Tenison, C., Fincham, J., & Anderson, J. (2014). Detecting math problem solving strategies. An investigation into the use of retrospective self-reports, latency and fMRI data. *Neuropsychologia, 54,* 41-52.
DOI: http://dx.doi.org/10.1016/j.neuropsychologia.2013.12.011

Title: Detecting math problem solving strategies. An investigation into the use of retrospective self-reports, latency and fMRI data

Caitlin Tenison[1]*, Jon M. Fincham[1], John R. Anderson[1]

[1] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA. 15213

Correspondence to:

Caitlin Tenison

Department of Psychology ~ Baker Hall 342c

Pittsburgh, PA 15213

Email: ctenison@andrew.cmu.edu

Abstract: This research explores how to determine when mathematical problems are solved by retrieval versus computation strategies. Past research has indicated that verbal reports, solution latencies, and neural imaging all provide imperfect indicators of this distinction. Participants in the current study solved mathematical problems involving two distinct problem types, called ‘Pyramid’ and ‘Formula’ problems. Participants were given extensive training solving 3 select Pyramid and 3 select Formula problems. Trained problems were highly practiced, whereas untrained problems were not. The distinction between untrained and trained problems was observed in the data. Untrained problems took longer to solve, more often used procedural strategies and showed a greater activation in the horizontal intraparietal sulcus (HIPS) when compared to trained problems. A classifier fit to the neural distinction between trained-untrained problems successfully predicted training within and between the two problem types. We employed this classifier to generate a prediction of strategy use. By combining evidence from the classifier, problem solving latencies, and retrospective reports, we predicted the strategy used to solve each problem in the scanner and gained unexpected insight into the distinction between different strategies.

Keywords: Arithmetic, Problem solving, Strategy assessments, fMRI

Main Text:

1 Introduction

The ability to accurately identify the strategy a student uses to solve a math problem can help the teacher provide guidance and feedback; however, any number of different strategies could possibly be used to solve a single arithmetic problem. While some strategies can be identified by outward behaviors, such as finger counting, other strategies are more difficult to distinguish. Various methods of strategy detection vary in trade offs between reliability, reactivity, and ease of use. The current study investigates three different indicators of strategy use: verbal reports, response latency, and fMRI evidence. We will describe a method for combining these three indicators that has applicability beyond study of math cognition.

One basic division between types of strategies lies in the distinction between procedural and retrieval. Procedural strategies use mathematical computations, whereas, retrieval strategies involve the recall of facts from long-term memory. Procedural strategies are variants of calculations, such as, counting or transforming an unknown problem into known sub-problems (25+9= 20+5+9). Retrieval strategies, on the other hand, allow a person to arrive at solutions more quickly and accurately than when utilizing procedural strategies (LeFevre, J.-A., Sadesky, G. S., & Bisanz, J., 1996b; Siegler & Shrager, 1984). While complex procedural methods might include some retrieval within the mix of strategies used for problem solving, in this paper, we use retrieval to refer to those strategies in which answers are arrived at directly from memory. Previous developmental studies documented that as children gain experience solving problems they progress from the more mentally arduous procedural strategies to efficient retrieval (Geary, D. C., Brown, S. C., & Samaranayake, V. A. 1991; Imbo & Vandierendonck, 2007). Despite this evidence of progression towards retrieval, both children and adults do not consistently use just one strategy, but instead, a mix of strategies when solving problems (Campbell & Xue, 2001; Hecht, 2006; LeFevre et al., 1996b; Siegler, 1988). This means that even after training a participant to use a specific strategy, exclusive use of that strategy cannot be reliably assumed.

We will consider three measures of strategy use: participant report, problem solving latency, and brain response. Verbal strategy assessments are based on self-reports and are collected during (*concurrent*) or after (*retrospective*) the participants complete the task being studied. Within the problem solving process, the different methods of verbal strategy assessment vary in their *reliability* (the accuracy of the assessment) and their *reactivity* (the impact of the assessment on the participant's behavior during the experimental task). A second method for determining strategy use comes from the relation between strategy use and problem solving latency (LeFevre et al., 1996a; Siegler & Shrager, 1984). While retrieval tends to be faster than computation, often the latency distributions overlap making latency an imperfect indicator of strategy use. Finally, brain imaging studies show distinctions between retrieval and procedural strategies in areas such as the horizontal intraparietal sultucs (HIPS) and the angular gyrus (Delazer et al., 2003; Grabner et al. 2009). While a few fMRI studies have used verbal strategy assessments to label in-scanner trials, we are aware of no studies that have used fMRI findings to predict strategy use, leaving the accuracy of such a method untested (Cho, Ryali, Geary, & Menon, 2011; Grabner et al. 2011). The three measures discussed here provide information regarding the retrieval-procedural distinction; however, each method is an imperfect measure for assessing what participants are actually doing.

1.1 Validity of Strategy Assessment

When studying the use of math strategies, researchers have often relied on self-reports from the participants. Depending on the study constraints and the questions of interest, procedures for collecting strategy assessments vary widely. Several studies (Geary & Brown, 1991; LeFevre et al, 1996b) have assessed strategy by asking the participants to describe their methods for solving math problems immediately after reporting the solutions (concurrent assessments). Other concurrent assessments provide predetermined sets of strategies from which the participants must choose the method most similar to the one used (Campbell & Timm, 2000; Imbo & Vandierendonck, 2008; LeFevre et al, 1996a). An alternative to concurrent strategy assessments collects retrospective strategy assessments (RSAs) after the set of problems have been solved. RSAs have been used in fMRI studies where participants solved problems during the scanning session and then reported strategy use outside the scanner (Cho et al., 2011; Grabner et al., 2009). Although retrospective reports are generally considered less accurate than concurrent reports (Russo, Johnson, & Stephens, 1989**)**, concurrent strategy reports during a functional magnetic resonance imaging (fMRI) task are less feasible due to the risk of head movements and the fluctuations in the blood-oxygen-level-dependent (BOLD) signal caused by breathing changes associated with speaking (Birn, Smith, Jones, & Bandettini, 2008).

A debate exists as to the best methodology for accurately and unobtrusively eliciting reports of strategy use. Verbal self-reports have been challenged for causing reactivity in participant reports and for not being veridical assessments of actual mental processes (Kirk & Ashcraft, 2001; Russo et al., 1989). In a study of simple arithmetic, Kirk and Ashcraft (2001) found that participants reported an increase in the use of the experimenter's suggested strategy. While differences in latency confirmed the validity of the reports, the reactivity of the assessment on participant behavior meant the findings did not reflect natural strategy use. In addition, while RSAs do not influence strategy selection during the solving process, they present difficulties with respect to reliability. In particular, it is unclear how accurately and consistently participants remember what strategies are employed. This problem is aggravated by strategy variability. For instance, Siegler and Shrager (1984) found that over time children not only switched between strategies, but also, one third of the children in the experiment applied a different strategy to the same problem. This makes it unreliable to generalize strategy use information collected at the end of the task for items solved during a task.

While recognizing these issues, verbal strategy assessments remain an accepted standard and studies indicate behavioral and neuroimaging measures tend to agree (Grabner & De Smeldt, 2011; Imbo & Vandierendonck, 2008; Siegler & Shrager, 1984). Items reported as having been 'retrieved' displayed faster reaction times and fewer mistakes than those reported as solved via procedural strategies (LeFevre et al., 1996a; Siegler & Shrager, 1984). Additionally, the strong problem size effect of procedural strategies is significantly lessened when only retrieval strategies are reported as used (LeFevre et al., 1996a). Neuroimaging studies have explored the neural signatures of the different strategies participants report using. An event related potential (ERP) study found no evidence of the problem size effect in the ERPs for math problems participants reported as retrieval. Items reported as calculated, on the other hand, showed a problem size effect in the recorded ERPs (Núñez-Peña, Cortiñas, & Escera, 2006). Another more recent neuroimaging study by Grabner and De Smeldt (2011) specifically designed to test the validity of strategy assessment combined both electroencephalography (EEG) methodology and concurrent strategy assessment data. This study used event-related synchronization and desyncronization ERS/ERD to record frequency bands commonly associated with retrieval and procedural processes. The results indicated a significant relationship between the strategies reported and the frequency bands of the corresponding problems.

