A neural network model of learning mathematical equivalence

Kevin W. Mickey (kmickey@stanford.edu)
James L. McClelland (mcclelland@stanford.edu)
Department of Psychology, Stanford University

Abstract

The typical structure of equations influences how we learn the meaning of the equal sign. Previous studies have shown that as students gain experience with addition problems, they actually perform worse on certain problems, before eventually improving. We seek to explain this trajectory with gradual implicit learning, without explicit representation of strategies or principles. Our parallel distributed processing model is nevertheless able to simulate several phenomena observed in how children learn mathematical equivalence: not only how successful performance develops, but also what strategies are used and how equations are encoded.

Keywords: Cognitive development; mathematical equivalence; problem solving; neural network model.

Introduction

What kind of knowledge is required to solve mathematical problems? Explicit knowledge of abstract principles, while valuable, may be neither sufficient nor necessary for correct performance. For example, students who know explicit rules of precedence of algebraic operations may fail to apply them (Kellman & Massey, 2013), and children often discover new strategies without being able to justify them or even acknowledge their use (Siegler & Araya, 2005). Here we explore how behavior that appears in accordance with abstract principles can emerge from gradual, implicit learning. Applying previous work that has ranged from acquisition of inflectional morphology (Rumelhart & McClelland, 1989) to implicit knowledge of a balance beam (McClelland, 1989), we offer a model that learns to solve simple equations without using explicit strategies or principles of equivalence.

The ubiquity of the equal sign indicates how fundamental equivalence is. This general concept underlies knowledge in many areas, including non-mathematical tasks. Piagetian conservation tasks (Piaget, Inhelder, & Szeminska, 1960) test whether a child recognizes the invariance of a quantity (weight, length, number, etc) after some transformation. Likewise, we can transform an arithmetic expression like $5+1$ into the number 6 or another expression (e.g., $4+2$), and the quantity being represented is the same. A *relational understanding* of the equal sign captures the idea that both sides of an equation represent the same quantity, even if they appear in different forms. Before learning this, children often first develop what has been called an *operational understanding* of the equal sign (Perry, Church, & Goldin-Meadow, 1988). An expression like $3+4=\_$ may be treated as though it is a request for action, namely the addition of the numbers. This response may even be given in cases where the equal sign is not the final operand: Students may see $3+4+9=9+\_$ and answer 25. A number of studies have focused on this class of equations called equivalence problems (McNeil, 2007; McNeil & Alibali, 2004; McNeil, Fyfe, Peterson, Dunwiddie, & Brletic-Shipley, 2011; Perry, 1991; Perry et al., 1988; Rittle-Johnson & Alibali, 1999; Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011).

Students may have such strong expectations about equations that they incorrectly reconstruct an equation after viewing for five seconds (McNeil & Alibali, 2004). Reconstruction was worst when the prompt, a blank, was in the final position and the equal sign was in a nontraditional location. Students who solved such problems incorrectly by adding all the numbers tended to make conceptual errors (such as moving the equal sign to become the final operator), but did not make more errors on the numbers themselves.

It is essential for both cognitive scientists and educators to understand better how children develop the ability to solve equivalence problems. Students with a relational understanding of the equal sign perform better on algebraic problems, even after controlling for performance on standardized math tests (Knuth, Stephens, McNeil, & Alibali, 2006). Differences in algebraic understanding, in turn, predict lifelong outcomes in education and employment (National Research Council, 1998). Efforts to reform mathematics education are sometimes motivated by international comparisons, and students in countries including the United States have particular difficulty with nonroutine equivalence problems, compared to students the same age in some other countries (Li, Capraro, & Capraro, 2008).

In this paper, a simple neural network model will be used to simulate the development of children's ability to solve equivalence problems. The model treats algebraic problem solving as an acquired skill, emerging slowly from practice solving example problems. Several studies including McNeil et al. (2006) have found an overwhelming bias in the structure of equations presented in textbooks, favoring the operational notion of the equal sign, and we use a training set that is biased according to such empirical estimates. As a result, the equations that have typical structure are learned first, and on atypical problems with the blank at the end, the model adopts an add-all strategy, ignoring the location of the equal sign. The model fits the qualitative pattern observed in children's problem solving, and it also simulates the biased reconstruction of such equations seen in children who make the add-all response. The model goes beyond previous informal verbal theories and offers a mechanistic explanation of how biases in mathematical textbooks influence cognitive development.

