenger, V.A., Jung, K.J., Carter, C.S., & Anderson, J.R. (2005). An information-processing model of three cortical regions: Evidence in episodic memory retrieval, *NeuroImage, 25*, 21-33.

Tomasino, B., Borroni, P., Isaja, A., & Rumiati, R. (2005). The role of primary motor cortex in mental rotation: A TMS study. *Cognitive Neuropsychology, 22*(3-4), 348-363.

Wagner, A.D., Maril, A., Bjork, R.A., & Schacter, D.L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral prefrontal cortex. *NeuroImage, 14*, 1337-1347.

Wagner, A.D., Pare-Blagoev, E.J., Clark, J., & Poldrack, R.A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31*(2), 329-338.

Windischberger, C., Lamm, C., Bauer, H., & Moser, E. (2003). Human motor cortex activity during mental rotation. *NeuroImage, 20*(1), 225-232.

Appendix

Geometry Theorems

Reflexivity
- Every line segment is congruent to itself.
- Every angle is congruent to itself.

Vertical Angles
- Vertical angles are congruent.

Parallel Lines
- If parallel lines are cut by a transversal, then corresponding angles are congruent.
- If two lines are cut by a transversal, and a pair of corresponding angles are congruent, then the lines are parallel.
- If parallel lines are cut by a transversal, then alternate interior angles are congruent.
- If two lines are cut by a transversal, and the alternate interior angles are congruent, the lines are parallel.

Triangle Congruence
- Side-side-side (SSS)
- Side-angle-side (SAS)
- Angle-side-angle (ASA)
- Angle-angle-side (AAS)
- Hypotenuse-leg for right triangles (HL)
- Corresponding parts of congruent triangles are congruent.

Isosceles Triangles
- If two sides of a triangle are congruent, then the angles opposite those sides are congruent.
- If two angles of a triangle are congruent, then the sides opposite those angles are congruent.

Acknowledgements

This research was supported by NSF ROLE grant REC-0087396 to Anderson.

Correspondence concerning this article should be addressed to Yvonne Kao, Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. Electronic mail may be sent to ykao@andrew.cmu.edu. Phone: (412) 268-8113. Fax: (412) 268-2798.

Table 1

Summary of Verbal Protocol Data

	Difficulty		
	1-Inference	3-Inferences	Not Provable
Number of Problems	49	39	42
Statements/Problem	3.41	5.54	6.36
Goal/Problem	1.10	1.33	1.55
Givens/Problem	0.41	0.97	1.07
Inferences/Problem	0.43	1.69	2.00
Conclusions/Problem	0.96	0.97	0.93
Other/Problem	0.47	0.56	0.79

Table 2

F-statistics for Within-Participants Difficulty Effects and Difficulty x Scan Interactions in Predefined Regions

	Hemisphere			
	Right		Left	
	Difficulty	Difficulty x Scan	Difficulty	Difficulty x Scan
Motor	$F(1.5,21.0) = 10.73^{**}$	$F(3.3,48.1) = 5.15^{**}$	$F(1.4,19.4) = 11.09^{**}$	$F(3.7,51.3) = 26.86^{***}$
Anterior Cingulate	$F(1.5,20.2) = 23.50^{***}$	$F(4.1,57.4) = 51.07^{***}$	$F(1.3,18.6) = 10.90^{**}$	$F(2.6,36.1) = 28.05^{***}$
Parietal	$F(1.8,24.5) = 77.99^{***}$	$F(3.8,52.6) = 57.58^{***}$	$F(1.8,24.5) = 89.56^{***}$	$F(3.6,50.9) = 65.06^{***}$
Prefrontal	$F(1.3,18.4) = 18.39^{***}$	$F(2.4,34.2) = 27.00^{***}$	$F(1.5,21.0) = 20.23^{***}$	$F(3.1,43.9) = 24.94^{***}$
Caudate	$F(1.6,22.2) = 4.63^{*}$	$F(2.9,40.8) = 4.72^{**}$	$F(1.5,21.2) = 4.41^{*}$	$F(2.7,37.8) = 3.90^{*}$
Fusiform	$F(1.7,23.8) = 84.85^{***}$	$F(2.9,39.9) = 61.33^{***}$	$F(1.3,18.4) = 85.20^{***}$	$F(2.5,34.3) = 57.72^{***}$

Note. Degrees of freedom and *p*-values were computed using the Greenhouse-Geisser correction.