1.2 Arithmetic training effects and the brain

Several fMRI studies of strategy use have used math-training paradigms to influence the strategies participants chose to use (Delazer et al., 2005; Ischebeck, Zamarian, Egger, Schocke, & Delazer, 2007). As mathematical problems are practiced, there tends to be a strategy switch from procedural to retrieval (Imbo & Vandierendonck, 2008; Logan & Klapp, 1991). The Distribution of Associations (DOA) model (Siegler & Shrager, 1984) explains how practice solving arithmetic problems results in an increased association between problem and solution and a decrease in confusion between similar math facts. Procedural strategies are often used when the solution to a problem cannot be quickly or confidently remembered, whereas, direct retrieval occurs when there is a strong association between the problem and solution. The DOA model has been supported by findings from several experiments, as well as observed in various participant populations (Campbell & Timm, 2000; Geary & Brown, 1991; Imbo & Vandierendonck, 2008; Reder, 1988).

Practice of problem solving generally results in a switch from procedural to retrieval strategy use, and is associated with changes in activation in different areas of the brain. Imaging studies investigating the effects of training (practice) on solving math problems have reported clear distinctions in neural activity between trained and untrained problems. Several math training studies (Delazer et al., 2003; Delazer et al., 2005; Grabner et al., 2009; Ishebeck et al., 2006) report that untrained, in contrast to trained problems, activates the frontal-parietal network. The parietal region commonly observed in arithmetic studies, the horizontal intraparietal sulcus (HIPS) is generally associated with calculation processes, magnitude comparison and counting (Dehaene, Piazza, Pinel, & Cohen Kadosh, 2003; Cohen Kadosh, Lammertyn, & Izard, 2008). The frontal regions observed in arithmetic studies include the lateral inferior prefrontal cortex, which is generally associated with more complex computations that require a degree of planning and working memory (Anderson, Betts, Ferris, & Fincham, 2011). The comparison between retrieval and calculation shows that the practiced problems involved less activity in these regions and more activity in the angular gyrus (Delazer et al, 2003).

1.3 Brain imaging to detect specific cognitive processes

While there have been studies which used brain imaging to investigate the reliability of verbal strategy reports (e.g., Grabner et al., 2009; Grabner et al., 2011), these studies have not taken advantage of the power of multi-voxel pattern analysis

(MVPA). There have been applications of MVPA, however, that predict other kinds of distinct cognitive processes (Damarla & Just, 2012; Pereira, Mitchell, & Botvinick, 2009; Shinkareva et al., 2008). In practice, a classifier is first trained on a distinction of interest using a sample of fMRI data. It is then tested on a separate data set to see how well it can detect the same distinction. How well the classifier distinguishes items reflects both the uniqueness of the two elements within a set being compared and the similarity between the sets of data used for training and testing. Most of the research on classification of brain activity requires a definition of "ground truth". Simply put, the experimenter knows what the participant is doing. It is natural to use verbal reports to assess the ground truth for the classification of fMRI data, but, as we have noted, there are questions concerning the validity of these reports. Even if there is a solid ground truth, these classification methods are imperfect and do not always correctly classify events. As with the previously discussed methods of evidence, classification also offers an imperfect measure of strategy use.

1.4 Current Study

We used a 2x2 within-subject factorial design where we manipulated training (trained instances vs. untrained, novel instances) and problem type (two distinct problems each requiring different computational procedures to solve). We used both training and problem type as means of manipulating the strategies that participants used during the scan task. We contrasted trained and untrained items to examine the effect of training on latency, brain activation, and RSA. In line with previous studies, we expected untrained items to take significantly longer for participants to solve than trained items (Delazer et al., 2003; Imbo & Vandierendonck, 2008). Furthermore, we expected reduction in activation of the HIPS region for trained problems in comparison to untrained problems (Delazer et al., 2003; Delazer et al., 2005; Ishebeck et al, 2006). Numerous findings about the effect of training on strategy use lead us to expect RSAs to predominantly indicate the use of retrieval strategies when solving trained problems and procedural strategies when solving untrained problems (Imbo & Vandierendonck, 2008; Logan & Klapp, 1991). Thus, in our latency, fMRI, and RSA measures we expected the contrast between trained and untrained problems mirror the contrast between retrieval or computation strategies.

The current study used two distinct problem types; one (referred to as ‘Pyramid problems’) involved adding several numbers together and the other type (referred to as ‘Formula problems’) involved the use of division, subtraction and addition. We chose to use two problem types in our study to test if our method of classifying problem solving strategy is problem specific or can generalize. Previous studies have found that neural distinctions between different problem types decrease when problems are solved by retrieval (Delazer et al., 2009; Ishebeck et al, 2006), suggesting that neural patterns converge in problem retrieval. We hypothesize that the shift from procedural to retrieval strategies will be generalizable across problem types.

The principal interest of this study was to form a prediction of strategy use for each item solved in the scanner. We trained a classifier to distinguish between trained and untrained problems. We expected the classifier to use neural evidence of retrieval in making this classification. For any given trial, the classifier computes the evidence that the trial belongs to the ‘trained’ category. This evidence can be used as an indicator of whether the trial involved retrieval. We also collected RSAs where participants tell us after the fMRI session whether a retrieval strategy was employed for particular problems.

Just as the classifier is an imperfect reflection of retrieval, such a report is also imperfect. This is particularly apparent given the research that indicates participants switch between computation and retrieval for the same problems on a trial-by-trial basis (Russo et al., 1989; Siegler and Shrager, 1984). Finally, we used the latency on a particular trial as another imperfect indicator of strategy use. We compared the classifier, problem solving latencies and retrospective reports and investigated how to best combine all three measures to improve our ability to assess strategy usage.

2 Material and Methods

The current study spanned two consecutive days. On the first day, the participant completed a behavioral training session. The second day, the participant returned for an fMRI scan.

2.1 Participants

Twenty university students (10 females; mean age 21.8/ SD 2.2) participated in the study. Participants were recruited by posting flyers on university online message boards. Participants gave informed written consent and received monetary compensation for their participation. All participants were right handed and had normal or corrected-to-normal vision. The university ethics board approved the study.

2.2 Materials

Each participant was trained on two novel operations associated with distinct procedural algorithms. The first type of problem, called a ‘Pyramid problem’, uses the same algorithm as in the experiment by Anderson et al. (2011). Pyramid problems have

'$' as an operator between two values, the base and the height (i.e., B$H, where B is the base and H the height). The base is the starting number and the height indicates the total number of terms to add together where each item is one less than the previous. For example, 11$4 would be expanded as 11+10+9+8, and thus evaluates to 38. The second type of problem, the 'Formula problem', uses '#' as an operator and is evaluated according to the equation V#H=(V/H)+(H/2)-0.5. This formula was specifically chosen because it represents a method of finding the base of a Pyramid problem given the height and the value of the Pyramid problem (i.e., 11$4=38 and 38#4=11). Participants were unaware of this relationship (as confirmed by strategy assessment after the study). The maximum base (B) was 10, and the smallest base for any problem set was height minus one. To control for difficulty between the two problem types, only problems with a height (H) of 3, 4 and 5 were used.

2.3 Training Procedure

During one 35-45 minute session, participants were initially taught how to solve both a Pyramid and Formula problem and then given practice on both to help them memorize answers for three problems from each problem type. The Pyramid problems were taught by explaining the role of the base and height of the problem. Participants were told to start with the base and then count down until the number of terms they had listed was equal to the height. After that, they were instructed to add these values together. To teach Formula Problems, the experimenter showed how the variables in the V#H expression mapped onto the equation previously shown, and then gave the participants time to study the equation.

The experimenter worked with the participant to solve three sample problems of each problem type before beginning the computer-based training. During the computer training participants were exposed to 6 distinct problems (3 Pyramid and 3 Formula problems). The memorized Pyramid and Formula problems were matched in difficulty, featuring base values of 3, 4, 5, and distinct height values. After each problem was solved, the participant received correctness feedback and instruction on how to solve the problem. This feedback featured the problem shown and the mathematical productions required to calculate the answer (either the iterative addition or the formula filled in with the appropriate terms). Training was divided into 6 blocks; for each block, the participant was drilled on each problem 6 times. Over the course of training, the participant solved each problem 36 times. In between blocks, the participant practiced using a numeric keypad with a similar layout as the one used during the fMRI scan. Since the participant cannot see the keypad during the fMRI scan, a cover was placed over the keypad halfway through training in order for the participant to practice using the keypad by touch only.

