JOURNAL OF COGNITION AND DEVELOPMENT, *6*(1), 65–88

How Mathematics Propels the Development of Physical Knowledge

Daniel L. Schwartz
School of Education
Stanford University

Taylor Martin
College of Education
The University of Texas at Austin

Jay Pfaffman
College of Education
The University of Tennessee

Three studies examined whether mathematics can propel the development of physical understanding. In Experiment 1, 10-year-olds solved balance scale problems that used easy-to-count discrete quantities or hard-to-count continuous quantities. Discrete quantities led to age typical performances. Continuous quantities caused performances like those of 5-year-olds. In Experiment 2, 11-year-olds solved problems with feedback. They were encouraged to use math or words to justify their answers. Children who used math developed an understanding superior to most adults, whereas children who used words did not. In Experiment 3, 9-year-olds solved problems with or without prompts to use math. Children encouraged to use math exhibited greater qualitative understanding, even though they were unable to discover metric proportions. The results indicate it is possible to design symbolic experiences to propel the development of physical understanding, thereby relating developmental psychology to instructional theory.

Mathematics, diagrams, and other explicit representations help scientists discover and organize complex empirical relations and this also may be true for the developing child. The proposal that external representations and organizing activities contribute to development comes from Vygotsky's (1978) foundational insight that

Requests for reprints should be sent to Daniel L. Schwartz, School of Education, 485 Lasuen Mall, Stanford, CA 94305–3096. E-mail: Daniel.Schwartz@stanford.edu

cultural forms mediate experience. The following studies examine the significance of mathematics, a cultural form, in helping children develop complex physical knowledge about balance. In these studies, mathematics does not help children learn primitive perceptual categories of causality, such as weight, torque, or balance. Instead, mathematics helps children structure complex causal relations; for example, when children need to coordinate multiple parameters of distance and weight to determine balance.

Research on the development of physical concepts often emphasizes qualitative physical intuition (e.g., Baillargeon, 1994; Spelke, 2000; Stavy & Tirosh, 2000). For example, diSessa (1983) argued that people have perceptual primitives, like springiness and torque. Development involves sorting among the primitives to find the most general and explanatory. In those accounts that do consider mathematics, mathematics only lends precision to situations that are previously understood qualitatively (e.g., Ahl, Moore, & Dixon, 1992; Dixon & Moore, 1996; Halford, 1993; Resnick, 1992; White, 1993).

For basic spatial and physical relations, a qualitative understanding that "more X causes more Y" may be sufficient (e.g., Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Starkey, 1992). However, the challenge of advanced development may not involve inducing qualitative relations available to perception (e.g., Spinillo & Bryant, 1991). Instead, the challenge may involve determining how to structure the multiple, perceptually distinct parameters of more complex casual situations. This may be difficult without reliance on cultural tools like mathematics. As the Nobel physicist Feynman (1965) stated, "it is impossible to explain honestly the beauties of the laws of nature in a way that people can feel, without their having some deep understanding in mathematics. I am sorry but this seems to be the case" (p. 39).

Cross-cultural work has shown the importance of mathematics for the acquisition of complex quantitative understanding. Saxe (2001), for example, documented how the introduction of a cash economy required Papua villagers to adapt their body-based counting system so that they could reason about addition and doubling. Those who did not learn to use the new notational system were limited in their computational abilities. We are unaware, however, of cross-cultural or experimental evidence that shows whether mathematics influences the development of physical knowledge, and if so, what specific representational properties of mathematics might support this development. In a review of experimental research on proportional reasoning tasks, Surber and Haines (1987) stated, "The ease with which a variable can be quantified may influence the strategy subjects employ, especially when they are on the verge of discovering metric proportions" (p. 37). They did not provide evidence for this claim or explore its implications for development. Similarly Siegler (1981) observed that children seem to use their knowledge about the effects of transformations on number to learn about the transformations' effects on liquid and solid quantities" (p. 62). Yet why this would be the case was not at focus given his emphasis on identifying children's cognitive structures rather than the material conditions of their development.

There are several reasons that the numerical notations and operations of mathematics may cultivate the development of physical understanding. Children perceive quantities at an early age and, therefore, mathematics may harness a "privileged" domain of cognition (Gelman & Gallistel, 1978; Huttenlocher, Jordan, & Levine, 1994). Case and Okamoto (1996), for example, proposed that some perceptually based mathematical representations, like a number line, tap into central conceptual structures of magnitude and complement children's natural mode of thinking and development (cf. Ginsburg, 1982). Here, we focus on nonperceptual properties of mathematics that may support the development of physical understanding.

One property is that the use of mathematics in physical situations presupposes measurement. Measurement transforms difficult to relate perceptual quantities into a common numerical ontology that supports precise comparisons and relations. The weights and distances of a balance scale, for example, involve different perceptual modalities (haptic and visual), which make them difficult to relate analytically. However, by converting them both to numerical measures, they can enter into a relation, as in $2 + 3$ or 2×3, and children can reason without immediate reference to different types of perceptual experiences.

A second, related property is what Bruner (1966) called, "compactibility." Numerals provide a compact representation that can alleviate working memory burdens. For example, when thinking about a balance scale correctly, children need to combine the weights on each side of the fulcrum and their distances. Numbers can efficiently represent the relevant quantities, at least compared to maintaining mental images.

A third property is that arithmetic provides familiar structures for organizing multiple parameters. Addition, subtraction, multiplication, and division provide ready-made candidates for organizing complexity. For example, Schwartz and Moore (1998) found that children reasoned proportionally about water and juice concentrate when the associated numbers were within their arithmetic competence but not when there were difficult numbers or no numbers at all.

Finally, arithmetic provides a mental technology for trying out different possible structures. A challenge for learning and development is what to do when one's prior knowledge fails. Even if failure is often a precondition of progress, failure does not include the mechanisms for constructing new knowledge. It is important to have alternatives that one can try out. For example, imagine a balance scale that has two stacks of weights on one side and one stack of weights on the other. It is difficult to infer the operative rule, unless one generates specific values and explores different arithmetic operations for combining those values. The multiple operations and possible configurations of arithmetic offer a way to explore alternative structures.

To evaluate the significance of mathematics for the development of physical understanding, we chose a physical device that has a well-documented progression of development. These studies used the balance scale task in which children decide if a beam will tilt left, right, or balance given the placement of weights on either side of the fulcrum. Our strategy was to show that mathematics leads to greater development as measured by performance on the balance scale task, and given that evidence, to look more closely at the children's reasoning to discern why mathematics made a difference.