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 3

Summary of Model Fit Statistics

		Statistic	
Region	Magnitude parameter	χ^2	Correlation
Right Motor	---	210.02***	---
Left Motor	6.35	287.69***	0.71
Right ACC	0.60	181.19***	0.92
Left ACC	0.66	143.15***	0.91
Right Parietal	3.98	89.51***	0.97
Left Parietal	4.24	125.74***	0.97
Right Prefrontal	1.30	40.86	0.97
Left Prefrontal	1.08	45.22†	0.97
Right Caudate	3.77	59.33**	0.89
Left Caudate	4.01	38.55	0.92
Right Fusiform	9.89	484.66***	0.82
Left Fusiform	9.19	464.81***	0.84

Note. $^{\dagger}p < .10$, $*p < .05$, $**p < .01$, $***p < .001$.

Table 4

Table 4

Results of Difficulty × Scan Exploratory Analysis

Region of interest	Brodmann area(s)	Voxel count	Stereotaxic coordinates (mm)			Difficulty condition		
			x	y	z	1 inference	3 inferences	Not provable
a. Right prefrontal	6, 8, 9	618	49	12	36	2.58%	4.44%	7.60%
b. Anterior cingulate	6, 8, 24, 32	415	0	14	40	2.28%	3.52%	5.87%
c. Right anterior prefrontal	10	87	33	50	13	2.68%	4.33%	6.94%
d. Left anterior prefrontal	10, 46	82	-34	51	12	2.29%	4.30%	7.30%
e. Right insula	13, 47	181	35	16	6	2.01%	2.58%	5.29%
f. Left insula	13	69	-32	15	7	1.61%	2.42%	4.65%
g. Left thalamus	--	81	-12	-17	10	1.69%	2.56%	4.27%

Figure 1

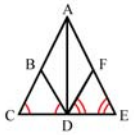
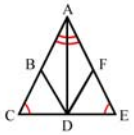
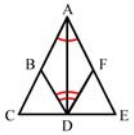
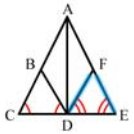
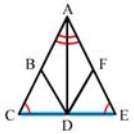
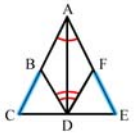
Highlight	Difficulty: 1 inference	3 inferences	Not provable
Goal pair not highlight	 FD ≡ FE	 CD ≡ ED	 BC ≡ FE
Goal pair highlighted	 FD ≡ FE	 CD ≡ ED	 BC ≡ FE

Fig. 1. Example geometry proof problems for each level of the difficulty and highlight factors.

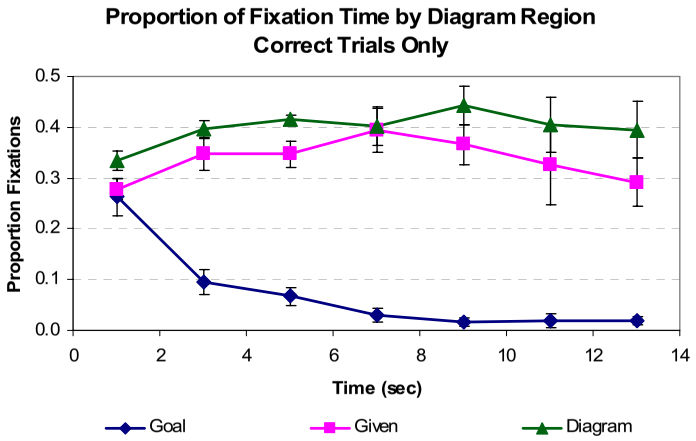


Fig. 2. Mean proportion of fixation time by region in the first 14 seconds of a trial.