2.4 Scanning Procedure

On the day following the training session, each participant completed 6 fMRI imaging blocks. Each block contained 3 randomly ordered sets of 7 Pyramid problems (3 trained and 4 untrained), 7 Formula problems (3 trained and 4 untrained) and 7 Addition and Division problems. Untrained problems matched the trained problems in difficulty. The first problem of each set was omitted from statistical analysis to avoid warm-up effects and effects of switching operations. Untrained problems were repeated twice (once per 3 imaging blocks) and trained problems were repeated six times (once per imaging block). Problems were presented in a slow event-related design. During the solving stage of the trial, the participant had a maximum of 30 seconds to indicate knowledge of a solution by pressing the return key on the numeric keypad. After pressing 'return', participants then had 5 seconds to input a solution and press the return key. Because there is no backspace key on the keyboard, participants were unable to correct any typos made when entering answers. After answering the problem, the participant was given correctness feedback and information about how the problem should have been solved. This feedback was the same as that which was received during the previous day's training. In between problems, a 12 second repetition detection task was presented onscreen to prevent metacognitive reflection on the previous problem and allow the hemodynamic response to return to baseline. In the repetition detection task, participants were asked to judge if the letter in the center of the screen was the same as the previous letter seen. Because the participant determined solving and input time, the length of the scan task was variable across participants and trials.

2.5 Strategy Assessment

Following the completion of all 6 imaging blocks, participants exited the scanner and then solved 16 paper-based problems which included all the trained and some of the untrained problems that they had solved in the scanner. Problems were presented on paper, and included 8 Formula problems (3 trained, 5 untrained) followed by 8 Pyramid problems (3 trained, 5 untrained). Participants were asked to report the solution to the problem and then explain how it was solved. The researcher recorded the participant's verbal report. The participants' strategy self-reports were coded based on five categories:

retrieval, calculated using method taught, guessed, calculated using an alternative method and started to solve for the answer but retrieved the solution half way through calculation [from here on referred to as the “partial” strategy]. We used these codes as an indicator of how these problems had been solved in the scanner.

2.6 fMRI Data Acquisition

Images were acquired using gradient echo-echo planar image (EPI) acquisition on a Siemens 3T Verio Scanner using a 32 channel RF head coil, with 2 s. repetition time (TR), 30 ms. echo time (TE), 79° flip angle, and 20 cm. field of view (FOV). The experiment acquired 34 axial slices on each TR using a 3.2 mm thick, 64×64 matrix. This produces voxels that are 3.2 mm high and 3.125 x 3.125 mm^2. The anterior commissure-posterior commissure (AC-PC) line was on the 11th slice from the bottom scan slice. Acquired images were pre-processed and analyzed using AFNI (Cox, 1996; Cox & Hyde, 1997). Functional images were motion-corrected using 6-parameter 3D registration. All images were then slice-time centered at 1 sec and co-registered to a common reference structural MRI by means of a 12-parameter 3D registration and smoothed with an 6 mm full-width-half-maximum 3D Gaussian filter to accommodate individual differences in anatomy.

2.7 fMRI Analysis

We processed the fMRI data used for region of interest (ROI) analysis and classification analysis in different ways. For the ROI analysis, the two predefined regions we focused on were the left and right HIPS from the triple-code theory (Dehaene, 1997) expecting to replicate the finding of less activation for trained problems. To locate these regions we used the coordinates from the meta-analysis of Cohen Kadosh, Lammertyn, and Izard (2008) (Maxima TC: -31, -50, 45). Imaging data were analyzed using a GLM. The design matrix consisted of 9 model regressors and a baseline model of an order-4 polynomial to account for general signal drift. The 9 model regressors corresponded to the following: the solution periods of correct trained and untrained Pyramid and Formula problems (4 regressors) and Addition and Division problems (2 regressors); response activity for solution entry (1 regressor across all problem types); the feedback period (1 regressor) and finally a single regressor for the solution period of incorrect problems. The design matrix regressors were constructed by convolving the boxcar functions of these variables with a hemodynamic function (assuming the standard SPM hemodynamic response – Friston et al., 1998)[1]. The GLM yielded 9 beta weights per voxel for each participant. Our focus was on those representing solution periods for the correct Pyramid and Formula problems. Our ROI analysis was done on the averaged beta weights of voxels within the HIPs regions.

2.8 fMRI Classification Analysis

To prepare the fMRI data for a linear discriminate analysis (LDA), we went through a number of steps to restrict the set of features used by the classifier to avoid over-fitting the data and impacting the reliability of our results (Pereira et al., 2009). For the first step we subdivided the brain into 4x4x4 voxel cubes (a voxel is 3.2 x 3.125 x 3.125mm) over 32 slices of the 64x64 acquisition matrix to create an initial 408 'mega-voxel' ROIs (Anderson, Betts, Ferris, & Fincham, 2010). We chose to divide the brain in

[1] The difference between two gamma functions used was gamma (6,1) – gamma (16,1)

this manner because it maximizes the trade offs between two objectives. First, we did not choose larger anatomical ROIs because we want a whole brain perspective that uses variation within regions. Second, we did not choose a smaller voxel-level analysis because we want to find patterns that exist across participants and using the 4x4x4 mega-voxels smooths the brain making it more likely that the mega-voxels overlap between participants.

The second step was to eliminate regions that had highly variable fMRI signals. A measure of variability was calculated for each of the 6 imaging blocks by dividing the block range by the mean. ROIs containing more than 15 TRs across all participants that fluctuated more than 15% during a block were eliminated. The reduced sample comprised 266, 4x4x4 voxel regions of raw data. Previous work using the 408 voxels found reducing the voxels by half had no effect on classifier accuracy reflecting the correlations that exist among voxels (Anderson et al., 2010)[2]. The majority of the regions eliminated from the analysis were the most dorsal and ventral slices or on the edges of the other slices. For the 266 regions, we estimated 23 regressors for each subject for each block: one input regressor for all the trials, one feedback regressor for all trials, and 21 solving period regressors, one for each problem. We constructed the design matrix regressors by convolving the boxcar functions of each of the regressors with a hemodynamic function (see footnote 1). Each boxcar function's duration was equal to the time the participant spent on the trial (Grinband et al., 2008). A GLM was used to estimate the betas for each problem as well as the input and feedback periods of the scan block. Combining our results across blocks we get an estimate of engagement (a beta

[2] This finding was echoed in the current experiment. Running the classification described in section 3.4.1 on all 408 ROIs rather than the 266 regions, we continued to be able to distinguish the training distinction (mean d-prime= 1.37, *t(19)* =8.27, *p*<.0001).

from the GLM) during problem solving for each of the 6 x 21 = 106 trials and these are the values that we will use in our classification analyses

As a third step of dimensional reduction, we used a Principle Components Analysis (PCA) to create uncorrelated variables from linear combinations of the ROI activity. We chose this method of dimensional reduction for three reasons: the ease with which we can map classifier results back to the brain, the reduction in computational cost of running the classifier, and because it eliminates redundancies in the data (Ford et al., 2003; Thomaz et al., 2004). The PCA was performed on the z-scores of the beta values for the 266 regions. Using z-scores rather than raw values allowed for comparison across subjects. To eliminate fluctuations in the BOLD signal that were physiologically implausible, z-scores were Winsorized such that scores greater than 5 or less than -5 were changed to 5 or -5 respectively[3]. The combination of both trimming the regions that fluctuated excessively and Winsorizing the data is a commonly accepted technique for removing the largest outliers (usually due to signal drop out) while maintaining the “true” outliers that deviate less severely (Kettaneh et al., 2005; Marchini & Ripley, 2000).

We then preformed an LDA on the first 50 factors extracted from the PCA. Both the Kaiser-Guttman rule and an analysis of the scree plot indicate that the first 20 components capture the most the variance in our data. We chose, however, to use the first 50 factors because work pairing PCA with LDA suggest it is better to include more factors (Yang &Yang, 2002). We used the LDA to identify which of these factors contributed to distinguishing the categories. Because we were interested in identifying

[3] To assess the effect of Winzoring on our classification analysis we ran the classification described in section 3.4.1 un-Winzorized data. We continued to be able to distinguish the training distinction (mean d-prime= 1.42, *t(19)* =9.66, *p*<.0001), with a hit rate of 63.8% and a false alarm rate of 24.7%. Eliminating Winzoring adds noise, which has negative, but not substantial effects on our classifier.

similar features that exist across participants, we used a leave-one-out cross-validation method. We trained on all but one participant and then tested on the remaining participant. Besides returning a predicted category for each item, an LDA generates a continuously varying evidence measure for category membership and a posterior probability that an item is from a category. Both of these measures were used in subsequent analyses.