Modeling Approach

This paper explores the utility of parallel-distributed processing (PDP) models for capturing the fascinating learning trajectories and behavioral phenomena seen in the domain of mathematical cognition, where rule-based manipulation of structured expressions is often assumed (Anderson & Lebiere, 1998) but may not always be necessary. PDP models have previously been used to explore interesting developmental trajectories in other domains, and also to capture perceptual reconstruction errors when inputs violate typical patterns. McClelland (1989) presented a model of learning about the role of weight and distance from a fulcrum in balance-scale problems, which coincidentally tap into broadly similar notions of equivalence and are often used as an instructional metaphor for teaching mathematical equivalence. The model demonstrated stage-like developmental progression seen in children and previously characterized by rules, even though the underlying learning from each new experience was gradual and continuous.

Rogers, Rakison, and McClelland (2004) discussed how parallel distributed processing principles could explain a number of U-shaped developmental phenomena in language and other domains. For instance, many but not all past-tense verbs in English end in -ed, and a PDP model (Rumelhart & McClelland, 1989) initially declines in performance due to over-regularization, before learning correct morphological patterns. In the domain of semantic knowledge, Rogers, Lambon-Ralph, et al. (2004) used a recurrent network to simulate the typicalization of atypical features of items in a delayed picture-copying task. To model the similar task of encoding and reconstructing mathematical equations (while also filling in a missing number in any of three positions), we likewise employed a fully recurrent neural network model.

Method

Table 1: Problem types, with their frequencies in parentheses. The frequencies were obtained by treating two empirical estimates as independent: the bias for more equations with operation(s) on the left than on the right (90.52%) and the bias for more prompts that ask for the sum than for either addend (60.65%).

	Operation on left		Operation on right	
Prompt 1	$\_ + b = c$	(.178)	$\_ = b + c$	(.057)
Prompt 2	$a + \_ = c$	(.178)	$a = \_ + c$	(.019)
Prompt 3	$a + b = \_$	(.549)	$a = b + \_$	(.019)

Training Set

The training set consisted of 330 addition equivalence problems, presented in random order in each epoch. Each problem consisted of an expression containing a slot for each of two addends and a slot for their sum; two of the three slots were filled with a single-digit number, and the third slot was left blank, serving as a prompt for completing a valid mathematical statement. The digits ranged from 0–9; because the sum was constrained to a single digit, there were 55 different problems containing 30 unique pairs of addends if their order is ignored. Each problem was presented with the two addends either on the left or the right side of the equal sign. Combining this with three positions for the prompt, there were six syntactic types (shown in Table 1) of each of the 55 different problems.

We conducted simulations in which the different syntactic types of problems in the training set were sampled according to biased, empirical frequencies, as well as a control simulation run using equal frequencies for all problem types. As students work towards algebra and more advanced mathematics, they may see Operation=Answer less often, but we chose a stationary distribution to show that U-shaped learning need not depend on changes in the distribution of training examples. We used frequencies[1] found by Capraro et al. (2011) in a second-grade US textbook (Charles, Crown, & Fennell, 2007). Of problems that had operations on one side of the equal sign and a prompt for an answer on the other side, they found that 90.52% had operations on the left and only 9.48% had operations on the right. Accordingly, we sampled operation on left problems with a frequency of 90.52%. Of the problems with all of the operations on the left[2], 60.65% involved a prompt to solve for the sum and 39.35% for an addend. Based on this, we sampled sum prompt problems with the relative frequency of 60.65%; the remaining cases involved either the first or the second of the two addends with equal frequency (19.68%). Table 1 shows the probability of the model seeing each type of problem.

Representations

The network contained five visible pools of units: three for the numbers and two for the operators. The operator pools each contained two units: one for '+' and one for '='. We considered several different ways of representing the numbers. One possible choice of representation was localist, where each numerical value was represented by a one unit out of ten in each pool. However, because localist patterns are orthogonal, the network would not be able to generalize its knowledge of addition across different numbers; furthermore, childen come to addition problems with prior knowledge of numerical magnitude, so we sought a representation that would represent similar magnitudes with similar patterns. If we applied a Gaussian distribution of activation across these localist units, we could create a linear similarity structure, allowing the network could infer that results from operations involving 4 should fall between those for 3 and 5, but this introduced the need to specify a value for the spread of the Gaussian. To avoid this, we opted for a 'thermometer' representation in which the number of units active in the pool

[1] Frequencies of equation structures have varied between countries (Capraro et al., 2011; Li et al., 2008), between decades (Capraro, Capraro, Younes, Han, & Garner, 2012), and between grade levels (McNeil et al., 2006; Rittle-Johnson et al., 2011; Powell, 2012).

