Detecting Math Strategy Use During Learning

Caitlin Tenison (ctenison@andrew.cmu.edu)
Department of Psychology
Pittsburgh, PA 15206 USA

John R. Anderson (ja0s@andrew.cmu.edu)
Department of Psychology
Pittsburgh, PA 15206 USA

Abstract

The ability to accurately assess math problem solving strategy is an important part of understanding the effects of practice. Unfortunately the measures researchers trust are often unreliable and ill suited for studying the effects of practice. In the current study we are interested in identifying intermediary strategies that emerge as people switch from computational to retrieval strategies. To build a more accurate assessment of strategy we combine latency, neural evidence, and verbal reports using a mixture model. We compare the model's predictions of strategy use with concurrent assessments collected during the problem solving. The results suggest that while participants consider a partial computation-retrieval strategy, distinct from pure computation, our model finds no evidence of such a partial state; however, distinction is found between early and well-practiced retrieval. These results suggest a discrepancy between the distinctions people make when reporting strategy use and the distinctions in the cognitive processes underlying strategy use.

Keywords: fMRI; Mixture Model; Problem Solving; Strategy use.

Introduction

Tools to assess math strategy use are critical to the study of math problem solving. A researcher can glean only so much from knowing a person's solution to a task because it provides little information about the processes that were used to arrive at the solution. Take, for example, the problem of adding up all the numbers from 1 to 100. The solver could mentally keep a running total, adding each number, or creating a formula to arrive at the answer (i.e., 100*(100+1)/2). The strategy a student uses to solve a math problem reflects a valuable measure of their understanding of the mathematical concepts underlying the problem.

As students gain practice working with problems, the strategies they use to solve the problems change. Practice often causes participants to switch from strategies that use calculations (referred to here as computational strategies) to strategies that involve recall of previously learned facts (referred to here as retrieval strategies) (Imbo & Vandierendonck, 2008; Ischebeck et al., 2007). According to the adaptive strategy choice model, the shift to retrieval strategies arises out of an increased association between the math problem and the solution such that a participant can retrieve the answer from memory (Siegler & Shipley, 1995). Work studying children learning arithmetic suggests that strategies emerge and/or decline in use through a mix of metacognitive strategy discovery and associative mechanisms of gradual learning (Shrager and Siegler, 1998). This idea, summarized by Siegler's 'overlapping waves theory', describes the gradual changes in childrens' strategy use over time from less efficient to more efficient strategies. The changes in strategy use are an important feature for understanding learning, and consequently the ability to accurately assess these changes is necessary for the study of math learning.

The different methods for assessing strategy use have tradeoffs, when being used to assess a dynamic learning task. Assessing strategy use becomes especially difficult when studying math learning in the fMRI scanner. A verbal protocol in the context of an fMRI study cannot be collected without impacting the quality of the data. Speaking modulates breathing, which in turn has been shown to have an effect on the blood-oxygen-level-dependent (BOLD) response (Birn, Smith, Jones, & Bandettini, 2008). A number of experiments have explored means to simplify concurrent verbal assessment to reduce its reactivity. For instance, in several studies participants were provided with a list of strategies after each problem and encouraged to choose the option that best represented the strategy that they used (Campbell & Timm, 2000; Grabner et al., 2011; Imbo & Vandierendonck, 2008). In support of the effectiveness of this technique, Grabner et al. (2011) found similar brain responses for items reported to be solved with the same strategy. This method of concurrent assessment, however, has two flaws. First, suggesting alternative strategies may alter the participant's problem-solving methodology, and second, a participant is forced to choose among the provided strategies, which may not include the specific method used in problem solving. These two flaws can be avoided by use of a retrospective strategy assessment.

Retrospective strategy assessments (RSAs) are a less reactive form of strategy assessment, but are also less accurate (Russo, Johnson, & Stephens, 1989). During retrospective strategy assessments, researchers ask the participant to report strategy use after the entire task has been completed, often with a list of problems to help cue memory (Grabner et al., 2009). The advantage of RSAs is that task data remain unaffected by the assessment of strategy. Additionally, the RSA allows for a more detailed report on specific strategy than concurrent assessments. Nevertheless, this form of assessment is ill suited for dynamic learning tasks in which solution strategies change

with practice because they only provide information about the final strategy state of a problem. Thus, while RSAs do not provide item-by-item information, they can still be useful for identifying general strategy patterns.