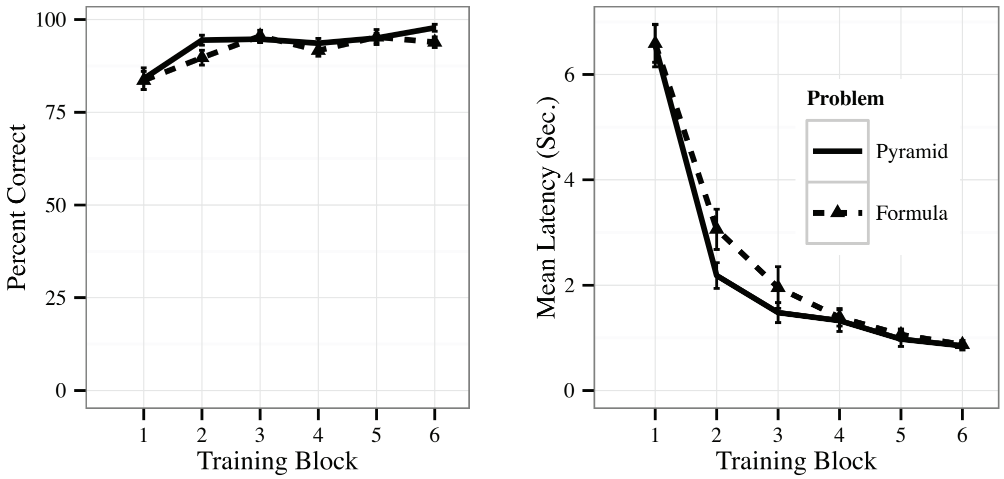

Figure 1. Average problem solving latency and average accuracy scores of each block in the training task with error bars representing standard error.

3 Results

3.1 Behavioral Results

Our measure of latency was calculated as the time between the appearance of the problem on the screen and the point at which the participant pressed the return key to indicate readiness to input the solution. We only considered correct trials in our analysis of latency and the fMRI data. Figure 1 shows average performance during training. A repeated measures ANOVA on solution times of correctly solved problems with training block (block 1 through block 6) and operation (Pyramid and Formula) as factors showed a significant effect of training block, *F*(5,95) = 192.52, *p*<.0001, and a significant interaction between block and problem type, *F*(5,95)=2.72, *p*< .05, but no significant effect of problem type, *F*(1,19)=2.49, *p*=.13[4]. A similar repeated measures ANOVA on accuracy found a significant effect of training block, *F*(5,95)=12.96, *p*<.001, a marginally significant effect of problem type, *F*(1,19)=4.19, *p* =.06, and no interaction, *F*(5,95)=1.28, *p*=.27. Participants reached a high level of performance that we would associate with retrieval.

Table 1
Mean solution time (total time in parentheses) and accuracy scores of the trained and untrained problems of both operation types.

	Trained		Untrained	
	Latency	Accuracy	Latency	Accuracy
Pyramid	2.59 s (4.34 s)	93.10%	5.88 s (7.84 s)	71.50%
Formula	3.35 s (4.83 s)	91.10%	5.68 s (7.33 s)	82.60%

Table 1 shows the results from the fMRI session. Participants were slightly less accurate and considerably slower for trained problems on the day 2 scan task than they were on the last block of the day 1 training. Whereas the training task featured only 6 problems repeated several times, the scanner task consisted of a mix of both trained and untrained problems, making it harder for participants to identify the problems for which they had memorized answers. Nevertheless, participants were still considerably more accurate and much faster on trained problems than on untrained problems. For correct problem solving latencies there is a main effect of training, *F* (1,19)=44.3, *p*<.0001, with

[4] Subjects averaged 1.32 (SD= .29) seconds to type in their answers after solving each problem. A similar analysis of total time (solution plus typing) yields similar statistics: a significant effect of training block, *F*(5,95) = 243.5, *p*<.001 and a significant interaction between block and problem type, *F*(5,95)=2.49, *p*< 0.04 but no significant effect of problem type, *F*(1,19)=0, *p*>.99. We focus on solution time.

trained problems being solved faster than untrained problems. There was, however, no significant main effect of problem type, *F* (1,19)=2.17, *p*=.17, or interaction, *F* (1,19)=3.155, *p* = .92. Accuracy scores showed a significant main effect of training, *F* (1,19)=35.9, *p*≤.001 and problem type, *F* (1,19)=6.47, *p*<.05. There was a cross-over interaction of problem type by training, *F* (1,19)=11.2, *p*<.005: for untrained problems, accuracy was 9% better on Formula problems, but for trained problems accuracy was 2% worse for Formula problems.

3.2 Strategy Assessment

We coded the strategy assessments for 5 strategy types: direct retrieval of the solution, calculation according to the taught strategy, a mix of retrieval and calculation (“partial), guessing, and calculating the problem using an untaught method. Of the untrained problems, 90.8% (Pyramid: 90%, Formula: 92%) were reported as solved by the taught method, 3% were reported as solved using a partial strategy, 2.5% were reported as retrieval and less than 3.5% fell into the other categories. Of the trained problems, 60.5% (Pyramid: 64%, Formula: 57%)were reported as retrieved, 25.2% as calculated according to the taught strategy, 10.9% as partial and 3.4% in the remaining categories.

3.3 Imaging Data – The Effects of Training

We used the Cohen Kadosh et al. (2008) ROI for the HIPS to investigate the effect of training on two different problem types[5]. We ran a hemisphere (left, right) X

[5] An exploratory analysis of the effects of training across problem types is reported in the Appendix A.

problem type (Pyramid, Formula) X training (trained, untrained) repeated measures ANOVA on the HIPS. There were significant main effects of hemisphere, *F* (1,19) = 36.7, *p*<.0001, which reflected less activation in the right than the left hemisphere, as well as a significant training-by-hemisphere interaction, *F* (1,19)=16.2, *p*<.001, which reflected a greater decrease in the left hemisphere activation. There was a significant effect of training, *F* (1,19)=9.2, *p*<.01, but not problem type, *F* (1,19)=1.4, *p*=.25, and a marginally significant operation-by-training interaction, *F* (1, 19)=3.9, *p*=.06. Figure 2 displays a decrease in activation for trained problems in the left HIPS. [6]

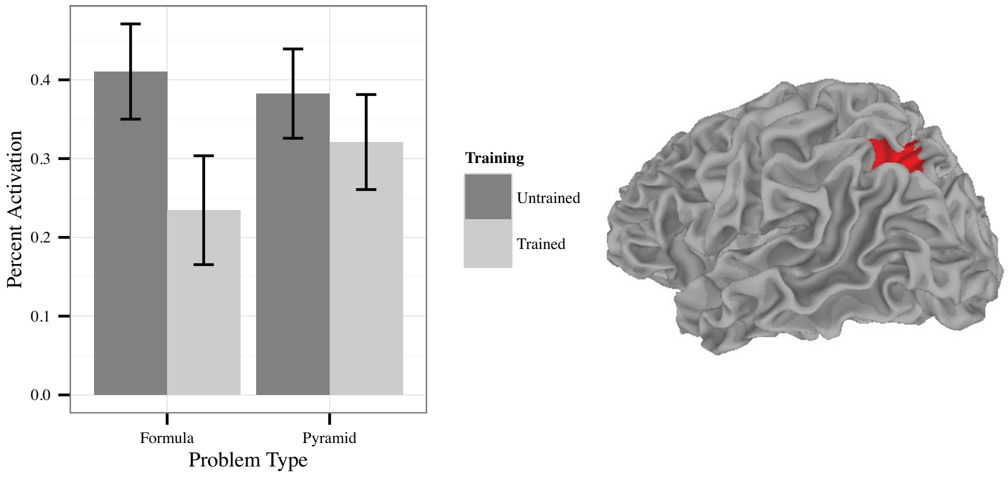

Figure 2. Mean percent change in activation from baseline in the left HIPS for trained and untrained pyramid and formula problems with error bars representing standard error.

3.4.1 Classification of Training

An LDA was conducted using the first 50 factors from the PCA to predict whether or not a problem was trained or untrained. The classifier was trained on all but one participant and then tested on that one participant who had been left out of the training set. We calculated for each participant the proportion of hits (the percentage of

[6] Appendix B reports the results for the angular gyrus.

problems correctly classified) and the proportion of false alarms (the percentage of problems incorrectly classified). From these two numbers we calculated a d-prime measure of discriminability (Wickens, 2001). We calculated the mean d-prime over all participants and performed a t-test of whether or not the d-primes were significantly greater than zero.

We first ran an LDA combining data from both problem types. This classifier was successful in distinguishing the trained and untrained problems, mean d-prime= 1.58, *t(19)* =10.2, *p*<.0001[7], with average hit rates of 63.86% and false alarm rates of 20.1%. The classifier produced positive d-prime in its predictions for each of the 20 participants. We examined the classifier to find out how it was making its discrimination. Classification of the trained and untrained categories is done on the basis of a weighted sum of the 50 factors, which can be viewed as the brain evidence for the problem being untrained. We took the coefficients associated with the factors and calculated weights for each of the 266 brain regions. Figure 3 shows the positive values of these weights for the slice of the brain that contains the HIPS. The most positive regions overlap the HIPS, which our ROI analysis showed as having the greater activation during untrained problems.