[2] There were not enough problems with operations on the right for reliable frequency estimation.

corresponds to the value of the number (no units are active for 0, only the first for 1, the first and second for 2, etc.). Zorzi, Stoianov, and Umiltá (2005) argue that such a representation naturally maps to lower-level processes and enumeration procedures. To equate the amount of activation across different numerical values, there was a second thermometer in each pool, whose units reflected the value of the number by turning off rather than on.

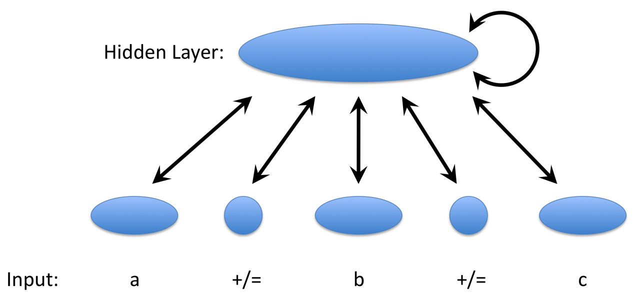

Figure 1: The model architecture. There are five visible pools (number, operator, number, operator, number), which act as both input and output of the model. These visible pools have bidirectional connections to a hidden layer, whose units also have connections between themselves.

Architecture, Learning, and Aggregation Over Replications

The architecture of this model is shown in Figure 1. The ovals represent pools of units and the arrows represent weights connecting every unit in one pool to every unit in the other pool. The strength of the net input to each unit (plus a fixed bias of -2) determines its activation according to a nonlinear (sigmoidal) function. The weights were initialized to small random values. The model is fully recurrent, and so activation propagated through the model gradually over seven time intervals, divided into sub-intervals or ticks to approximate continuity (Plaut, McClelland, Seidenberg, & Patterson, 1996). The units in all but the prompted slot were clamped to the specified input values for three intervals, during which activation propagates to other units. There were four intervals of further processing; during the last two, the activations of units were compared to their correct values (corresponding to the full equation with the prompted slot filled in). All the weights were adjusted in order to reduce the discrepancy between activation and target according to the (recurrent) back-propagation algorithm (Rumelhart, Hinton, & Williams, 1989).

The simulation results reported below used a learning rate of 0.0001. Initial learning with thermometer representations was faster than with localist representations because similarity allows transfer of knowledge; however, later phases of learning were slower because it is more difficult to distinguish representations that are partly overlapping than those that are orthogonal.

All results below reflect the average behavior of 12 simulations run with the same parameters, but we randomized initial weights, biased training set, and order of training experiences. Each run showed the same qualitative effects.

Results

Problem Solving as an Effect of Training Frequency

The control training set (fixed to sample each problem once an epoch) led unsurprisingly to learning the six problem types at an equal pace. We predicted that the biased training set (randomly sampling each epoch according to empirical frequencies) would show faster learning for the typical, operation on left problems, and slower learning for the atypical, operation on right problems. The learning trajectories for each problem type using the biased training set is shown in Figure 2. The model showed the predicted main effect of syntactic order, but there was an interaction. Between epochs 60 and 120, the model becomes worse at $a = b + \_$ problems, all while gaining practice and learning from feedback. This key problem type had the lowest frequency (3.73%) relative to its partner type with the same prompt position but the more typical operation on left structure, because both the bias for operation on left structure and the bias to prompt for the sum favored its alternative.

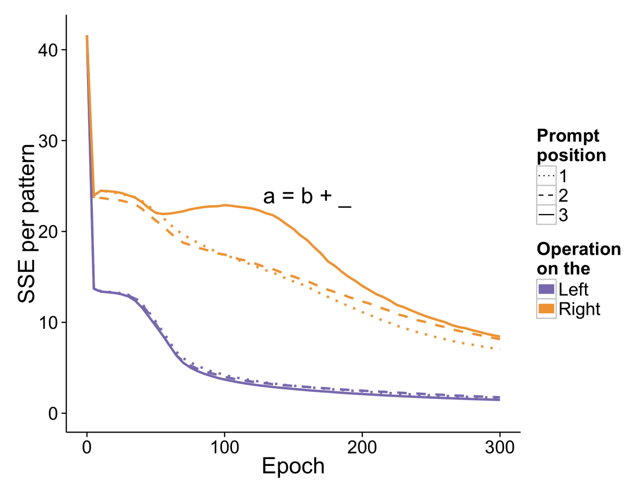

Figure 2: The trajectory of error over the time course of training. Error is the mean sum of squared error for each pattern. Time is measured in epochs, which each consist of 330 trials. The performance is separated by problem type (shown in Table 1).