There are multiple characterizations of the developmental plateaus revealed by the balance scale (e.g., Case, 1985; Metz, 1993; Piaget, 1954; Surber & Haines, 1987). However, the goal here is not to characterize the many ways children can (mis)understand the balance scale. Instead, the goal is to document the value of mathematics for propelling cognitive change, whatever the ultimate internal form may be. Siegler's (1981) widely adopted rule levels provided a parsimonious marker of developmental change. These rules characterize the child's ability to attend to and coordinate two relevant dimensions of information, distance, and weight. At Rule 0, children guess at random, unable to coordinate even within a single dimension. At Rule 1, children pay attention to weight exclusively and therefore make mistakes when the distances are unequal. At Rule 2, children pay attention to the distances but only if the weights are equivalent. If the weights differ, they ignore the distances. At Rule 3, children attend to both weight and distance. They no longer have a weight bias. They correctly reason when the weights or distances (or both) are equal but when the weights and distances both differ, they "muddle through" the problems, often operating at chance. Finally, at Rule 4, children apply metric proportional reasoning and compare the ratios of weight and distance.

We conducted three studies with 9- to 11-year-olds. One can anticipate their rule levels without any interventions by looking at Table 1, which shows the distribution of rule levels by age as found in Siegler's (1981) original sample. Our first experiment hindered the application of mathematical knowledge by making it difficult for children to measure the weights and distances. The children should do worse than their age-level norms. The second and third experiments encouraged children to use mathematics to justify their answers to help learn about the balance scale. The children should do better than their age-level norms. In addition, in the

TABLE 1
Age Norms for Rule Attainment (Adapted from Siegler, 1981)

	Percentages of Children at Each Rule Level				
Age	*Rule 0*	*Rule 1*	*Rule 2*	*Rule 3*	*Rule 4*
5 years	5%	85%	5%	5%	—
8 years	—	10%	35%	45%	10%
12 years	10%	—	15%	60%	15%
Adult	—	—	5%	65%	30%

Note. At Rule 0, children guess. At Rule 1, children only pay attention to weight. At Rule 2, children pay attention to distance, when weights are equal. At Rule 3, children always pay attention to weight and distance but perform at chance when both distances and weights differ. At Rule 4, children solve all problems correctly.

latter two experiments, we examined children's explanations. These explanations help reveal why mathematics can support development, and their analyses may offer the most important contribution of this work.

EXPERIMENT 1

In this study, we made it difficult to measure the relevant parameters of the balance scale. We predicted this would make the children less likely to consider weight and distance simultaneously. Children completed paper-and-pencil balance scale problems in one of two conditions. In the baseline peg condition, children solved standard problems that showed discrete quantities of weight and distance. In the beaker condition, children solved identical problems, except that the quantities were continuous. We assumed that the continuous quantities would be difficult to measure and enumerate in symbolic form and this would block the use of mathematical representations. Prior research documents children's difficulty with mass quantities compared to count quantities. Nevertheless, it seemed worth starting with a simple but untested instance of problem solving with mass quantities in the context of our account.

Figure 1 provides examples from the two conditions. The peg condition portrayed a balance scale with four pegs per side and a maximum of three weights per peg. The beaker condition used two beakers, each filled to one of three levels to indicate weight. The beakers rested at the same four distances as the peg condition, though the distances were unmarked. Children chose from three options to predict

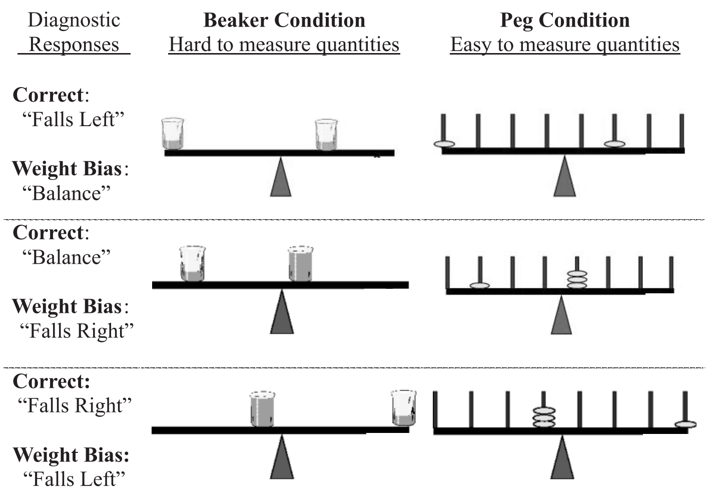

FIGURE 1 Three types of problem used to assess child understanding. For each problem, the correct answer conflicts with an answer that only considers weight. Children either solved problems involving hard-to-measure, continuous quantities or easy-to-count, discrete quantities.

what would happen to the scale once it was free to move: fall down to the left, fall down to the right, balance.

We designed the problems to assess whether the children reasoned about both the weight and distance dimensions, or if they reasoned only about weight, which we will call a "weight bias." There were three problem types.

1. Same weights at different distances: The top of Figure 1 shows this problem type. Children with a weight bias give a balance response because there are equal weights (Rule 1). More sophisticated children use the different distances to correctly predict the direction of tilt (Rule 2 and up).
2. Different weights at different distances that balance: The middle of Figure 1 shows this problem type. Children who cannot solve the first problem type avoid a balance response for this problem because there are different amounts of weight. They predict a tilt to the side with the more weights (Rule 1). Even children who can solve the first problem type may predict a tilt to the side with more weights because the use of unequal weights increases problem complexity and leads them to disregard distance (Rule 2). More sophisticated children who consider both weight and distance are willing to predict balance, though at chance levels (Rule 3), unless at the most advanced level (Rule 4).
3. Different weights at different distances that tilt toward less weight: Children should behave similarly to the second problem type.

Our prediction was that children in the beaker condition would be correct less often and exhibit more of a weight bias. Without access to measurement and the symbols of mathematics, beaker children would have to coordinate differences in weight and distance qualitatively, which we propose is difficult to do.

Method

Participants. Two classes of beginning 5th-grade children in suburban Tennessee participated ($M = 10.7$ years, $SD = 0.50$). Of those children, 33 returned permission slips that allow us to report their results, 19 in the beaker condition and 14 in the peg condition. There were 17 girls and 16 boys, equally distributed except one.

Design. Children were randomly assigned to a condition. For each of the problem types previously described, children were supposed to solve four separate instances. Due to experimenter error, the beaker condition included five instances of the third problem type (different weights and distances that tilt to less weight). There were no discernible consequences, and otherwise, children solved four instances for each problem type.