[7] Throughout, significance levels are calculated for 2-tailed t-tests.

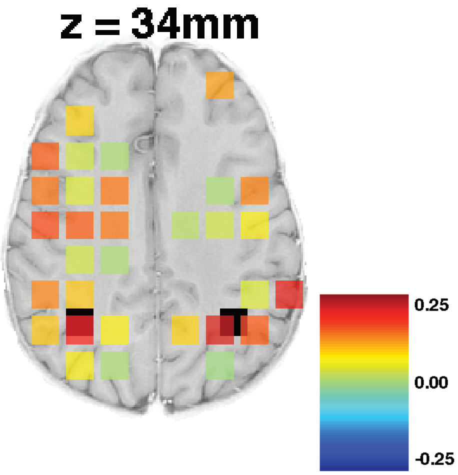

Figure 3. Distinguishing regions used in the untrained vs. trained classification (pyramid and formula together). These are the voxels that are more active for untrained than for trained problems. Locus of the HIPS is marked by the dark squares. The z value is for x=y=0 in Talairach coordinates.

Given that there was a significant effect of training on the HIPS ROI, one might wonder how well an LDA would do using only that region. The classifier was not very successful in distinguishing the trained and untrained problems, mean d-prime= .021, *t(19)* =2.09, *p*=.10, yielding a hit rate of 82.31% but with a false alarm rate of 80.27%. This poor performance is most likely due to considerable trial-to-trial variability in activation. A classifier based on whole brain activation can calibrate activation in any one region relative to activation elsewhere and is thus more robust to variability between trials.

3.4.2 Classification of Training – Similarity Between Problem Types

We were interested in whether or not the same features distinguished trained from untrained problems for the two problem types. To test this we crossed training on

Pyramid and Formula problems with testing on Pyramid versus Formula problems. In all cases, we trained on 19 participants and tested on the 20th. The results were:

Train on Pyramid, Test on Pyramid: d-prime = 1.60, *t(19)* =10.87, *p<0.001*

Train on Pyramid, Test on Formula: d-prime = 1.47, *t(19)* =9.02, *p<0.001*

Train on Formula, Test on Pyramid: d-prime = 1.27, *t(19)* =7.63, *p<0.001*

Train on Formula, Test on Formula: d-prime = 1.28, *t(19)* =8.46, *p<0.001*

In all cases, the effects were highly significant with 19 or 20 out of 20 participants showing positive d-primes. To test if there were significant differences in prediction success, we took the 20 d-primes for each case and subjected them to a 2x2 within-participant ANOVA. The effect of training source approached significance, *F*(1,19) = 3.72, p < .10, but there was no effect of test source, *F*(1,19)=0.57, nor a significant interaction between the factors, *F*(1,19) = 1.40; p > .25. Thus, while Pyramid data may have provided a more reliable case for training, it seems that the same information is being used to discriminate between trained and untrained problems for both Pyramid and Formula problems.

3.5 Validating the Retrospective Strategy Report and use of Classification to detect Strategy Use

Among correctly solved problems, all three measures (reports, latencies, and classifier evidence) show clear differences between trained and untrained problems:

1. Solution time. Across both problem types participants average 2.96 seconds to solve trained problems and 5.81 seconds to solve untrained problems.

2. Retrospective reports. Participants report retrieving solutions for 60% of the trained problems and for 2% of the untrained problems.
3. Classification of imaging data. The classifier provides a measure of evidence that each trial belongs to the trained or untrained category. Figure 4 shows distribution for the trained and untrained problems, with negative values indicating evidence for untrained and positive values indicating evidence for trained. One can take a positive value on this dimension as also evidence for retrieval.

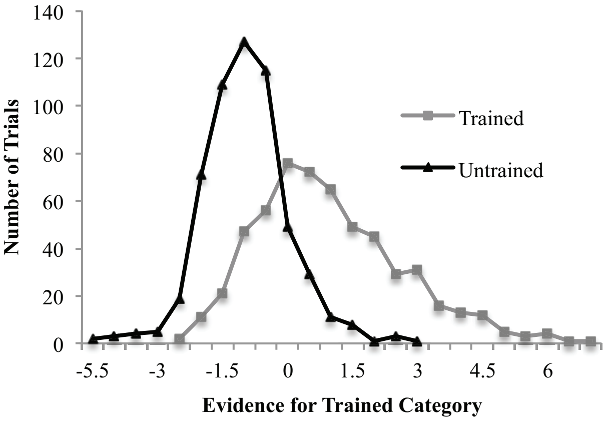

Figure 4. Classifier generated distributions of evidence for training on the trained and untrained data set.

Additionally, we investigated the convergence of latency, RSA, and the classification scores analysis for both trained and untrained problems. The RSAs were collected on all of the trained problems, but only a subset of the untrained problems was assessed. We eliminated trials in the scanner that did not appear in the RSA. We examined the relationship between these reports and the time it took participants to answer the problems during the scanning session. For all the correct problems, Table 2 reports the average times for problems reported as retrieved versus those problems reported as calculated. Participants were 2.53 seconds faster for Pyramid problems reported as retrieved and 1.62 seconds faster for Formula problems reported as retrieved. We built a multilevel model to test this relationship while taking into account the effect of individual differences and training. Furthermore the multilevel model allowed us to deal with the problem that different participants reported different numbers of problems retrieved. This model examined how retrospective reports and training affected latency taking into account the effects of individual differences. To specifically test the significance of the relationship between latency and retrospective reports we used the Markov Chain Monte Carlo method. This is a recommended method to arrive at confidence intervals of models that include random effects (Baayen, Davidson, and Bates, 2008; Goldstein and Browne, 2002). Retrospective reports of retrieval were associated with decreased latency for both Pyramid and Formula problems. Pyramid problems showed an estimated coefficient of -.56, T^2 (19)= -3.1, $p<0.005$, and Formula problems showed an estimated coefficient of -.63, T^2 (19)= -4.3, $p<0.005$.

Table 2
The average in-scanner latencies and evidence scores of trained and untrained problems divided based on retrospective strategy reports

	Reported Retrieval		Reported Procedural	
	Latency	Evidence	Latency	Evidence
Pyramid	2.19 s	*1.82*	4.72 s	*-0.07*
Formula	3.03 s	*1.21*	4.64 s	*-0.21*

We similarly investigated the relationship between problem solving latency and the classifier's evidence of training. We trained the classifier on the distinction between trained and untrained problems using the leave-one-out LDA mentioned previously. We used the classifier to calculate for each item the evidence that it belongs to the trained category subtracted from the evidence that it belongs to the untrained set. We built a multilevel model examining how evidence scores and training affected latency taking into account the effects of individual differences. For both Pyramid and Formula problems, problems classified as trained were faster (Pyramid coefficient of -.28, T^2 (19)= -9, $p<0.005$, Formula coefficient of-.36, T^2 (19)= -9.9, $p<0.005$) .

Finally, we investigated whether there was an association between the classifier evidence and participant reports (Table 2). Both Pyramid and Formula problems were reported as retrieved averaged higher evidence scores than those reported as not retrieved. We ran two multilevel models exploring how retrospective reports and problem type affected evidence, taking into account individual participants. Both problem types showed a significant positive relationship between report of retrieval and evidence scores (Pyramid coefficient of 1.15, T^2 (19)= 4.4, $p<0.005$, Formula coefficient of .76, T^2 (19)= 3.9, $p<0.005$)

Table 3
The distribution of the trained problems according to the retrospective reports (RSA), whether the classifier had positive evidence that the item was trained, and whether the problem solving latency was shorter than the mean solution time (3s).

		Solution Time	
RSA	Classifier	Short	Long
Calculate	Untrained	13	70
	Trained	81	31
Retrieve	Untrained	38	48
	Trained	223	55

3.6 Retrieval Identification for the Trained Problems

Within the trained problems category, each of the three data sources provides evidence of a participant's strategy on every scanner trial. For all the correct problems, Table 3 shows the following distribution of the trained problems: 1) according to retrospective self-report, 2) whether the classifier had positive evidence that the item belonged to the trained set, and 3) whether the problem solving latency was greater than the overall mean, which was 3 seconds. There are some consistencies among these dimensions; however, there are quite a few cases where the sources seem to conflict as to predicting if a problem was retrieved. For instance, on 48 trials (8.6% of trained problems) participants reported retrieval strategies; yet, these problems feature long latencies (M = 5.54 s, SD = 1.88) and the classifier predicted that the problems were calculated (mean -.72, Figure 4). Similarly, on 81 trials (14.5% of trained problems) participants reported the use of procedural strategies; yet, these problems feature fast latencies (M = 1.98 s, SD = .5) and strong classifier evidence for retrieval (M = 2.08, SD = 1.53). More generally, nearly half of the cases (47.6%) involve two of the measures voting one way, and the other measure voting in the opposite direction.