Both concurrent and RSAs assume that participants are fully aware and able to describe the strategies they use. To validate participant reports, researchers can use other indicators of the cognitive processes underlying strategy use. Problem solving time and fMRI data can be applied as an indirect method for assessing strategy use. Numerous studies that compared computational and retrieval strategies found that the time it takes to solve a problem significantly differs based on the strategy used, with retrieval taking much less time than computation (Campbell & Timm, 2000). Reaction time data—while a robust predictor of computational versus retrieval strategies—has limited sensitivity when used to detect intermediate or mixed strategies. The amount of time it takes to execute strategies changes with practice, so distinguishing between a fast computation and a slow retrieval using latency is difficult (Delaney, Reder, Staszewski, & Ritter, 1998).

Using fMRI to measure the brain's representation of these different strategies offers an indirect means of assessment as well. fMRI studies of math training have found distinctions in neural responses to highly practiced problems (problems thought to be solved by retrieval) and problems that are novel. Several math-training studies report that novel problems—in contrast to practiced problems—activate the frontal-parietal network where as the reverse contrast increase activity in the angular gyrus (Arsalidou & Taylor, 2011; Delazer et al., 2005; Ishebeck et al., 2006). In cases where the participant may not be aware of gradual changes in strategy, latency and fMRI measures may provide insight into these changes.

In a previous study, we combined different sources of data to increase the accuracy of the assessment of strategy use within the fMRI scanner (Tenison, Fincham & Anderson, 2014). With this method we identified 4 classes of strategies used to solve practiced and novel math problems; we built this model using fMRI data, problem-solving latency, and retrospective strategy assessment data (Tenison et al., 2014). Based on RSAs and latency data, we proposed that these states might indicate that participants used a very slow computational strategy, a very fast retrieval, or two intermediary strategies (one fast computational, and one slow retrieval). The retrospective reports from this study suggested that when participants practice solving problems, these intermediary strategies emerge from computation and are used until the participant is confident enough to retrieve the answer. These reports seem to suggest that adults experience overlapping waves of strategy use transitioning to an intermediary strategy before shifting to retrieval strategies.

The present study aims to test this model and shed light on these intermediary strategies. In the current experiment, we scan participants while they learn math problems to gain a more accurate picture of how intermediary strategies emerge and decline with practice. The current study uses a multiple-choice concurrent strategy assessment to act as a measure of 'ground truth' to test our model against. We hypothesized that the model will distinguish four strategies, a finding that is in alignment with previous research (Tenison et al., 2014). Furthermore, we predicted that participants would switch from using computational strategies, to intermediate strategies, to retrieval strategies as they gain practice. We used the concurrent reports to explore the sensitivity with which this assessment method can detect strategies used to solve these problems.

Methods

Participants

Twenty university students (9 females; mean age 22/ SD 2.3) participated in the study. Participants gave informed written consent and received monetary compensation for their participation. All participants were right handed. The university ethics board approved the study.

Stimuli and experimental design

To investigate the change in strategy that occurs when learning a new type of operation, we trained participants on a novel operation. This operation, called a 'Pyramid problem', uses the same algorithm as that in the prior experiment by Tenison et al. (2014). To solve these problems participants must keep a running total in their head as they add together several integers. For example, 11$4 would be expanded to 11+10+9+8, and thus, the correct value is 38. We controlled for difficulty between experimental conditions.

We used two assessments of strategy use, a concurrent and a retrospective assessment. The concurrent assessment presented a list of strategies from which participants were encouraged to choose the strategy that best matched the one used to solve the previous problem. We compiled the strategy options by considering the frequency of reported use in a previous experiment using the same problems (Tenison et al. 2014) The instructions at the start of each scan included definitions of the different available strategy choices. "Retrieve" was defined as remembering the answer; "calculate" was defined as using arithmetic to find the answer; "partial" was described as partially calculating and partially remembering the Pyramid problem. Participants were instructed to indicate if they used an "other" strategy for any strategy that did not fit within the first 3 categories. The concurrent assessment was presented on the screen after participants finished entering their answer to the Pyramid problem. Participants were asked, "How did you solve the problem?" and were given the choices of "1) Retrieve 2) Calculate 3) Partial 4) Other". Only one participant indicated use of the "other" strategy, information from the retrospective report indicated that an abbreviated computation strategy had been used. For our later analysis, we recoded this as the "calculate" strategy

since the strategy described was similar to other participants' reports of computation.

Following the completion of the scan, participants completed a retrospective strategy assessment in which they solved 15 paper-based problems that included all the practiced and some of the novel problems. After solving each problem participants were instructed to write down a detailed explanation of how the problem was solved. The participants' strategy self-reports were coded based on three categories: retrieval, calculated, and partial.