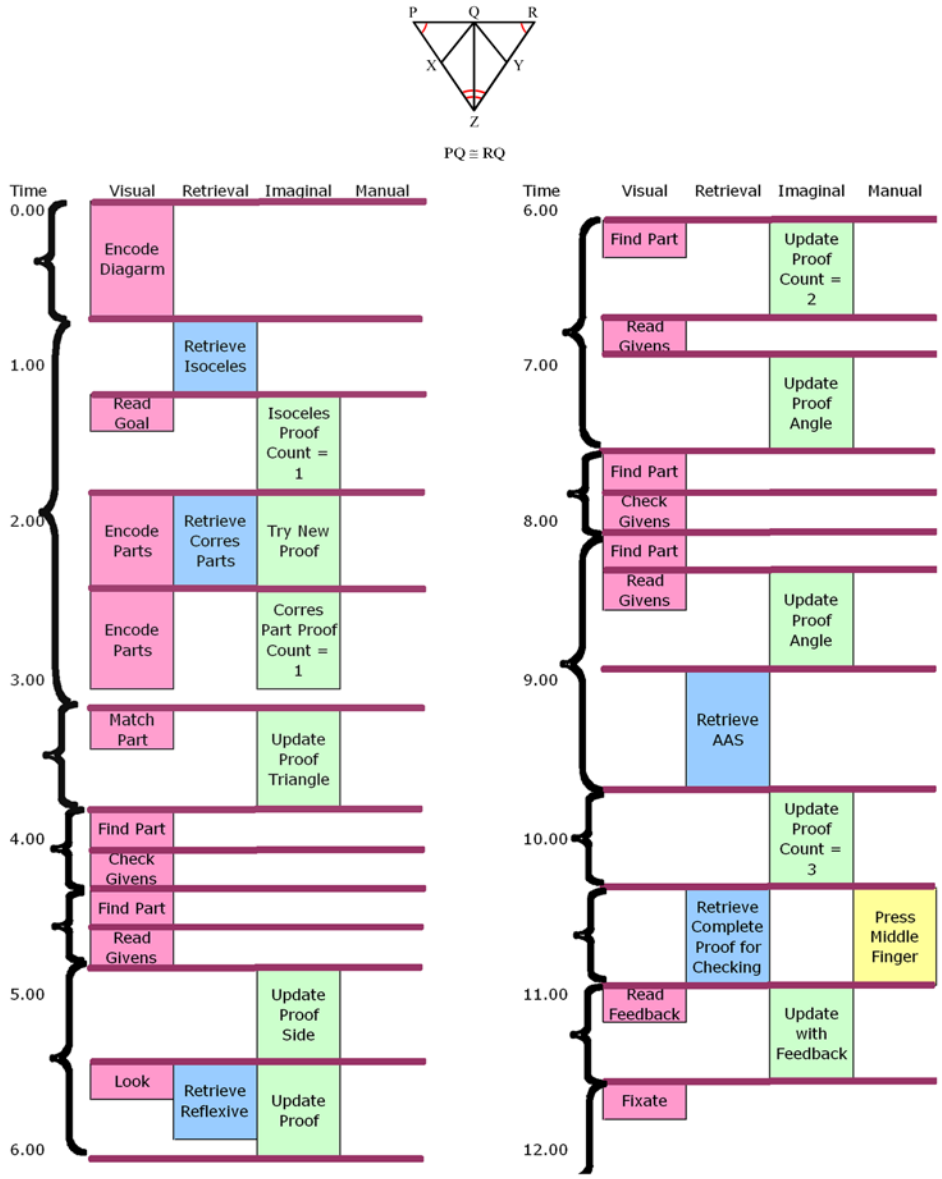


Fig. 3. Activation of the visual, retrieval, imaginal, and manual modules while solving the example problem. Time is given in seconds. Lengths of boxes reflect approximate times the modules are engaged. The horizontal lines represent the firing of production rules. Brackets indicate subtasks of activity controlled by a setting of a goal.