How can one best combine these three sources of information (RSAs, latencies and classification evidence scores) to predict how a problem was solved? Essentially, the solution process is a hidden state that is associated probabilistically with these three observable measures. We used expectation maximization to fit a mixture model (Aitkin & Rubin, 1985; Bailey & Elkna, 1994) to these three measures to identify the hidden process states for the trained problems. A mixture model assumes that partially correlated measures arise from a mixture of states in which the measures are independent. We suspected that there were two such states, retrieval and calculation, and that we were observing a mixture in which some trials reflected retrieval and some trials reflected calculation; however, not to prejudge the matter we searched for the best number of states over the range 1 to 10.

Each latent state *i* is associated with a probability p_i of reporting retrieval, a distribution F_i of latencies, and a distribution G_i of evidence from the classifier generated probabilities. We assumed that the latency distribution F_i will be shaped as a gamma with shape v_i and scale a_i and that the classification evidence distribution was distributed as a normal with mean μ_i and standard deviation σ_i. Then the probability of a trial *j* with reported strategy r_j, latency t_j, and evidence e_j, is:

$$\text{Prob}(r_j,, t_j, e_j) = \sum\nolimits_{\text{i}} [s_i * P_i\,(r_j) * F_i\,(t_j, v_i, a_i) * G_i\,(e_j, \mu_i, \sigma_i)\ ]$$

where s_i is the probability of being in state i and $P_i(r_j) = p_i$ if the RSA r_j is retrieval and 1- p_i if it is compute. Each state requires estimating 6 parameters – s_i, p_i, v_i and a_i , μ_i and σ_i. To assure that expectation maximization did not get stuck on a local maximum, we ran 12,000 parameter searches from random starting points for each number of states.

To our surprise, the best model involved 3 states rather than 2. For comparison, Table 4 reports the parameters estimated and the measures of goodness of fit for models with states varying from 1 to 4. More states mean more parameters, resulting in better fitting models and higher log likelihoods. To correct for this we calculated the Bayesian information criterion (BIC) that penalizes the fits for number of parameters (Lewandowsky & Farrell, 2011):

$$\mathrm{BIC} = -2 \ln L + k \ln n,$$

where *L* is the likelihood value from maximum likelihood estimation, *k* is the number of free parameters, and *n* is the total number of observations contributing to the data. The 3-state model results in the lowest (best) BIC value.

Table 4
Parameter estimates, measures of goodness of fit and Bayesian information criterion (BIC) reports for three models.

	Proportion	Report	Latency (Gamma)		Evidence (Normal)		Log	Number of	BIC
	of Trials	Probability	Mean	SD	Mean	SD	Likelihood	Parameters	Measure
1-State	1.000	65.1%	3.00	1.75	1.16	1.87	-2541	5	5118
2-State	0.533	78.7%	1.76	0.60	2.27	1.78	-2334	11	4749
	0.467	49.7%	4.41	1.90	-0.11	0.90			
3-State	0.217	88.2%	1.29	0.25	3.03	2.06	-2298	17	4722
	0.325	67.4%	2.16	0.53	1.81	1.17			
	0.457	52.5%	4.41	2.07	-0.20	0.91			
4-State	0.320	86.7%	1.47	0.55	2.57	2.13	-2284	23	4738
	0.216	63.3%	2.18	0.45	1.86	0.92			
	0.340	54.3%	3.66	1.25	0.16	0.70			
	0.124	42.3%	6.54	2.00	-0.94	0.67			

Figure 5a shows the inferred latency distributions for the 3-state model in Table 4, and Figure 5b shows the inferred evidence distributions for the three states. The figures also illustrate how the inferred distributions combine to match the observed distributions. Just as the self- reports are only partially diagnostic, the 3-state model implies both classifier evidence and latency are also only partially diagnostic. This can be seen in the overlap of the inferred distributions where some retrieved items have longer times than some compute items, and some retrieved items have less evidence than some compute items.

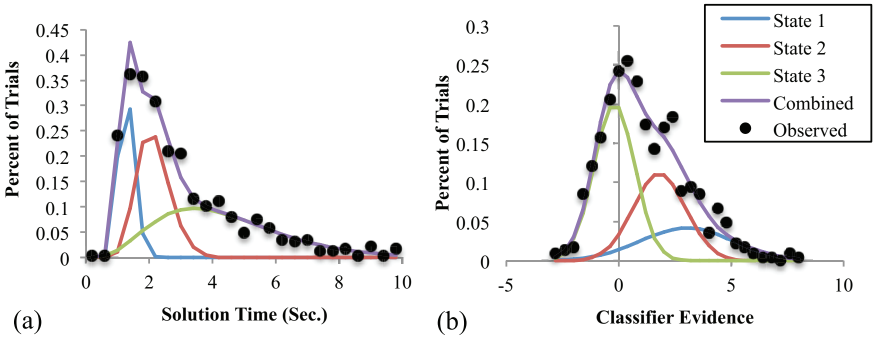

Figure 5. Figure 5a (Left) shows inferred latency distributions for the 3-state model in Table 4 and Figure 5b (Right) shows the inferred evidence distributions.

3.7 Investigating the effect of including untrained data in the model

We next included the untrained data into our data set to check the validity of our 3- state predictions. While trained data contains a range in strategy use, we know that untrained problems are unfamiliar to the participants and are solved using procedural strategies. Since we only collected reports on 6 untrained problems per participant, we eliminated two-thirds of our untrained problem set from the data used by our mixture model. The mixture model discussed in section 3.6, when run on a combination of trained and untrained data best fit four states. Table 5a shows the average value of the evidence scores, latency, and retrospective reports given for each of the four states. For every problem, the model outputs the probability of belonging to each of the states. To achieve a single state assignment for every problem, we assigned each problem to the state category that the model indicated had highest probability of representing that problem. Table 5b compares how the trained and untrained problems assigned to the four states, best fit by the second model, relate to the trained problems assigned to the three states, best fit by the first model (Table 4). It is apparent that the model of all the data has recovered the original three states of the trained data model and then added a fourth state mostly occupied by untrained problems. Comparing the parameters for the two models also indicates that the first three states of the model of all the data (Table 5) are found in the three states of the trained data model (Table 4). The largest deviation from perfect correspondence between the models is the 60 trained problems assigned to states 3 by the trained data model that are now classified in State four. The fourth state emerges with descriptors for the most extreme cases of calculation, those with the longest latency, lowest reports of retrieval and highest evidence of calculation.

Table 5

(a) Parameter estimates for the four state model fit on trained and untrained data.

			Latency (Gamma)		Evidence (Normal)	
4 States	Proportion of Trials	Report Probability	Mean	SD	Mean	SD
1	0.25	86.6%	1.4	0.3	2.6	2.11
2	0.23	65.4%	2.4	0.52	1.6	0.95
3	0.29	37.7%	4.08	1.4	-0.01	0.6
4	0.23	13.0%	6.76	3.14	-1.28	0.74

(b) Number of problems in each of the combination of 3 states (Table 4) and each of the 4 states (Table 5a).

		4 States All Problems			
		1	2	3	4
3 States of Trained Problems	1	173	5	0	0
	2	2	144	24	0
	3	1	0	157	60
Untrained		2	3	37	95

Table 5b examines the distribution of the different states among the trained and untrained problems. As we would expect, untrained problems are for the most part predicted to be in either the third or fourth state. Trained problems, on the other hand, consist of a mixture of procedural and retrieval strategies. In Table 5a we see that the dominant feature in state 1 is rapid response – almost all the items in this category show latencies under 2 seconds, matching the speed that participants showed at the end of training (Figure 1). At the other extreme, State four is dominated by evidence having virtually no case where the evidence indicates retrieval. States two and three, on the other hand, represent more ambiguous cases of retrieval and calculation.