The experiment used a numeric keypad with the number arranged in a standard keyboard format. Participants used the keypad to type out the answers to the math problems and to indicate the problem solving strategies that were used.

Scanning Procedure

Participants completed 6 fMRI scans. Participants were exposed to a set of highly practiced problems (We will refer to these as *practiced problems*) and set of limited practice problems (We will refer to these as *novel problems*). This allowed us to contrast the difference in strategy use between the two sets. In total, the experiment featured 3 practiced problems, 18 novel problems and 6 warm up problems. Each scan began with a warm up problem of which the response was discarded and then a random mix of 6 novel problems and 18 practiced problems (3 problems with each problem repeated 6 times). Novel problems matched the practiced problems in difficulty. During the experiment, participants saw the novel problems twice over the course of 6 scans (thus the novel problem was repeated when participants had solved over 50 problems and were unlikely to remember having seen the problem before). Participants also completed a concurrent strategy assessment after each problem during the 2nd, 4th, and 6th scans. Participants did not complete a concurrent assessment on the 1st, 3rd and 5th scans. The alternating of scans featuring or not featuring an assessment allowed the experimenter to check the reactivity of the assessment (no reaction was found).

Pyramid problems were presented on the screen following a 2 second fixation period. Once the problem appeared on the screen, the participant was allowed a maximum of 30 seconds to indicate knowledge of a solution by pressing the return key on the numeric keypad. After pressing 'return', participants had 5 seconds to input a solution using the keypad and press the return key. After answering the problem, the participant was given correctness feedback and information about how the problem should have been solved. If it was the 2nd, 4th, or 6th scan, a screen appeared that asked participants, "How did you solve the problem?" Participants were given 5 seconds to select the number that best corresponded to the strategy used. At the end of each problem solving trial, a 12 second 1-back task was presented onscreen to prevent metacognitive reflection on the previous problem and allow the hemodynamic response of the brain to return to baseline. Problem solving time was defined as the time between the appearance of the math problem and the point at which the participant indicated a readiness to input the answer.

MRI data acquisition

Images were acquired using gradient echo-echo planar image acquisition on a Siemens 3T Verio Scanner using a 32 channel RF head coil, with 2 s. repetition time (TR), 30 ms. echo time, 79° flip angle, and 20 cm. field of view. 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 mm2. The anterior commissure-posterior commissure line was on the 11th slice from the bottom scan slice. Acquired images were pre-processed and analyzed using AFNI (Cox, 1996). 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.

fMRI Analysis

To create a single measure of strategy use from the fMRI data we used a classification analysis to quantify how similar a given trial was to other retrieval trials. Without a direct report of retrieval, we trained our classifier on the distinction between practiced and novel problems, since we knew novel problem could not be solved by retrieval, whereas most practiced problems would be solved by retrieval. For the purposes of this paper, we will summarize the processing steps applied to our data, explicit justification of our actions are reported in the Tenison et al. (2014). 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' regions of interest (ROIs) (Anderson, Betts, Ferris, & Fincham, 2010). 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 288, 4x4x4 voxel regions of raw data. 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 288 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[1]. A GLM was used to estimate the beta values 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 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, we performed dimensionality reduction using Principle Components Analysis (PCA), which creates a set of uncorrelated variables from linear combinations of the ROI activity. The PCA was performed on the z-scores of the beta values for the 288 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. We then preformed a linear discriminate analysis (LDA) on the first 50 factors extracted from the PCA. We used the LDA to identify which of these factors contributed to distinguishing between practiced and novel problems. Because we were interested in identifying 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 analysis.

Results

Effects of Practice

Practicing a problem had clear effects on the speed and strategy with which participants solved the pyramid problems. A repeated measures ANOVA indicates that the time to solve practiced problems decreased, but the time to solve novel problems remained constant. The analysis revealed a significant main effect of problem group, $F(1,18)=69.28$, $p<0.0001$, scan block, $F(5,90)=18.66$, $p<0.0001$, and a significant problem by scan block interaction, $F(5,90)=14.95$, $p<0.0001$. There were corresponding changes in the strategy use reported for the practiced problems. There was an increase in reports of retrieval over the three blocks on which concurrent reports were obtained, F $(2,38)=42.04$, $p<0.0001$, a decrease in reports of both computation, F $(2,38)=8.396$, $p=0.001$, and partial strategies, F $(2,38)=18.598$, $p<0.0001$. On novel problems, participants did not indicate any significant changes in their use of strategy during the experiment. Retrospective reports echo the overall differences in strategy use between practiced and novel problems reported in the concurrent assessments. Table 1 shows the mean percentages of strategy use reported for both problem types on the concurrent and retrospective assessment.