Procedures. Each teacher received equal numbers of peg and beaker packets randomly mingled in a single stack. The teacher handed out the packets, which the children completed individually as seatwork. Each randomly ordered page had a single problem. For the peg condition, there were 12 problems, plus 4 filler problems. For the beaker condition, there were 13 problems, plus 3 filler problems. A coversheet showed a problem with equal weights (or equally filled beakers) at equal distances from the fulcrum. The instructions described the balance scale and told the children their task was to predict whether the scale would balance, fall down to the left, or fall down to the right. The instructions stated that the correct answer to the sample problem was balance and that the children should circle one of the three choices for the subsequent problems.

Results

An alpha level of .05 was used throughout for all statistical tests.

Figure 2 indicates that the beaker children performed worse than the peg children did. The top panel shows the mean percentage of correct answers for each of the three problem types. The bottom panel shows the mean percentage of weight-biased responses. The correct responses help indicate whether children focused solely on the weights because correct answers directly competed with weight-biased answers. The weight-biased responses also help indicate whether children focused solely on weight because an incorrect response could be weight-biased. The peg children were correct more often and less biased across the three problem types.

For each child we found the percentage of correct responses and the percentage of biased responses for each of the three problem types. This created six within-subjects measures (see Figure 2). Condition was a between-subject factor. Using a multivariate analysis of variance (MANOVA), there was a significant effect of condition, $F(6, 26) = 2.73$, $t^2 = 0.63$. Within the umbrella of the MANOVA, we compared the two conditions on each measure for each problem type. Figure 2 indicates significant condition effects with asterisks. For the problems with equal weights at unequal distances from the fulcrum, the beaker children made fewer correct “tilt” answers than the peg children, $F(1, 31) = 5.9$, $MSE = 0.14$, and more weight-biased “balance” answers than the peg children, $F(1, 31) = 4.8$, $MSE = 0.14$. Thus, the beaker children ignored the distance dimension. For the problems with different weights and distances that balanced, the beaker children were correct less often than the peg children, $F(1, 31) = 7.7$, $MSE = 0.04$. The beaker children had difficulty reasoning how different weights could balance. However, incorrect peg children were also inclined to use a weight-biased answer, and there was not a significant difference in the percentage of weight-biased responses, $F(1, 31) = 2.2$, $MSE = 0.12$, $p < .15$. For the remaining problem type (different weights

PERCENTAGE OF CORRECT ANSWERS

	Problem Type					
						
	Same Weight “Tilt Left”		Different Weight “Balance”		Different Weight “Tilt Left”	
Condition	M	SD	M	SD	M	SD
Pegs	60.7*	(.42)	25.0*	(.27)	32.1	(.33)
Beakers	28.9	(.33)	05.2	(.12)	28.7	(.26)

PERCENTAGE OF WEIGHT-BIASED ANSWERS

	Problem Type					
						
	Same Weight “Balance”		Different Weight “Tilt Right”		Different Weight “Tilt Right”	
Condition	M	SD	M	SD	M	SD
Pegs	32.1*	(.40)	57.1	(.37)	57.1	(.31)
Beakers	60.5	(.35)	75.4	(.33)	68.2	(.29)

Note: Asterisk (*) indicates significant difference between conditions.

FIGURE 2 The mean percentage of correct responses and weight-biased responses broken out by condition and problem type. The top of each table displays a representative of each problem type plus answers in quotations that are correct (top panel) and weight-biased (bottom panel).

and distances that tilt toward less weight), there were no significant differences between conditions (*F*s < 1.0).

To coordinate these findings with other research, we reassess the results as rule levels. Few children were 100% consistent with one rule across the problems, as might be expected in a whole-class, paper-and-pencil task. If we required 100% consistency, six children could be assigned to a rule. However, at a 75% fit, all of the children could be assigned. For example, children at Rule 1 had to exhibit a weight bias for (a) at least three of the four problems with equal weights at different distances and (b) 75% or more of the remaining problems. For the beaker condition, the frequencies were as follows: Rule 1 = 68%, Rule 2 = 22%, Rule 3 = 5%, Rule 4 = 5%. For the peg condition, the frequencies were: Rule 1 = 29%, Rule 2 = 43%, Rule 3 = 14%, Rule 4 = 14%. The beaker condition is significantly below the peg condition in a one-tailed test, $\chi^2(3, N = 33) = 6.6, p = <.05$. The beaker children were also further below their age norms than the peg children (refer to Table 1), though this comparison should be interpreted in light of our imperfect assignment to rule level.

Discussion

Children in two conditions received isomorphic problems. The peg condition used discrete quantities. The beaker condition used continuous quantities. Children in the beaker condition exhibited less developed physical reasoning. We hypothesize that the beaker children did not measure the quantities into precise values that organized the weights and distances into a common ontology (number), and therefore, they had difficulty representing and coordinating the amounts. Given that they did not apply a preformed schema to these problems, they needed to reason with perceptual magnitudes. When perceiving differences on the weight dimension, they stopped looking at the distance dimension and they exhibited a weight bias. It is difficult to maintain a perceptual representation with sufficient precision to coordinate thinking about forces interacting at a distance from each other (Schwartz & Black, 1996). The results, however, are merely consistent with this hypothesis and do not prove it. Thus far, we have shown only that physical situations that are difficult to measure lead to poorer performance. In the following studies, we look for direct evidence that the conversion into numbers helps children consider the complex relations within a physical problem.

An alternative hypothesis is that the beaker children had difficulty perceiving the differences in the quantities. For example, the children may have been unable to see that the beakers had different amounts of water. This alternative has limited appeal. The children discerned the quantities in the beakers because they answered based on their equivalence or difference (i.e., the weight bias). With regard to distance, some problems made the differences pronounced (e.g., fourth position on the left side vs. first position on the right side). The beaker children did not improve on these problems. Also, 32% of the beaker children used the distances at some point in their thinking (Rule 2 or better) and it is difficult to fathom why the other 68% would have less perceptual acuity. Instead, it seems that the beaker children could perceive the relevant differences but they could not measure them with sufficient precision to coordinate the two dimensions of perceptual information symbolically and keep them in mind. Without measurement into numbers, they ignored distance in their reasoning.

EXPERIMENT 2

Experiment 1 tested the prediction that when it is more difficult to use mathematics, children will exhibit less advanced understanding. Experiment 2 tested the hypothesis that when children are encouraged to use mathematics, their understanding will become more advanced. The central manipulation was that children in the explain condition were encouraged to justify their answers in words, whereas children in the math condition were encouraged to justify their answers using mathematics.