3.8 Contribution of the three sources to the mixture model

The previous results confirm the model’s ability to distinguish strategic states. To assess the contribution of the different measures on the predictions made by the model we investigate how well each measure alone could predict the values generated by the mixture model. We used the output of multi-source mixture model run on trained and untrained data to assign each trial to one of four states. We ran three multinomial logistic regressions; one on each data source separately to explore how well each explained the full model's four state predictions. The BIC scores for the models indicated that latency best fit the data (BIC 838.56) followed by the evidence scores (BIC 1,161.1) and lastly by retrospective reports (BIC 1,714.4). While no measure alone fit the 4-state model, this analysis tells us how well each measure could predict four states.

We next looked at the ability of the various combinations of sources to predict the state solution we obtained when using all sources. Using the original parameters we estimated from all sources we then reran the model with different combinations of 0, 1, 2, or 3 sources of information. Assigning each problem to the most likely state we looked at the match between the states discovered and the state solution obtained with all four sources. Table 6 shows the results. While all sources contribute to the state identification, it is clear that the retrospective reports contribute less than either evidence scores or latency.

Table6
The percentage of the total mixture model accounted for by using different combinations of the three measures

		Evidence Scores Used	
Latency Used	Reports Used	No	Yes
No	No	31.0%	72.7%
	Yes	41.8%	74.8%
Yes	No	76.1%	92.3%
	Yes	81.4%	100.0%

4 Discussion

The current study investigated the effect of training on problem solving strategy. We studied two different problem types, Pyramid and Formula problems, each of which employed unique operations. Participants received training on 3 Pyramid and 3 Formula problems prior to the scan. During the scan session, participants solved a mix of trained problems and untrained problems that had not been seen before. The effect of training suggests that while learning did take place, it was inconsistent among trained problems. As a result, we were challenged to explore a method for predicting strategy using problem solving latency, retrospective strategy reports and fMRI data.

4.1 Effect of Training

This study used a training paradigm in order to influence the problem solving strategies employed by participants. Consistent with prior research (Delazer et al., 2005; Imbo & Vandierendonck, 2008; Ischebeck et al., 2006), trained problems had decreased problem solving latency and increased accuracy when compared to untrained problems. When trained problems were randomly mixed with untrained problems during the scan task, there was an increase in latency and decrease in accuracy; however, trained problems continued to be more quickly and accurately solved than untrained problems. After the scan task, participants reported using procedural strategies in about 40% of trained trials. Previous training studies showed greater reductions in latency and lower reports of procedural strategy use than our study (Delazer et al., 2003, Delazer et al., 2005). We trained participants in a single session, whereas other studies used 5 or more sessions to train participants (Delazer et al., 2003, Delazer et al., 2005). We chose not to ‘over train’ our trained problems in order to create problem sets that varied in familiarity. Even after many years of schooling, participants have been found to solve common arithmetic problems like 8+6 or 8*6 using mixtures of retrieval and procedural strategies

(Campbell and Timm, 2000; Hecht, 2006). Our goal in this study was to build a method for assessing strategy in the scanner that was robust enough to study math problem solving without requiring that participants be trained to such a degree that their distribution of strategies used becomes artificial.

The distinction between trained and untrained problems was also identified in our ROI analysis of the HIPS, a region that contributes to the classifier and is highly implicated in arithmetic and counting (Dehaene et al., 2003). In agreement with other training studies (Delazer et al., 2003; Ischebeck et al., 2006), HIPS showed a significant main effect of training with no difference between the two problem types. While a classifier using only the HIPS was not reliable, a classifier using a larger network of regions including the HIPS was successful in distinguishing between trained and untrained problems. It is worth noting that neither the HIPS nor angular gyrus was identified in our exploratory analysis contrasting trained and untrained problems (Appendix A). First, these regions, though an important in distinguishing retrieval and calculation, are two of many regions that provide information useful in making this distinction (Arsalidou and Taylor, 2011; Dehaene et al., 2003; Grabner et al., 2009). Second, retrospective reports indicate that the trained problems were solved by a mixture of strategies, it is unclear the percentage of retrieval problems that would need to be present in this sample in order for us to find robust training effects in the exploratory analysis.

We compared classifiers that were trained and tested on the same operation or trained on one operation and tested on the other. There was no significant difference in precision suggesting that the effects of training are the same for the two problem types.

This finding indicates that the distinction between trained/untrained used by the classifier is independent from problem type.

4.2 Convergent validity of strategy assessment

This study found that the three methods of assessing strategy use (in-scanner latency data, strategy assessment reports, and classifier predictions of retrieval) are related to one another. The relationship between reports and latency is in agreement with other studies (Grabner & DeSmedt, 2011; LeFevre et al, 1996a) that show alignment between concurrent strategy assessment and supporting latency data. Also in line with previous research, this study finds participants are significantly slower to solve problems when using procedural strategies than when using retrieval (Goldman, Mertz, & Pellegrino, 1989; Grabner et al. 2011; Imbo & Vandierendonck, 2008). Additionally, there was a significant relationship between latency data and the evidence scores generated by the classifier. The trained problems showed substantially increased latency in the scanner (as compared to the training task) and many had latencies overlapping with untrained problems. Similarly, the classifier found overlapping distributions of evidence (Figure 4). Finally, we found a relationship between strategy assessment reports and classifier generated evidence scores. Previously, both concurrent and RSAs have been verified by observing established neural patterns of different strategies that reflect the reported use of strategies (Grabner & De Smedt, 2011; Grabner et al. 2009).

4.3 Predicting Strategy Use from three data sources

While we found similarities among the three measures, it is clear no single method offered a perfect measure for assessing strategy. This raised the question of how these three measures might be combined to create a more accurate predictor of strategy use. We combined classifier-generated evidence scores, the verbal reports of retrospective strategy use, and the in-scanner latency data to detect strategy use by identifying the best mixture model. The BIC values indicated that within the trained problem set the 3-state model did the best job fitting the data. To provide a convergent test, we ran the mixture model on a combination of untrained and trained problems and found the model best fit four states. Three of the four states fit by this model corresponded closely to the three states fit by the model of only trained problems. The additional fourth state mainly is composed of the untrained problems. The addition of untrained data allows the model to separate the calculation state into what appears to be fast and slow calculations. The fact that untrained problems are predominately represented by these states is evidence for the reliability of the differentiations made by the model.

Examining the distribution of the different latencies, fMRI evidence and retrospective reports in the different state categories give some insight into what the states represent. The three states of the mixture model run only on the trained data fall into two retrieval-like states and one calculation state. Of the two retrieval states, one encompasses the more obvious cases of retrial whereas the other contains problems with slightly longer latencies, lower evidence scores and fewer reports of retrieval. This second retrieval strategy state may reflect problems that involve a strong ‘feeling of knowing’ yet require additional work to recall the solution (Reder and Ritter, 1992). The Delazer et al. (2003) math training study reported that, despite 5 days of training, some participants used partial retrieval strategies as well as full retrieval strategies. While that study was unable to test that hypothesis, it seems possible that our model is disentangling the distinction between fully or partially retrieved math facts.

4.4 Contribution of the three data sources in predicting strategy

Having identified that the models appear to be detecting separate groups of strategies, we can then assess the contribution of the three measures in the model. We looked at how well each measure by itself predicts the four states predicted by the mixture model of trained and untrained data. A comparison of the BIC scores of the three logistic regressions of the measures indicates latency scores best predicted the four states, followed by evidence scores and finally retrospective reports. Next, we analyzed the ability of different combinations of measures to generate the predictions of the 4-state mixture model. The performance of each data source in predicting the output of the full mixture model shows that latency is only about 4% more accurate than the imaging classifier evidence (see Table 6). When the two measures are combined together there is a 16% boost in accuracy, and we account for 92% of the full model's prediction. Retrospective reports contribute very little to the ability to distinguish strategy, thus supporting claims that retrospective reports are generally inaccurate (Russo et al., 1989). Taken together these tests suggest that latency provides the best source of data when distinguishing strategies: however, it is apparent that without the addition of the other data sources, the model is restricted in its ability to detect distinctions between strategies.

4.5 Conclusion

In this study we used selective training in order to generate sets of problems likely to be solved by either procedural (untrained) or retrieval (trained) strategies. While training did influence the speed, accuracy, strategy use, and brain activation patterns, participants continued to use procedural strategies to solve some of the trained problems. We used the distinction between trained and untrained problem sets to train a classifier. When we applied this classifier to the trained problem set, it appeared that the distinction identified was that between different problem solving strategies. The classification of fMRI data generated evidence scores that were significantly correlated with latency data from the scan task and aligned with the retrospective reports. Using the measures of brain activity, self-report and behavioral data, we discovered a 3-state model that gives the best account of the trained problems and a 4-state model that gives the best account of all problems. These states appear to represent distinct problem solving strategies. By combining activation data, latency measures, and RSA, a clearer picture of strategy use emerges.