Table 1: Percent Strategy Use Reported

Strategy	Concurrent		Retrospective	
	Practice	Novel	Practice	Novel
Retrieval	81.7	1.3	76.4	4.8
Computation	6.8	89	16.4	85.6
Partial	11.4	9.6	7.4	9.6

Classifier Performance and Output

As described in the methods, we trained a classifier to distinguish trained trials from untrained trials. We used leave-one-subject-out classification technique in which we trained the classifier on the distinction for all participants on one and then tested its ability to predict trials for the remaining participant. The classifier was highly robust in the cross subjects tests, predicting all subjects better than chance. The average d-prime measure of performance for a particular subject in this analysis was 1.71, $t(19) = 14.5$, $p<0.001$, with a hit rate of 60% and a false alarm rate of 11%. Mapping the weights from the classifier back to the brain we can observe areas associated with computation and retrieval used in this classification (Figure 1). For purposes of further use in this paper, the major contribution of this classifier is that it labels each trial with the probability that it was trained. We will use this evidence score as one source of information about the strategy used to solve a problem.

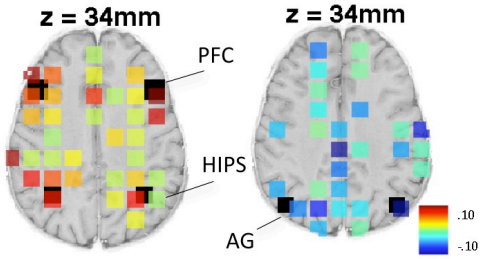

Figure 1: Warm voxels are more active for untrained problems, cool voxels are more active for trained problems. Locus of the prefrontal cortex (PFC), horizontal intraparaital sulcus (HIPS) are marked on the left graph, angular gyrus (AG) is marked on the right. The z-value is for x=y=0 in Talairach coordinates.

Modeling Strategy Use

We employed a type of mixture model called a 'location model' that used three measures of strategy use (latency, fMRI evidences scores, and retrospective reports) in order to predict the strategic state representative for each problem. This model identifies the hidden states that are associated probabilistically with the three observable measures. We use expectation maximization to fit the model to these three measures to identify the hidden process states (Aitkin & Rubin, 1985; Bailey & Elkna, 1994).

Applying this model to the data best fit 6 states, however, a detailed investigation of these states indicated that the

[1] Assuming the standard SPM hemodynamic response – Friston et al. (1998), the difference between two gamma functions used was gamma (6,1) – gamma (16,1)

model was separating states to capture an underlying correlation between the evidence scores and latency (log latency is correlated .71 with the evidence scores). The mixture model treats these two dimensions as independent and to capture the correlation it was creating various states along the latency-evidence continuum. Therefore, we decided to combine and orthogonalize these two measures by use of a PCA. The first component of the PCA proved to carry all the information accounting for 88.7% of the variance. This first component can be taken as a general "strength" measure and then used it, in combination with the retrospective reports, to train the mixture.

Rather than fitting the 4 states fit in Tenison et al. (2014), we best fit 3 states. Our mixture model assigns to each trial a probability that it is in one of the 3 possible states. We assigned each trial to the highest probability state (the mean probability of these states are .73). Using this classification, Table 2 shows the mean latencies, evidence scores, and percent of problems with a retrospective report of retrieval for the three resulting states. In addition, Table 2 reports the proportion of problems in each state that were practiced. States 1 and 2 are almost exclusively practiced trials with the major difference being slightly slower latencies and higher evidence scores. Problems in States 1 and 2 are solved quickly, retrospectively reported as retrieval, and have low fMRI evidence of computation. State 3 contains a majority of untrained problems, long latencies, high evidence of computation and few retrospective reports of retrieval. It seems clear that State 1 is a retrieval state and State 3 a calculate state. The status of State 2 is somewhat ambiguous.

Table 2: Parameter estimates for the 3 State Model

	State1	State2	State3
Latency (sec)	1.3	2.5	7
Evidence	-2.1	-.19	.71
% Retrieval	94%	79.5%	14.4%
% Practiced	99.8%	96.5%	39%

A major interest of this experiment was to use the concurrent reports to check the validity of the state assignments obtained with the non-obtrusive measures of latency, brain evidence, and retrospective reports. We have these concurrent reports for problems solved on the 2nd, 4th, and 6th block. Table 3 shows the state assignments of the problems that were assessed.

Table 3: State Assignments for Concurrently Assessed Items.