To implement the experiment and give the children a chance to further their understanding, we developed the software shown in Figure 3. Each row of the interface presents increasingly difficult problems. When children clicked on a problem, they received a screen that enlarged the problem and required children to choose from: falls down to the left, balances, falls down to the right. Beneath the choices was a first-try "justification" box that implemented the major difference between the conditions. For the explain condition the justification box stated, "Explain your answer." For the math condition it stated, "Show your math." The system did not interpret the justifications. If children made a wrong choice, they simply received the correct answer, and then they had a chance to make a second-try justification

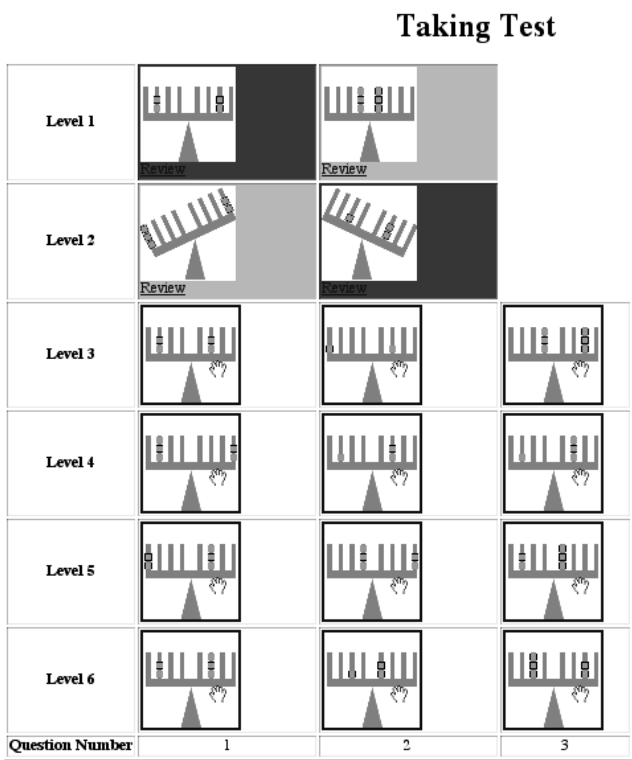

FIGURE 3 The at-a-glance interface children used to work on balance scale problems. (All problems used weights and pegs.) Children selected a problem, predicted its behavior, and justified their answers using words or math. The interface updated after each prediction to show the correct answer and the child's accuracy. In this example, the child has answered problems in the first two rows and made one correct answer in each case.

for why this answer was correct. After each problem, the system returned to the main interface. The interface showed the correct behavior of the scale, it colored the square green or red to indicate the correctness of the child's initial choice, and it showed the first-try justification (removed from the figure for legibility). The hope was that the interface would help children induce patterns and learn.

Before using the system, the explain and math children received brief on-line examples of what it means to explain or use math to justify an answer. These examples did not use a balance scale. The math children also received instruction on how to count (i.e., measure) the elements of the balance scale. The goal was to maximize the effect of using mathematics to determine whether and how it can propel development. In Experiment 3, these extra supports were removed to ensure that our training was not the sole reason that mathematics helped.

The children's understanding was evaluated in individual exit interviews. Children solved a new set of balance scale problems and explained their reasoning. To ensure the math children had not simply stumbled onto a successful formula without physical understanding, all children also solved two new kinds of problems shown in Figure 4. For a physical model problem, the balance scale was missing one of the pegs. If the children blindly apply a formula without physical understanding, they disregard the distance and simply count the number of pegs. For a transfer problem, the scale had weights on three pegs. For the example in Figure 4, the correct answer is "balance" and the math is: $2 \times 4 = (2 \times 1) + (3 \times 2)$. We did not expect children to find the correct answers; they had no exposure to the behavior of these systems and they did not receive feedback. We explicitly created a problem for which the children would not have a formulaic answer and they would have to reason about the physics of a new situation. If the math children exhibited advanced reasoning for the transfer problem, it would show that the use of mathematics had given them a deeper understanding, even when they could not use mathematics to derive a correct answer.

Overall, we expected the math children to show more advanced rule levels for the standard balance scale task and to exhibit more complex thinking for the transfer problems. Given this result, we also expected their typed justifications to yield evidence for how mathematics helped the children learn about the balance scale.

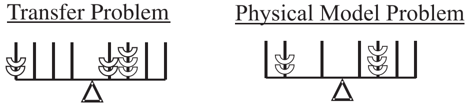

FIGURE 4 Problems used to determine if children reasoned about a physical device and whether they could transfer their learning to a more complex problem. To reason physically about the problem on the right, children had to consider distance (or the weight of the missing peg) instead of blindly count pegs. Transfer problems included weights on three pegs to determine if children reasoned about weight and distance for novel, complex problems.

These latter data were analyzed post hoc, so we defer further explanation to the Results section.

Method

Participants. Four new 5th-grade classrooms in suburban Tennessee participated at the end of the school year ($M = 11.3$ years, $SD = 0.54$). Children were randomly assigned to the explain ($n = 29$) and math ($n = 30$) conditions. There were 30 girls and 29 boys, equally distributed except one. A subset of 51 children completed the transfer task: 26 and 25 for the explain and math conditions, respectively.

Design. The explain and math conditions comprised the only factor. The study included three a priori dependent measures. One measured which of the four rule levels children exhibited at posttest. Another tested whether children simply counted pegs without reference to their relative distances. The third measured whether children transferred their learning to three problems that included weights on three pegs. Post hoc analyses characterized child justifications during learning.

Procedure. Each class had four computers. Children participated when a computer became available. Once seated, a child received an online introduction. The introduction began with an image of a balance scale and stated the children would solve problems about a balance scale like the one in the picture. The next page stated they would have to justify their answers. To give examples of what this meant, children saw two graphically presented problems in turn. One problem asked who had more total candy given groups of red, blue, and green candies. The other problem asked who would win a race given different speeds and starting positions. For each problem, children read two different examples of justifications. The explain condition used qualitative verbal justifications. For example, “Ben has more for two types of candy, but Mike has more types of candy. Mike has more.” The math condition provided examples of mathematical justifications. For example, “Mike: 4 + 4 + 4 = 12. Ben: 5 + 5 = 10. 12 is more than 10. Mike has more.” The sample math justifications were not analogous to the justifications children used for the balance scale. The introduction for the math condition also included a final page that showed how to measure the balance scale. To show how to count distance, the screen displayed numbers above the pegs (i.e., 4 3 2 1 || 1 2 3 4). To show how to count weight, the screen displayed numbers with arrows pointing to each weight’s position in its stack (i.e., bottom weight equals 1, next weight up equals 2, etc.). The example did not show how to use the numbers.