5.1 Appendix A

Exploratory analysis – The effects of training

The results of the whole-brain contrast between trained and untrained problems are listed in the below table. This analysis combines problem types, and focuses on the general distinction of training. We found significant activation clusters (p<.005), by using an extent threshold 20 voxels and running a simulation to find a brainwise alpha .05 (Cox, 1996).

Significant activation clusters (brainwise p<.05) from the contrast of trained-untrained problems across both problem types.

Brain area	X	Y	Z	voxels	t-value
L Cuneus	-2	-67	30	719	4.38
L Middle Occipital G	-26	-93	17	328	4.79
L Precentral G	-33	-20	65	299	3.58
R Middle Occipital G	32	-91	7	164	5.98
R Fusiform G	32	-60	-9	155	3.81
L Postcentral G	-57	-15	16	82	3.19

Coordinates reported in Talairach coordinate space given by AFNI. Anatomical labels based on the the Talairach-Tournoux atlas (Talairach and Tournoux, 1988). The label represents the location of peak activation. Significant activation clusters (p<.005, extent threshold 20 voxels, ran a simulation to find a brainwise alpha .05) (Cox, 1996) Abbreviations: L, left hemisphere; R, right hemisphere; G, Gyrus.

5.2 Appendix B

Imaging Data – The Effects of Training in Angular Gyrus

We used the Dehaene et al. (2003) specification of the left angular gyrus (AG) as located at -41, -66, 36 in Talairach coordinates. We ran a problem type (Pyramid, Formula) X training (trained, untrained) repeated measures ANOVA on the left AG. There were no significant main effects. There was a moderately significant operation-by-training interaction, *F* (1, 19)=5.88, *p*=0.03, with Pyramid problems showing a slight increase in AG activation (.014 to .09%) with training and Formula problems showing a very slight decrease in activation (.031 to .026%). As these percentages indicate, the absolute level of AG activation was quite low. This interaction may be due to differences between these operation types in the strategies used to solve trained and untrained problems. The percentage reports of retrieval for trained Pyramid problems is 7.7% higher than those of trained Formula problems.

Acknowledgements

This work was supported by the National Science Foundation grant DRL-1007945 . We would like to thank Bob Siegler for his comments on this manuscript.

References

Aitkin, M., & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67-75.

Alibali, M. W. (1999). How children change their minds: strategy change can be gradual or abrupt. *Developmental Psychology*, *35*(1), 127–145.

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

Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. (2010). Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences*, *107*(15), 7018–23.

Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. (2011). Cognitive and Metacognitive activity in mathematical problem solving: prefrontal and parietal patterns. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(1), 52–67.

Arsalidou, M., & Taylor, M. J. (2011). Is 2+ 2= 4? Meta-analyses of brain areas needed for numbers and calculations. *Neuroimage*, *54*(3), 2382-2393.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Baroody, A. (1999). The roles of estimation and the commutativity principle in the development of third graders' mental multiplication. *Journal of Experimental Child Psychology*, *74*(3), 157–93.

Birn, R. M., Smith, M. A., Jones, T. B., & Bandettini, P. A. (2008). The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage*, *40*(2), 644-654.

Campbell, J. I. (1987). Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 109.

Campbell, J. I., & Timm, J. C. (2000). Adults' strategy choices for simple addition: effects of retrieval interference. *Psychonomic Bulletin and Review*, *7*(4), 692–9.

Campbell, J. I. D., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General*, *130*(2), 299–315.

Cho, S., Ryali, S., Geary, D. C., & Menon, V. (2012). How does a child solve 7 + 8? Decoding brain activity patterns associated with counting and retrieval strategies. *Developmental Science*, *14*(5), 989–1001.

Cohen Kadosh, L., Kadosh, R., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental, and comparative studies of magnitude representation. *Progress in Neurobiology*, *84*, 132–147.

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory and Cognition*, *19*(2), 151–8.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162-173.

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

Damarla, S. R., & Just, M. A. (2012). Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping*.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

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

Delazer, M., Domahs, F., Bartha, L., Brenneis, C., Lochy, A., Trieb, T., & Benke, T. (2003). Learning complex arithmetic—an fMRI study. *Cognitive Brain Research*, *18*(1), 76–88.

Delazer, M., Ischebeck, A., Domahs, F., Zamarian, L., Koppelstaetter, F., Siedentopf, C. M., Kaufmann, L., Benke, T., & Felber, S. (2005). Learning by strategies and learning by drill--evidence from an fMRI study. *NeuroImage*, *25*(3), 838–49.

Ford, J., Farid, H., Makedon, F., Flashman, L., McAllister, T., Megalooikonomou, V., & Saykin, A. (2003). Patient classification of fMRI activation maps. *Medical Image Computing and Computer-Assisted Intervention*, 58-65.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: characterizing differential responses. *NeuroImage*, *7*(1), 30-40.

Geary, D. C., & Brown, S. C. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in gifted, normal, and mathematically disabled children. *Developmental Psychology*, *27*(3), 398–406.

Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology*, *27*(5), 787–797.

Goldman, S. R., Mertz, D. L., & Pellegrino, J. W. (1989). Individual differences in extended practice functions and solution strategies for basic addition facts. *Journal of Educational Psychology*, *81*(4), 481–496.

Goldstein, H., & Browne, W. (2002). Multilevel Factor Analysis Modeling Using Markov Chain Monte Carlo Estimation. In G.A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (225). Mahwah, NJ: Psychology Press.

Grabner, R. H., & De Smedt, B. (2011). Neurophysiological evidence for the validity of verbal strategy reports in mental arithmetic. *Biological Psychology*, *87*(1), 128–36.

Grabner, R. H., Ischebeck, A., Reishofer, G., Koschutnig, K., Delazer, M., Ebner, F., & Neuper, C. (2009). Fact learning in complex arithmetic and figural-spatial tasks: the role of the angular gyrus and its relation to mathematical competence. *Human Brain Mapping*, *30*(9), 2936–52.

Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *Neuroimage*,*43*(3), 509-520.

Hecht, S. A. (2006). Group differences in adult simple arithmetic: good retrievers, not-so-good retrievers, and perfectionists. *Memory and Cognition*, *34*(1), 207–16.

Imbo, I., & Vandierendonck, A. (2007). The development of strategy use in elementary school children: working memory and individual differences. *Journal of Experimental Child Psychology*, *96*(4), 284–309.

Imbo, I., & Vandierendonck, A. (2008). Practice effects on strategy selection and strategy efficiency in simple mental arithmetic. *Psychological Research*, *72*(5), 528–41.

Ischebeck, A., Zamarian, L., Egger, K., Schocke, M., & Delazer, M. (2007). Imaging early practice effects in arithmetic. *NeuroImage*, *36*(3), 993–1003.

Ischebeck, A., Zamarian, L., Siedentopf, C., Koppelstätter, F., Benke, T., Felber, S., & Delazer, M. (2006). How specifically do we learn? Imaging the learning of multiplication and subtraction. *NeuroImage*, *30*(4), 1365–75.

Kettaneh, N., Berglund, A., & Wold, S. (2005). PCA and PLS with very large data sets. *Computational Statistics & Data Analysis*, *48*(1), 69-85.

Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 157–175.

Kadosh, C. R., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, *84*(2), 132–47.

LeFevre, J.-A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996a). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*(3), 284–306.

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996b). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216–230.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 179.

Marchini, J. L., & Ripley, B. D. (2000). A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, *12*(4), 366-380.

Núñez-Peña, M. I., Cortiñas, M., & Escera, C. (2006). Problem size effect and processing strategies in mental arithmetic. *NeuroReport*, *17*(4), 357–60.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. NeuroImage, 45(1), S199–209.

Reder, L. M. (1988). Strategic control of retrieval strategies. *The psychology of learning and motivation*, *22*, 227-259.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 435-451.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*(6), 759–69.

Siegler, R. S. (1988). Strategy Choice Procedures and the Development of Multiplication Skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. *Developing cognitive competence: New approaches to process modeling*, 31-76.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do. *Origins of Cognitive Skills*, 229-293.

Shinkareva, S. V, Mason, R. a, Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using FMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS one*, *3*(1), e1394.

Thomaz, C. E., Boardman, J. P., Hill, D. L., Hajnal, J. V., Edwards, D. D., Rutherford, M. A., & Rueckert, D. (2004). Using a maximum uncertainty lda-based approach to classify and analyse mr brain images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004* (pp. 291-300). Springer Berlin Heidelberg.

Wickens, T. D. (2001) *Elementary Signal Detection Theory*, OUP USA.