Figure 4

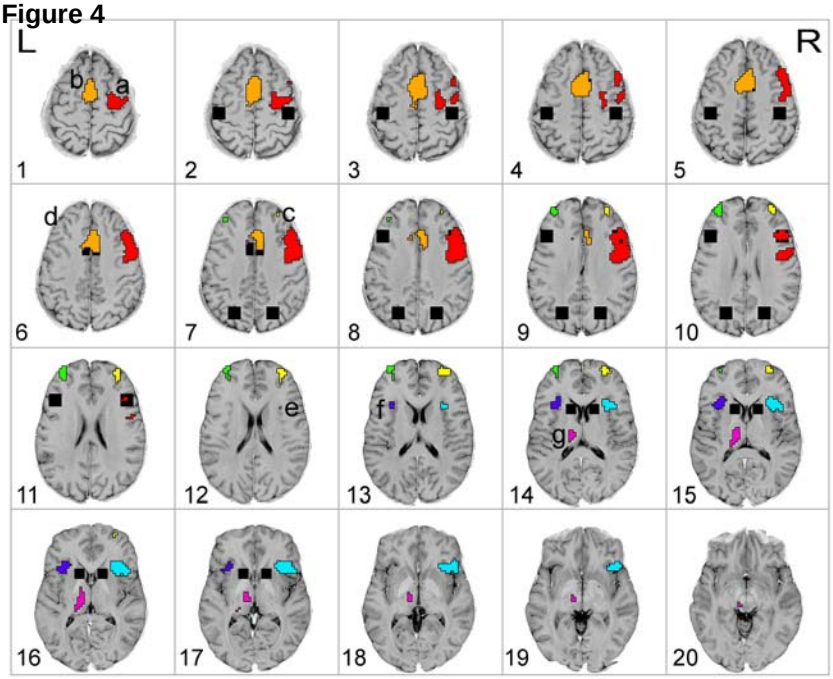


Fig. 4. Predefined and exploratory regions. Predefined regions are shown in black; exploratory regions a-g are shown in solid blocks of color with black borders. The exploratory regions are: a) right frontal—red, b) anterior cingulate—orange, c) right anterior prefrontal—yellow, d) left anterior prefrontal—green, e) right insula—light blue, f) left insula—indigo, and g) left thalamus—magenta.