After completing the introduction, children completed the learning phase previously described (approximately 30 min). The system (Figure 3) permitted children to select uncompleted problems from the visual interface and provided feedback on the accuracy of their multiple-choice answers, but not their justifications.

After the learning phase, children moved to a separate location for an audiotaped interview. The interviewer asked each child to answer and explain three new balance scale problems that were of the three types used in Experiment 1. These problems were used to locate a child's rule level (see Introduction to Experiment 1). A risk, however, was that children who were at Rule 3 could choose the correct answers by chance (unlike children at Rule 1 or 2, who choose weight-biased answers). To address this limitation, the interviewer used a fourth and fifth problem for children who correctly answered all the problems but would not explain how. This reduced the odds of mistakenly coding children at Rule 4 when they were really at Rule 3. The interview ended with the physical model problem (right side of Figure 4). Children who only counted the visible pegs to reach an answer apparently did not view the pegs as markers of distance. Children had to count the missing peg as though present or comment on the change in weight due to the peg's absence to get credit for treating the scale as a physical system.

After all the children completed the interview, the classes completed three transfer problems as paper-and-pencil seatwork (left side of Figure 4). For one problem, the scale tilts to the side with more total weights; for another, it tilts to the side with the maximum distance; and for another, it balances. Children circled their answers from the usual three choices. Children who only use weight always predict that the beam tilts to the side with more total weight for all three problems. Children who only use distance always predict a tilt to the side that has the weights the farthest out. Children who consider both dimensions predict at least one scale tilts toward more weight and one scale tilts toward more distance or choose balance.

Results

The interview included the physical model problem to determine if children were simply "pushing numbers." Two math children were caught by this trap question compared to zero explain children. We exclude them from the following analysis because their physical understanding is unknown. One math child developed an algorithm that added the pegs and weights on each side of the fulcrum (cf. Case, 1985). This solution, though not described by Siegler (1981), was counted as Rule 3 because the child considered both dimensions when the weights and distances were unequal.

Figure 5 shows that the math children exhibited higher rule levels than the explain children in the exit interview, $\chi^2(3, N = 57) = 15.6$, $p = <.01$. For the math condition, the mean Rule level was 3.4 (*SD* = 0.68), and for the explain condition, 2.7 (*SD* = 0.65). The difference is reliable; $F(1, 56) = 13.0$, $MSE = 0.44$.

The math children also did better on the transfer problems that used weights on three pegs. The percentage of children who considered both dimensions was 93% for the math condition and 36% for the explain condition. A revealing comparison includes only children who exhibited Rule 3 or better on the standard balance scale problem. Of those who completed the transfer task, there were 25 in the math condition and 20 in the explain condition. Of these children, 96% of the math children considered both dimensions compared to 65% of the explain children, $\chi^2(1, N = 45) = 7.3$, $p = <.01$. Thus, even though these explain children considered both dimensions for the original balance scale problem, many regressed to using a single dimension on the transfer problem.

The justifications during the learning phase suggest how mathematics helped. The children's justifications were coded post hoc for telling features. Two of the coding schemes required subjective agreement: The number of dimensions a child considered in a justification (weight or distance) and the number of distinct ways a child tried to compare and combine the information. The schemes are detailed in the following paragraphs. To ensure the reliability of each coding scheme, a primary coder worked with all the data and a secondary coder worked with a random 25% sample. In this study and the next, there was perfect agreement on the dimensions of information a child considered for each justification. The raters' tallies of the number of distinct ways each child combined the information were highly correlated for both studies ($r = .95$).

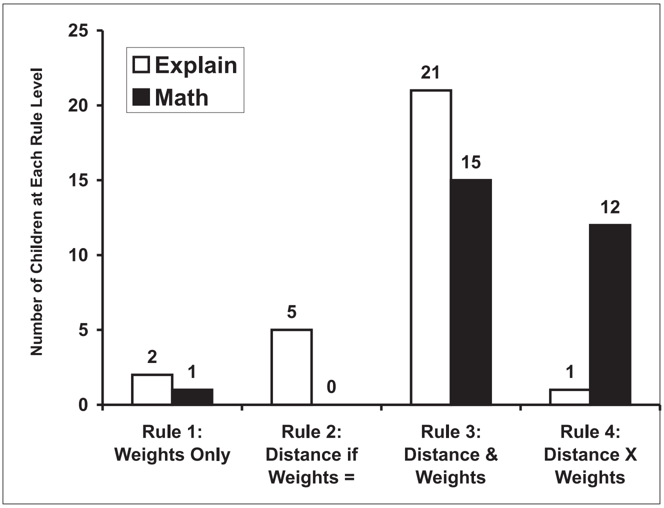

FIGURE 5 The frequency of rule attainment by condition (11-year-olds, Exp. 2).

First, we report how the children from both conditions had similar opportunities to learn. The groups made similar efforts to justify their first answers for each problem. The math and explain children typed relevant entries for 98% and 93% of the problems, respectively. A relevant entry included a reference to the balance scale, as opposed to blank entries; random typing; and comments like, "I hate this."

Children from the two groups made roughly the same number of errors leading to corrective feedback (math $M = 4.3$, $SD = 2.8$; explain $M = 5.3$, $SD = 2.0$). The math children did not exhibit greater accuracy, because they were exploring how to solve the problems (see following) and often did not reach insight until the end.

Finally, we examined whether the children noted the weight and distance for each problem. A weight-only coding reflected when a child only referred to the weights in a justification (e.g., "more weight on the right" or "3 < 4"). A distance-only coding reflected attention to only distance (e.g., "father out on right" or "1 < 4"). Both distance and weight reflected when a child referred to both weight and distance in a single justification (e.g., "less weight but farther out," "3 × 4 > 2 × 1"). To determine if children noticed both dimensions, we found the percentage of children who used both a justification that considered distance and a justification that considered weight across the problems. The math and explain children both considered the weights and distances, (90% and 93%, respectively), though not necessarily for the same justification.

The children from both conditions noticed the weight and distance dimensions and they made good faith efforts with the task and feedback. However, there were telling differences in their justifications. For example, all but one math child used digits (e.g., "3"), whereas 83% of the explain children exclusively used words (e.g., "three").

More important, children in the math condition considered more dimensions in a single justification. Previously, we noted that most of the children considered the weight dimension and the distance dimension when collapsing across different problems. Here, the analysis focuses on how many dimensions they incorporated into a single justification. For the math condition, 68% of the children had at least one justification that included both dimensions compared to 19% of the explain children, $\chi^2(1, N = 59) = 14.2$, $p < .01$. In contrast, the math children frequently included all the parameters of both dimensions, even if they did not use them correctly, as in the case of one child who wrote, "3 × 3 – 1 = 2 – 1 = 1 9 – 1 = 8."