Strategy Development

Having seen there was indeed a non-monotonic learning trajectory, we now show that the model first learned an add-all strategy and then gradually overcame this to produce correct responses consistent with the principle of equivalence, as observed in the developmental data. In order to classify the model's response, we calculated the cosine similar-

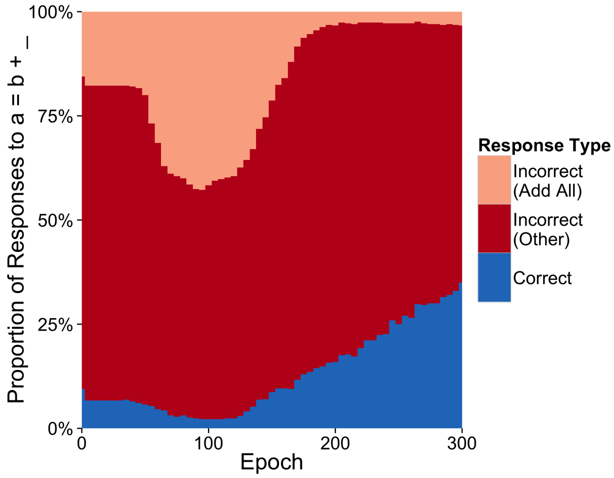

Figure 3: The distribution of responses to $a = b + \_$ (where $b \neq 0$) over the time course of training. Responses were coded as correct or incorrect, and incorrect responses were coded as 'add-all' if the response was within 1 of the sum, $a + b$. The add-all response would be correct if the operation were on the left (as is typical).

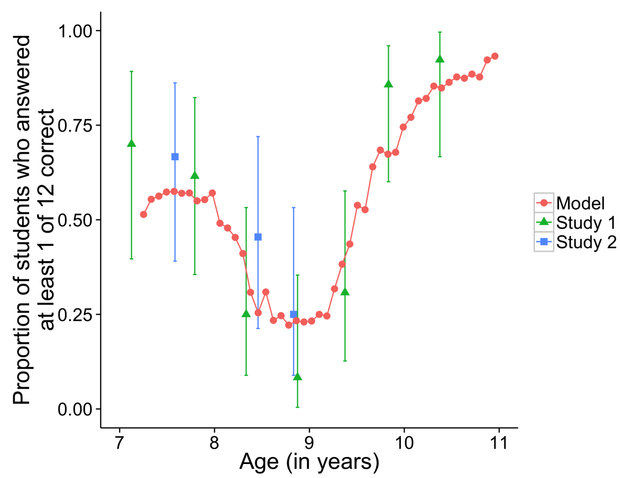

Figure 4: Model performance compared to human performance over time. Study 1 and Study 2 data are from McNeil (2007), with 95% Wilson confidence intervals shown.

ity between the model's output and the correct activation pattern corresponding to each of the 10 digits. We then chose the digit with the largest similarity as the model's response choice. In Figure 3, responses to our key "$a = b + \_$" problem are divided on the basis of whether they are correct, and the incorrect responses are subdivided on the basis of whether they are within 1 of the sum $a + b$ (the same margin used by McNeil (2007) to code children's strategies). The sum of a and b was the response generated from an add-all strategy as a result of mistakenly transferring their learning from "$a + b = \_$". Responses to such problems did not change much over the first 60 or so epochs, but then there was a large increase in add-all responses. The model actually performed below chance (10%) on "$a = b + \_$" before overcoming this to make progress toward the correct solution. This pattern has been noticed in children as well (McNeil, 2007), where 7-year-olds outperform 9-year-olds on final blank equivalence problems.