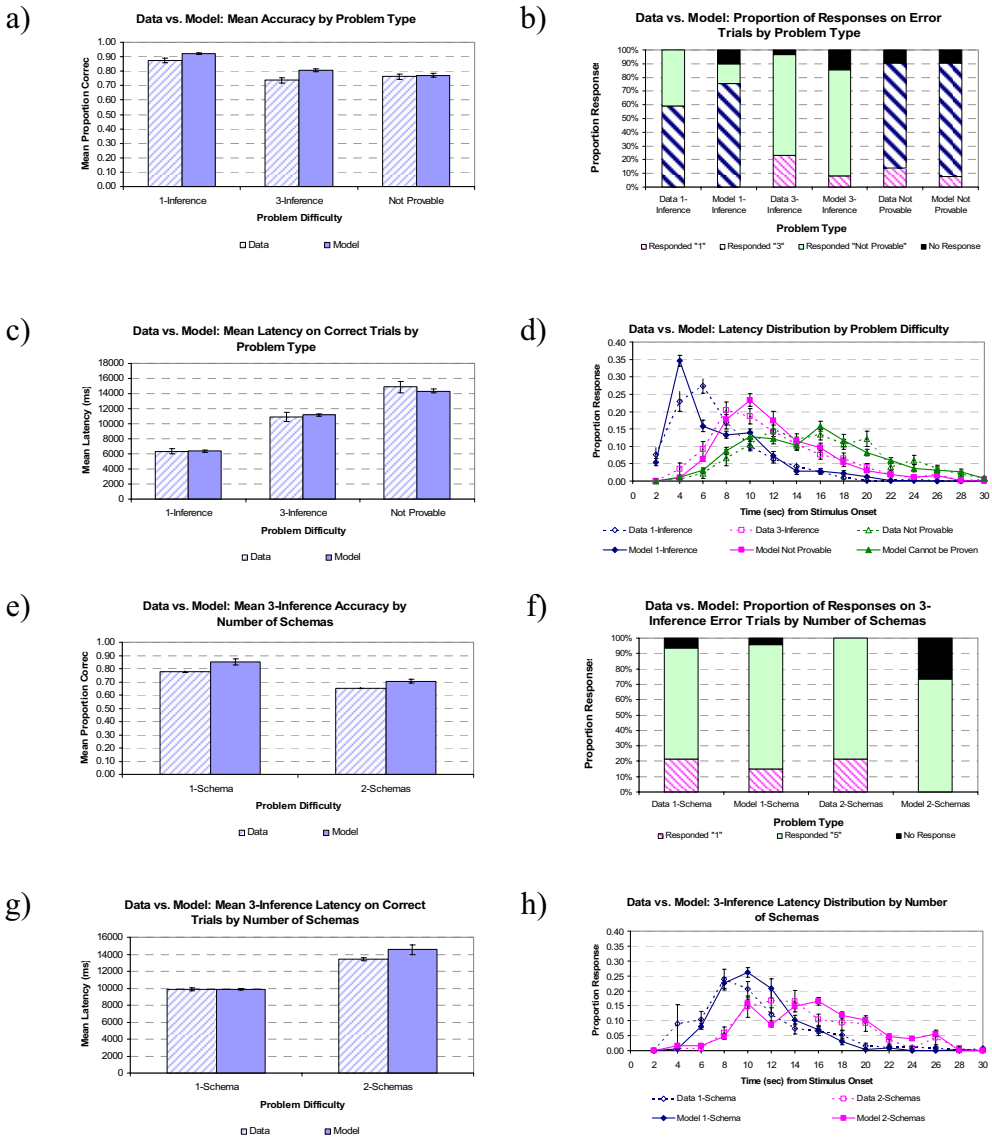


Fig. 5. Model vs. data in each difficulty condition for a) mean accuracy, b) error distribution, c) mean latency, and d) the latency distribution. Also, model vs. data for schema analysis of 3-inference problems for e) mean accuracy, f) error distribution, g) mean latency, and h) the latency distribution. Error bars represent standard error.