A second striking difference involved the exploration of new structural operations. We counted the number of times each child explored a new approach to solving the problem. A child took a new approach when she or he made a qualitative or quantitative comparison that combined or related the values in a new way, regardless of the specific parameters. For example, an explain child said, "greater weight" on one answer and "greater distance" on another. These are the *same* operation because they both use a qualitative subtraction (or comparison of magnitudes), even though they used different dimensions. In contrast, another child wrote, “more distance on the left,” and then on a subsequent problem switched to “distance and weight make more on the left.” The aggregation of weight and distance is a new operation. An example of multiple operations for the math condition comes from a child who began by subtracting the weights from the distances, then tried adding them, and then finally discovered that multiplication worked. Across the full problem set, explain children explored a new operation that differed from their initial operation an average of 0.3 times ($SD = 0.5$). Typically, when explain children were wrong and wrote a second-try justification, they simply highlighted the alternative dimension rather than trying a new operation (e.g., first justification: “more on the left,” second justification: “farther on the right”). In contrast, the math children tried 1.6 ($SD = 1.6$) new operations on average (excluding two children who discovered the correct solution immediately). The difference is reliable, $F(1, 56) = 16.7$, $MSE = 1.3$.

The exploration of new operations did not guarantee productive paths, however. For instance, one child began with “3 × 4 = 12” on one problem and concluded with “4 × 4 = 16 + 5 = 21” on the last problem.

Discussion

Children in the explain and math conditions received balance scale problems with feedback organized to maximize their chances of inducing the patterns of behavior. Even so, the children who were told to explain their understanding without constraint did not induce the underlying physical rule as well as children who were told to use mathematics in their explanations. Given the opportunity to use the symbol system of mathematics, the math children developed a superior physical understanding.

Evidence from the justifications indicated that the math children represented more of the problem parameters simultaneously and in comparable numerical formats. The explain children tended to switch between distance and weight rather than represent them simultaneously. Moreover, mathematics provided the children with a set of operators for combining the parameters in new ways. Coupled with the ease of keeping the four parameters in the mix, the operators allowed children to explore different structures that often led to a new way to understand the balance scale.

One alternative interpretation of the results is that the mathematics supported a syntactic trial and error that stumbled onto the correct answer. This did occur for two children who found the correct solution without understanding the physics of the problem. This documents a pitfall of mathematics; people can use algorithms without understanding the situation to which the algorithms apply (a well-known failing of much school-based instruction). Yet at the same time, this finding reveals the power of mathematics because these two children were able to invent a successful structure based on patterns, something the explain children did not do. More generally, the math children were attentive to the physics of the problem as indicated by their ability to reason about a problem that omitted a peg. In addition, they transferred their understanding to a new problem with weights on three pegs. The math children considered both weight and distance, whereas the explain children only considered a single dimension for these more complex problems.

A second alternative interpretation is that the instructions in the math treatment helped the children consider both dimensions by showing how to count pegs and weights. Siegler (1976), for example, showed that increasing the salience of the distance dimension helps younger children move from Rule 1 to Rule 2. By 10 years, however, children already attend to both dimensions when the weights are equal (assuming they are enumerable—see Experiment 1). The difference between the conditions was not whether the children noticed both dimensions; showing children how to count did not give the math children an unfair advantage that way. Rather, the difference between the conditions was in how the children used those dimensions. The explain children considered each dimension in turn, for example by stating that one side had more weight, and then on negative feedback, stating the other side had more distance. In contrast, the math children incorporated weight and distance into a single structure rather than successive answers.

EXPERIMENT 3

The next experiment tested whether the results of Experiment 2 would replicate with two modifications to ensure that we did not give the math children an unfair advantage. First, we removed the introduction that included sample justifications and counting instructions. We simply told the children that they had to invent math or use words to justify their answers. Second, we used a younger population. Rule 4 reasoning requires the use of mathematics, which is the level at which the math condition showed its greatest advantage in Experiment 2. A useful demonstration would show that mathematics also provides an advantage for qualitative reasoning (e.g., Rule 3) that falls short of full-blown, metric proportional reasoning. To make this demonstration, we involved younger children who were old enough to know arithmetic but were unlikely to discover the multiplicative relation. Our prediction was that the math children would still outperform the explain children, even though few would reach Rule 4. In this case, mathematics would help them consider and relate both dimensions simultaneously (Rule 3), and again, the justifications would help indicate why.

Method

Participants. Two 4th-grade classrooms from a semi-urban, California school participated at the end of the school year ($M = 9.1$ years, $SD = 0.46$). There were 23 girls and 22 boys, randomly assigned to conditions (explain $n = 23$, math $n = 22$) in nearly equal proportion.

Design and procedure. Children were randomly assigned to the explain and math conditions. Besides age, the major difference from Experiment 2 was that the children did not receive instructions or examples on how to count, use math, or explain their justifications. Instead, the introductory phase simply introduced the balance scale problem and stated that they would have to use math or words to justify their choice for what it would do. Another difference was that the children solved the transfer problems in the interview.

Results

Figure 6 indicates the math children performed at higher rule levels in the exit interview than the explain children, $\chi^2(1, N = 45) = 9.4$, $p = <.01$. The average rule level of the math condition ($M = 2.9$, $SD = 0.71$) was greater than the explain condition ($M = 2.3$, $SD = 0.77$), $F(1, 43) = 6.4$, $MSE = 0.55$. The math children rarely reached Rule 4. In the exit interview, none of the children used explicit calculations to solve the problems. Combined, these results suggest that the math children's advantage was not based on a specific mathematical procedure. Math simply led the children to consider both dimensions of the problem more frequently. No child in either condition fell for the trap question that tested whether they blindly counted pegs.