Comparing the Model to Human Performance

We now consider whether our model can match the U-shaped pattern of performance observed in empirical studies. McNeil (2007) presented 12 atypical equivalence problems to two groups of students (Study 1 and 2)[3]. Figure 4 compares these empirical results[4] to the performance of our model. We simulated this task by randomly sampling 12 problems of the key type $a = b + \_$ and scoring whether any responses were correct. We repeated this process 100 times for each run of the model at each epoch, and averaged the results. In order to translate epochs of the model into years of a child's age, we shifted and scaled this variable such that the sum of squared error between human and model data points was minimized; the best fit was obtained by setting $Years = 7.252 + 0.0161 * Epochs$. Using these values, our model explained 80.64% of the variance in the data from these studies. More importantly, our model and the McNeil (2007) studies showed very similar qualitative performance: poor but somewhat random responses which often include at least one that is correct, which first lead to consistently wrong responses following an add-all strategy, and then rise to a higher level with consistently correct responses.

Encoding and Reconstructing the Equation

Our model captures not only how students learn to solve equivalence problems, but also how students encode, maintain and reconstruct representations of the equation. McNeil and Alibali (2004) showed fourth graders (around ten years old) equations with different structures, and asked them to wait five seconds and then reconstruct the equation. Children tended to rearrange (erroneously) the operators to conform to their most frequent positions, and our model exhibited the same behavior. Figure 5 shows the final activation of the unit for an equal sign in the right (typical) position, at the end of a total of seven intervals of procesing, including four after the problem input was removed. This activation corresponds to the degree of belief that an equal sign had been present in the input in this position. The model saw operation on left problems much more frequently, and its outputs quickly came to conform to this bias; gradually, it overcame this bias as it continued to learn, reflecting the actual input presented.

Interestingly, for operation on right problems with a prompt in the final position, there was a perturbation in how

[3]The Study 2 students received brief lessons in order to match performance of similarly-aged children in Study 1.

[4]These data appear in Figures 2 and 3 from McNeil (2007).

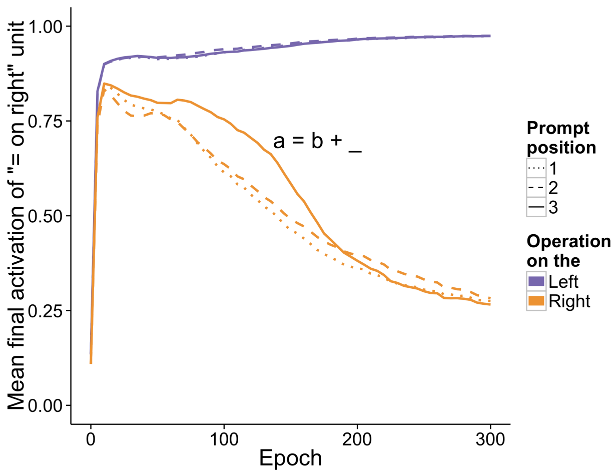

Figure 5: Model reconstruction of equation structure over the time course of training. The y-axis represents the mean final activation of the unit which, when active, signifies an equal sign in the right (typical) location. Each line corresponds to a different problem type. Activation of the unit for the plus sign (not shown) is approximately one minus the activation of the unit for the equal sign.

the model encodes the operators. At this point in its development, the model had a stronger tendency to misrepresent the operator in $a = b + \_$ problems than in other types of problems with operations on the right; the children tested by McNeil and Alibali (2004) also exhibited this tendency. The perturbation coincides with the increase in add-all responses after epoch 60 and is due to the interactive nature of the model: A similar network but with a non-interactive architecture did not show the same behavior. Interactive forces in the model go beyond mere frequency explanations and could offer a better understanding of the underlying cognitive mechanisms.

Discussion

In summary, a recurrent neural network can serve as a useful model of how we learn mathematical equivalence. First, we found that a biased but stationary environment, which (like textbooks) is strongly biased to equations with operations on the left and weakly biased to problems prompting for the sum, is enough to explain U-shaped development. We investigated the strategies of the model, and found that it adopted an add-all strategy like many children do before mastering equivalence problems. Finally, we looked at the maintenance and reconstruction of the operators (+ and =), pieces of the equation which were provided to the network at the beginning of every trial. The model is strongly biased at first to see the addition operator on the left and the equal sign on the right, and with its interactive nature, the add-all strategy caused the model to be even more likely to misperceive the equal sign when the prompt occurred in the right-most position.

Because of its detailed dynamics, the model makes predictions about the time course of processing that go beyond the granularity of the data from exising studies. The amount of time the model takes to reach some threshold and to complete a problem could be compared to human reaction times on equivalence problems. Eye-tracking studies may address at what stage of attention, encoding, retention and retrieval students mistakenly move the equal sign to its routine position. Some initial findings (J. Parvizi, personal communication, Nov. 18, 2013) from electrocorticographic recordings show that the neural activity that occurs while perceiving numbers or plus/minus signs does not occur while perceiving the equal sign in the context of a typical addition problem.