	State1	State2	State3
Calculate	1.6%	8.5%	67.5%
Partial	0.8%	13.7%	18.8%
Retrieval	97%	77.8%	13.8%

Consistent with the evidence in Table 3, participants overwhelmingly rate State 1 trials as retrieval and the other 85% of State 3 problems as calculate or partial. State 2 appears to be a slow retrieval state. The fact that almost 80% are called retrieval and less than 10% calculate suggests that this is really a retrieval state. Participants mean latencies in this state are rather slow (2.51 sec.) for a pure retrieval, suggesting some hesitancy in retrieving. Perhaps it takes them a moment to recognize that this is a problem they can retrieve. Figure 2 shows for each block the proportion of problems assigned in to each state. As participants gain practice they switch from slow retrieval to fast retrieval.

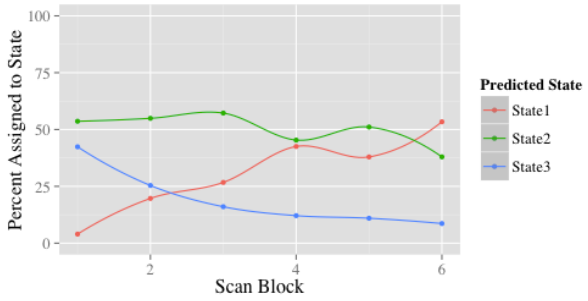

Figure 2
Percent of State Assignments for Trained Problems

Discussion

This experiment set out to gain a better understanding of the intermediary strategies that emerge as people transition from the use of computational strategies to retrieval strategies with practice. Our results indicate a distinction between early and late retrieval processes that people do not make when reporting their strategies.

Changes in problem-solving latencies and concurrent reports indicate that people increased in speed and changed their strategy use as they gained practice with problems. The retrospective report we collected echoed these results, indicating that by the end of the study, practiced problems were solved predominately by retrieval and novel problems were solved using computational strategies. Additionally, we were successfully able to apply a classifier to the fMRI data to distinguish practiced and novel problems. Among the variety of regions used to make this distinction were areas used for arithmetic computation and retrieval (Arsalidou & Taylor, 2011). Using the convergence of latency, fMRI evidence and the retrospective reports we fit a previously developed model. Immediately, we noted that by studying the active effects of learning, adjustments had to be made to our model that had been developed for a more static task. Training affected both the latency and the fMRI data; thus, we needed to run PCA to de-correlate these measures. With this adjustment our model fit 3 states: a computation state, a retrieval state, and an intermediary retrieval state. The low percentage of untrained problems in States 1 and 2 provide evidence that these states reflect the effects of training on retrieval strategy. Additionally, the gradual decline in State 2 and increase in State 1 assignments during

the task suggest that these states might reflect changes in retrieval due to practice.

Retrospective reports of a partial computation-retrieval strategy lead us to expect the model to identify such a state. Instead, the model identified a state that participants generally identified as retrieval in concurrent reports. This state contained items that took longer to solve and showed more neural similarity to computation than the fast retrieval state. It is possible that participants are executing a deliberate search of memory in order to make these retrievals. We did not gather any reports of such a strategy in the RSA and it seems unlikely that participants may even consider effortful retrieval as different from automatic retrieval. While retrieval and computation are distinct enough to recognize, people may not be aware of changes in computation and retrieval that occur when they are learning. In cases of gradual changes in strategy we suggest that measures such as problem solving latency and fMRI provide a more nuanced picture of strategy use. Future work could benefit from an ROI analysis of how math relevant regions distinguish the model-identified states. Such an analysis would be useful in further understanding the states and the cognitive processes they involve.

Acknowledgments

This work was supported by the National Science Foundation grant DRL-1007945 and by Carnegie Mellon University's Program in Interdisciplinary Education Research funded by IES grant R305B090023

References

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

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.

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.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. (Tech. Rep. CS94-351). San Diego, CA: UCSD, Department of Computer Science and Engineering.

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

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

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9(1),* 1-7.

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.

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.

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. *HBM, 30(9),* 2936–52.

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.

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. In Bower, G. (Ed.), *The psychology of learning and motivation* (227-259). San Diego: Academic Press.

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., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. Simon & G. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling.* New Jersey: Lawrence Erlbaum Associates.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9(5),* 405-410.

Tenison, C., Fincham, J. M., & Anderson, J. R. (2014). Detecting math problem solving strategies: An investigation into the use of retrospective self-reports, latency and fMRI data. *Neuropsychologia, 54,* 41-52.