Like Experiment 2, the math children exhibited superiority on the transfer problems. In this study, we administered the transfer problems during the exit inter-

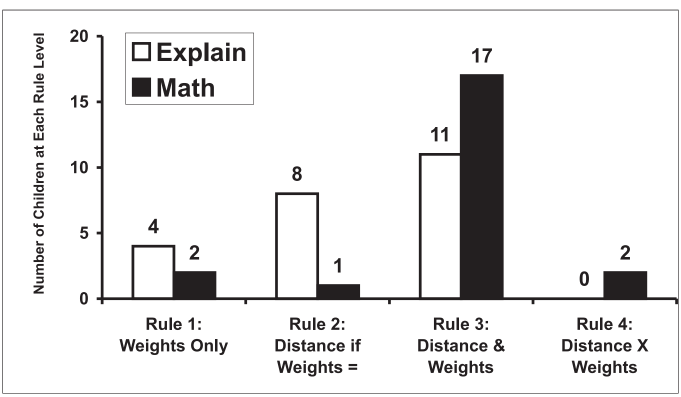

FIGURE 6 The frequency of rule attainment by condition (9-year-olds, Exp. 3).

view. This permitted us to differentiate three levels of sophistication on the transfer problems. Children considered only one dimension across transfer problems (math = 36%, explain = 70%); two dimensions across problems (math = 18%, explain = 9%), or two dimensions within the same problem as indicated by explanations that incorporated both dimensions (math = 45%, explain = 22%). The predicted advantage of the math condition is reliable, $\chi^2(1, N = 45) = 5.1$, $p = <.05$. When limiting the analysis to children who achieved Rule 3 or higher on the original problems (math $n = 19$, explain $n = 11$), 68% of the math children considered both dimensions compared to 54% of the explain children. The difference is not significant.

As in Experiment 2 children in both conditions made good efforts with the task and feedback, and they were attentive to both weights and distances, at least across problems in the learning phase. The children had similar error rates during learning, math ($M = 6.2$, $SD = 2.3$), explain ($M = 5.7$, $SD = 2.2$). All but one child entered meaningful justifications for each of the problems. Finally, 73% of the math and 83% of the explain children considered both dimensions across problems.

The telling differences between the conditions were similar to Experiment 2. Although not significant, 50% of the math children explicitly considered all four parameters of weight and distance at least once within a justification compared to 30% of the explain children. The math children also entertained an additional 1.2 ($SD = 1.6$) operations beyond their first solution, whereas the explain children considered none; $F(1, 43) = 13.5$, $MSE = 1.25$. Every single explain child used a qualitative subtraction structure for every single answer they justified.

Discussion

Conditions that facilitated the application of mathematics led to improved learning about the balance scale, even when there were no special instructions or examples. The math children were more likely to consider weight and distance simultaneously than the explain children, even though few math children had a successful procedure for combining those dimensions of information. In this light, the math helped the children develop a superior “qualitative” understanding. The justifications during learning again suggested two benefits of math: (a) a 20% gain in children who considered both dimensions simultaneously and (b) a twofold increase in the number of structures entertained for organizing those dimensions.

GENERAL DISCUSSION

Empirical Summary

Three studies demonstrated the value of number and mathematics for 9- to 11-year-old children’s development of physical understanding of the balance scale. Table 2 shows the relative performance of children from each condition

TABLE 2
Experimental Results Merged Into Siegler's (1981) Original Findings

		Percentages of Children at Each Rule Level				
Condition	*Age*	*Rule 0*	*Rule 1*	*Rule 2*	*Rule 3*	*Rule 4*
—	5 years	5%	85%	5%	5%	—
Beakers	10 years	—	68%	22%	5%	5%
Pegs	10 years	—	29%	43%	14%	14%
—	8 years	—	10%	35%	45%	10%
Explain	9 years	—	17%	35%	48%	—
Explain	11 years	—	7%	20%	70%	3%
—	12 years	10%	—	15%	60%	15%
Math	9 years	—	9%	5%	77%	9%
—	Adult	—		5%	65%	30%
Math	11 years	—	4%	—	45%	51%

Note. Labelled conditions introduce results from Experiments 1 to 3. At Rule 0, children guess. At Rule 1, children only pay attention to weight. At Rule 2, children pay attention to distance, when weights are equal. At Rule 3, children always pay attention to weight and distance but perform at chance when both distances and weights differ. At Rule 4, children solve all problems correctly.

when incorporated into Siegler's (1981) original norms. In Experiment 1, 10-year-old children who reasoned about hard-to-measure, continuous quantities in the beaker condition performed close to the level of 5-year-olds. Children who completed a standard peg and weight version of the task reasoned closer to their age-norms. In Experiment 2, 11-year-old children received feedback on the accuracy of their predictions across multiple problems. The math children saw how to measure the system and were encouraged to use math to justify their answers. They developed an understanding superior to many adults. Children in the explain condition received equivalent feedback but no instructions for how to measure the system or use math in their justifications. Their resultant understanding of the system was below that of the math children. In Experiment 3, 9-year-old children did not receive any instruction for how to measure the system or justify their answers, but otherwise replicated the protocol of Experiment 2. Children in the math condition again exceeded their age norms, though they did not learn to solve the problem mathematically. The math children considered weight and distance simultaneously more frequently than the explain children. The math children in Experiments 2 and 3 did not simply find a way to "push numbers." When the device changed to include weights on three pegs, the math children adapted their reasoning and continued to exhibit superiority over children in the explain conditions. In combination, these results support the claim that mathematics can propel the development of physical understanding in children.

Numbers and mathematical operations offered several helpful representational features. In Experiment 1, the children in the beaker condition could have used ordinal relations (i.e., more and less) to reason at the same level achieved by the children in the peg condition but they did not. Our a priori hypothesis was that the continuous quantities of the beaker condition would be difficult to measure, which in turn, would make it difficult to think simultaneously about weight and distance. In addition to reducing working memory demands (representing the digit "3" seems easier than representing three discrete weights in an image), numbers might make it easier to relate distance and weight by converting them to the same symbolic ontology. In support of this latter hypothesis, the justifications of the math children in Experiments 2 and 3 considered weight and distance simultaneously, whereas the justifications of the explain children tended to switch between weight and distance across justifications.

Mathematics also provided well-formed operators that served as candidate methods for organizing and combining the magnitudes of the problems. Experiments 2 and 3 revealed that children who used math explored more ways to combine the magnitudes. Children who explained in words perseverated on the qualitative comparatives *more*, *less*, *closer*, and *farther*.

A next step is to evaluate the highlighted properties of mathematics experimentally. Experiments 2 and 3 revealed that children in the math conditions considered more problem parameters simultaneously and explored more structuring operations but the experiments did not prove that these behaviors were responsible for the superior performance. Given the benefit of mathematics, it is now appropriate to compare different math treatments. For example, to evaluate the role of the external notations of mathematics, all children might receive instructions to use mathematics but only half of the children would be allowed to write down their mathematics.