Our current model is limited to addition problems that always involved three single-digit numbers, and processes its inputs in a single parallel settling process. A full account should allow for indefinitely complex expressions on both sides of the equal sign. It seems likely that a sequence of operations is often required to process such expressions, and this is outside the scope of our current model. Neural networks capable of processing inputs sequentially have been proposed for language processing (e.g., Elman, 1990; McClelland, St. John, & Taraban, 1989), and such models can exhibit biases and U-shaped patterns similar to those in the current model. The current model also does not have any gesture or speech layers. Asking students to define the equal sign or explain their performance (often using both speech and gesture) can identify where they are in the transition from operational understanding to relational understanding (Perry et al., 1988). How language and metacognition interact with problem solving remains an issue for further work.

The model presented in this paper applies PDP principles used in other domains to the learning of procedures that respect the principle of mathematical equivalence. Gradual implicit learning is shown capable of simulating the strategies underlying the U-shape in children's understanding. The interactive nature of syntax, semantics, and context explains both the biased reconstruction of equations and the development of correct performance. Based on our simulations, we recommend educators consider including more equations of varying structure in their curricula. Because the location of both the equal sign and the prompt influenced results, we encourage researchers to attend to not only how a solved equation appears but also how a student interacts with it (in this case, the prompt). This model offers potential improvements to educational materials, and it advances our theoretical understanding of core issues in the field of mathematical cognition.

Acknowledgements

We thank members of the PDP Lab for helpful discussion. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic compo-*

nents of thought. Mahwah, NJ: Lawrence Erlbaum Associates.

Capraro, R. M., Capraro, M. M., Yetkiner, E. Z., Corlu, S. M., Ozel, S., Ye, S., & Kim, H. (2011). An international perspective between problem types in textbooks and students' understanding of relational equality. *Mediterranean Journal for Research in Mathematics Education*, *10*(1–2), 187–213.

Capraro, R. M., Capraro, M. M., Younes, R., Han, S. Y., & Garner, K. (2012). Changes in equality problem types across four decades in four second and sixth grade textbook series. *Journal of Mathematics Education*, *5*(1), 166–189.

Charles, R., Crown, W., & Fennell, F. (2007). *Scott Foresman - Addison Wesley mathematics grade 2*. Upper Saddle River, NJ: Pearson.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 58, pp. 117–165). Amsterdam: Elsevier.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? evidence from solving equations. *Journal for Research in Mathematics Education*, *37*(4), 297–312.

Li, X., Capraro, M. M., & Capraro, R. M. (2008). Sources of differences in children's understandings of mathematical equality: Comparative analysis of teacher guides and student texts in china and the united states. *Cognition and Instruction*, *26*(2), 195–217.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 8–45). New York: Oxford University Press.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, *4*(3–4), 287–335.

McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology*, *43*(3), 687–695.

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science*, *28*(3), 451–466.

McNeil, N. M., Fyfe, E. R., Peterson, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4=2+2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development*, *82*(5), 1620-1633.

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-school students' understanding of the equal sign: The books they read can't help. *Cognition and Instruction*, *24*(3), 367–385.

National Research Council. (1998). *The nature and role of algebra in the k-14 curriculum*. Washington, DC: National Academy Press.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development*, *6*(4), 449–468.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, *3*(4), 359–400.

Piaget, J., Inhelder, B., & Szeminska, A. (1960). *The child's conception of geometry*. New York: Basic Books.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115.

Powell, S. R. (2012). Equations and the equal sign in elementary mathematics textbooks. *The Elementary School Journal*, *112*(4), 627–648.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, *91*(1), 175–189.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85–104.

Rogers, T. T., Lambon-Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Rogers, T. T., Rakison, D. H., & McClelland, J. L. (2004). U-shaped curves in development: A PDP approach. *Journal of Cognition and Development*, *5*(1), 137–145.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1989). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1989). On learning the past tenses of english verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

Siegler, R. S., & Araya, R. (2005). A computational model of conscious and unconscious strategy discovery. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 33, pp. 1–42). Oxford, UK: Elsevier.

Zorzi, M., Stoianov, I., & Umiltá, C. (2005). Computational modeling of numerical cognition. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 67–84). New York: Psychology Press.