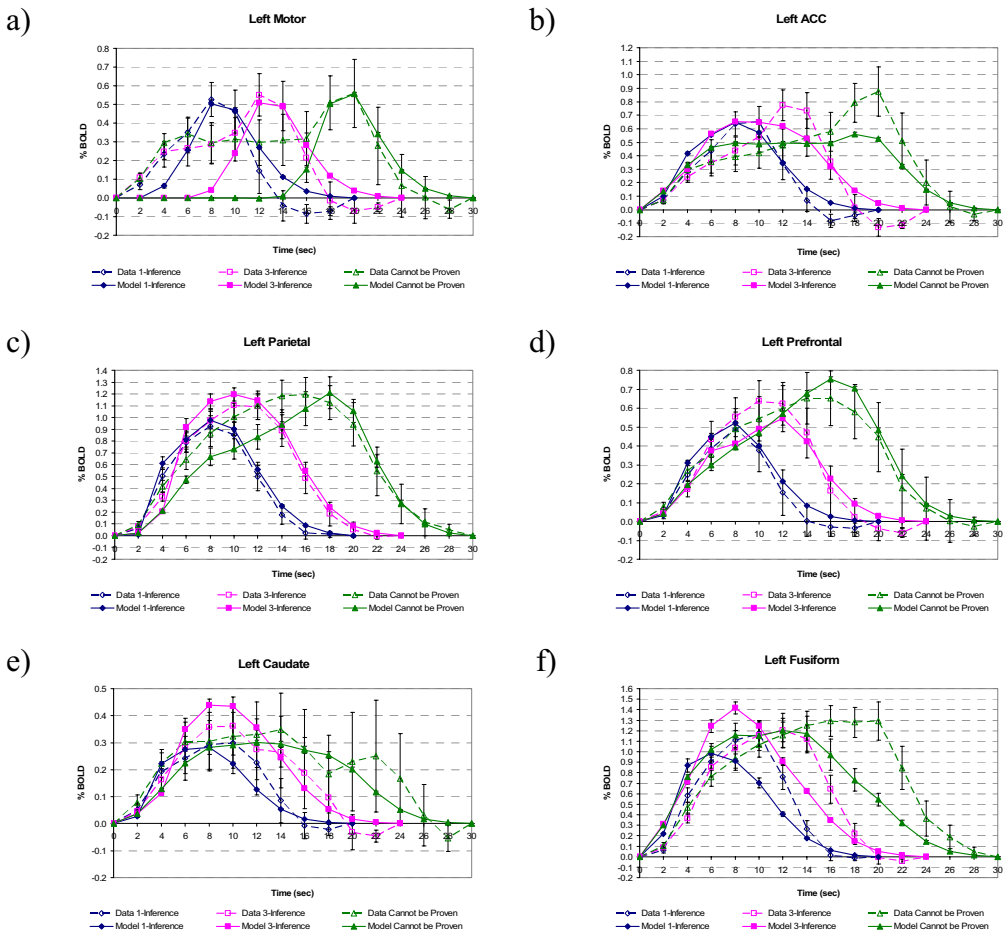


Fig. 6. Predicted and actual mean % BOLD change in the a) left motor, b) left cingulate, c) left parietal, d) left prefrontal, and e) left caudate predefined regions. Trials have been warped to the median length for each condition using the Split-Fincham method and then averaged. As a result, problem onset is always at time zero and the button-press response always occurs at the ninth data point from the end.

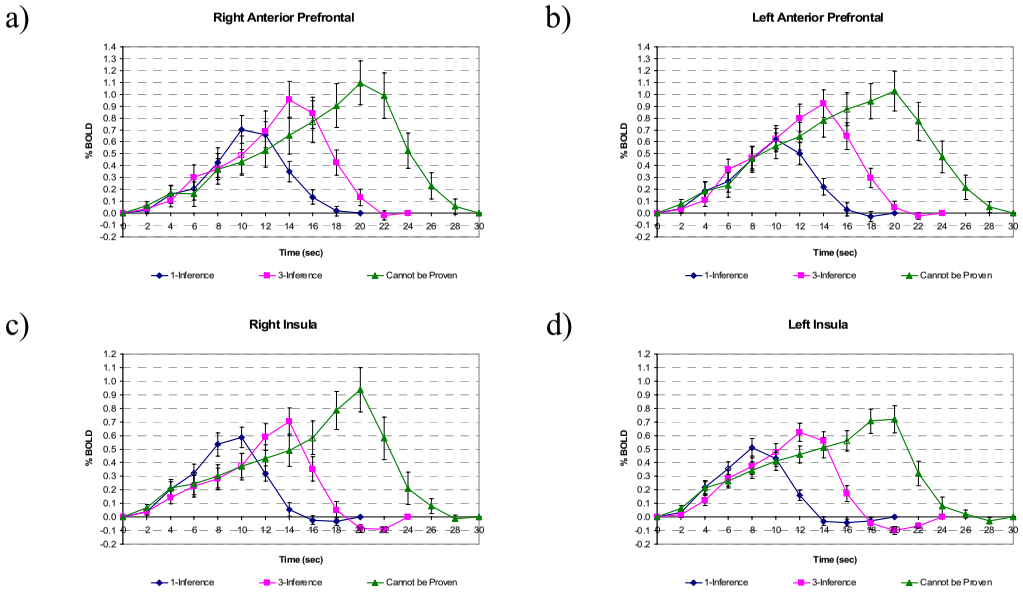


Fig. 7. Percent change in the BOLD response in four exploratory regions: a) right anterior prefrontal, b) left anterior prefrontal, c) right insula, and d) left insula.

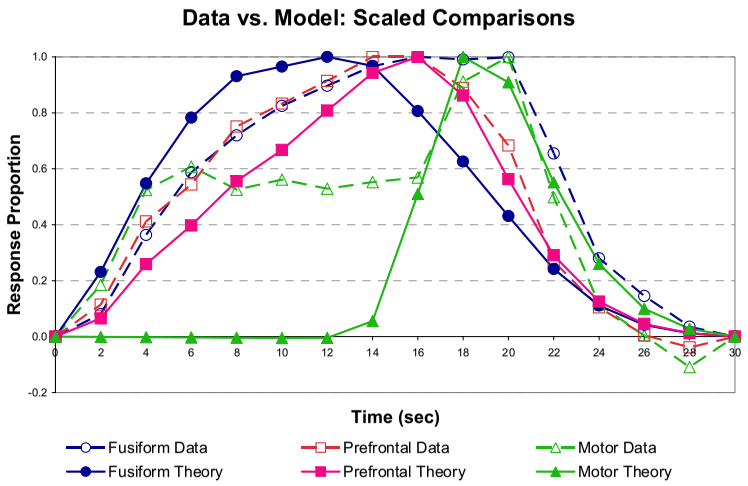


Fig. 8. Scaled comparison of data and model predictions in the fusiform, prefrontal, and motor regions for not-provable problems.