Learning in Development

There are many situations where people can apply proportional reasoning (e.g., densities, concentrations, fairness, etc.). The regularity of proportional situations, however, does not entail that children can learn or develop equally from all situations. Some situations, for example, may have simpler perceptual structures. Similarly, some situations may include cultural supports that help organize children's thinking. Cross-cultural work has demonstrated that cognitive development can vary according to specific cultural settings where instituted symbols and their associated interpretations occur (e.g., Cole, 1996; Geary, Bow-Thomas, Liu, & Siegler, 1996; Peng & Nisbett, 1999; Reed & Lave, 1979). This research took an experimental approach. It used homogeneous samples of children and hindered or facilitated their application of arithmetic. The goal was to demonstrate and begin to characterize the import of one cultural creation—mathematics—in propelling children's development on a standard benchmark task of physical understanding.

An issue for views of development that emphasize culturally mediated change is the generality of a given cognitive change. If specific situations give rise to learning and development, then the resulting knowledge is unlikely to generalize spontaneously beyond those specific situations. It seems unlikely that the children in Experiment 2 had a cognitive reorganization that ushered in universal proportional reasoning. Instead, the children developed a particular, culturally mediated organization of knowledge—a proportion—that happens to have general application. It is like learning the concept of variability. The concept is narrow but its range of application is broad and it changes the way people can reason about very many situations. By this account, we should not expect the successful children in Experiment 2 to exhibit spontaneous proportional reasoning whenever appropriate. Instead, the continued development of proportional reasoning is likely to be characterized by imperfect and poorly generalized applications.

A useful line of subsequent research would add an additional phase to the latter two experiments. In the appended phase, children would work with new proportional reasoning tasks such as predicting whether ratios of water and orange concentrate taste the same or different. Our hypothesis is that children from the explain and math groups would initially look the same. However, over time and with feedback, the children who had used math for the balance scale would learn the proportional structure of the orange juice task more readily than the children who never used math. They would be more prepared to transfer and learn from a new situation (Bransford & Schwartz, 1999).

CONCLUSION

Three studies showed that mathematics affects children's development of physical understanding and they began to reveal why mathematics has a positive effect. However, because we orchestrated a specific situation to promote cognitive development, there is a question of whether the results are ecologically valid (Kuhn, 1974). One response is that the application of mathematics in everyday activities is common, ranging from counting (Ginsburg, 1982) to sports (Nasir, 2001). A second response is that the results become ecologically valid to the extent that we can institute instructional settings that propel development. Based on these results, it seems that experiences with relevant symbol systems should be quite important, in addition to experiences with the physical situation itself. For example, in a pilot study, we asked half of the children to replicate the math condition from Experiment 3. The other half of the children replicated the explain condition, except they also worked with a physical balance scale. Afterward, we provided a brief classroom lecture on how to solve the original balance scale problem plus the transfer problem with weights on three pegs. Assessments of child problem solving revealed that the math children learned more from the lecture and developed a more complex physical understanding. As fits our general story, providing children opportunities to explore the cultural representations of mathematics, and not just the physical phenomenon itself, can propel the development of physical understanding.

ACKNOWLEDGMENTS

We thank Giyoo Hatano for his wisdom and Gail DeBellis for her care.

This report is based on work supported by the National Science Foundation under BCS Grant 0214548 and REC Grant 0196238. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Ahl, V. A., Moore, C. F., & Dixon, J. A. (1992). Development of intuitive and numerical proportional reasoning. *Cognitive Development, 7,* 81–108.

Baillargeon, R. (1994). How do infants learn about the physical world*? Psychological Science, 3,* 133–140.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61–101.

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.

Case, R. (1985). *Intellectual development: Birth to adulthood.* New York: Academic.

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development, 61* (1–2, Serial No. 246).

Cole, M. (1996) *Cultural psychology: A once and future discipline*. Cambridge, MA: Harvard University Press.

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

diSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Gentner (Eds.), *Mental models* (pp. 15–33). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dixon, J. A., & Moore, C. F. (1996). The developmental role of intuitive principles in choosing mathematical strategies. *Developmental Psychology, 32,* 241–253.

Feynman, R. (1965). *The character of physical law.* Cambridge, MA: MIT Press.

Geary, K. C., Bow-Thomas, C. C., Liu, F., & Siegler, R. S. (1996). Development of arithmetical competencies in Chinese and American children: Influence of age, language and schooling. *Child Development, 67,* 2022–2044.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Ginsburg, H. P. (1982). *Children's arithmetic: How they learn it and how you teach it*. Austin, TX: PRO-ED.

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Huttenlocher, J., Jordan, N. C., & Levine, S. C. (1994). A mental model for early arithmetic. *Journal of Experimental Psychology: General, 123,* 284–296.

Kuhn, D. (1974). Inducing development experimentally: Comments on a research paradigm. *Developmental Psychology, 10,* 590–600.

Metz, K. E. (1993). Preschooler's developing knowledge of the pan balance: From new representation to transformed problem solving. *Cognition & Instruction, 11*, 31–93.

Nasir, N. (2001). "Points ain't everything": Emergent goals and average and percent understanding in the play of basketball among African-American students. *Anthropology and Education Quarterly, 31,* 283–305.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist*, *54,* 741–754.

Piaget, J. (1954). *The construction of reality in the child* (M. Cook, Trans.). New York: Basic Books.

Reed, H. J., & Lave, J. (1979). Arithmetic as a tool for investigating the relations between culture and cognition. *American Ethnologist, 6*, 568–582.

Resnick, L. B. (1992). From protoquantities to operators: Building mathematical competence on a foundation of everyday knowledge. In G. Leinhardt, R. T. Putnam, & R. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 373–429). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Saxe, G. B. (2001). Studying cognitive development in sociocultural context: The development of a practice-based approach. In A. Colby & D. Jessor (Eds.), *Ethnographic approaches to the study of human development* (pp. 275–303). Chicago: University of Chicago Press.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology, 30,* 154–219.

Schwartz, D. L., & Moore, J. L. (1998). The role of mathematics in explaining the material world: Mental models for proportional reasoning. *Cognitive Science, 22,* 471–516.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*(2, Serial No. 189).

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55,* 1233–1243.

Spinillo, A. G., & Bryant, P. (1991). Children's proportional judgments: The importance of "half." *Child Development, 62,* 427–440.

Starkey, P. (1992). The early development of numerical reasoning. *Cognition, 43,* 93–126.

Stavy, R., & Tirosh, D. (2000). *How students (mis-)understand science and mathematics.* New York: Teachers College Press.

Surber, C. F., & Haines, B. A. (1987). The growth of proportional reasoning: Methodological issues. *Annals of Child Development, 4*, 35–87.

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

White, B. (1993). ThinkerTools: Causal models, conceptual change and science education. *Cognition and Instruction, 10*, 1–100.