Journal of Educational Psychology
2014, Vol. 106, No. 2, 403–418

0022-0663/14/$12.00 DOI: 10.1037/a0034098

Applying Grounded Coordination Challenges to Concrete Learning Materials: A Study of Number Line Estimation

Jonathan M. Vitale, John B. Black, and Michael I. Swart
Teachers College, Columbia University

Do concrete learning materials promote strong learning outcomes, or do they simply make learning tasks more initially accessible? Although concrete materials may offer an intuitive foothold on a topic, research on desirable difficulties suggests that more challenging tasks facilitate greater retention and transfer. In the approach introduced here, grounded coordination challenges (GCCs) are embedded into the design of concrete learning materials to deliberately increase the difficulty of the learning task. More specifically, these challenges are intended to promote a deliberative process of mapping between perceptual elements of the materials. In 2 experiments the GCC approach was tested in a number line estimation task by comparing training with an "incongruent ruler"—which was designed to mismatch the length of an on-screen number line—to a "congruent ruler" (both experiments), or no ruler (the 1st experiment only). In both cases participants with the incongruent ruler were more likely to transfer knowledge to spatially transformed number lines. These results indicate that desirable difficulties facilitate learning in mathematical activities. Furthermore, the difficulties should emphasize a deliberate coordination process between critical features of the learning tool and the task. Implications for the design of learning activities that balance instructional support with conceptual challenge are discussed.

Keywords: math, estimation, desirable difficulties, instructional technology, manipulatives

Concrete learning materials—such as measurement tools, physical manipulatives, and pictorial representations—are widely accepted means of engaging young children in complex mathematics. Concrete features of learning materials include physical properties, and more broadly, any feature that helps the learner establish a link between the learning artifact and his or her own knowledge (Clements, 2000). On the basis of the work of developmental psychologists, such as Piaget (1954, 1970) and Bruner (1966), researchers focused on progressive educational reform have advocated, nearly universally, for the increased inclusion of concrete materials in educational settings—at times, with unrealistic expectation of their value (Ball, 1992; Clements, 2000).

In particular, although the overall effectiveness of curriculum incorporating manipulatives is generally positive (Sowell, 1989), a number of researchers have noted that a mixed empirical record suggests a need for deeper critical analysis (McNeil & Uttal, 2009). Unfortunately, many manipulatives prove ineffective for numerous practical reasons inherent in their design (Mix, 2009), thus making studies that compare manipulative-based and traditional curriculum difficult to interpret. Although manipulatives may be useful on the whole, some intuitive features may interfere with learning. For example, recent research suggests that materials with high realism may interfere with knowledge transfer when compared to more schematic materials (Kaminski, Sloutsky, & Heckler, 2008; Son & Goldstone, 2009).

Similarly, *ease of coordination* is another overassumed and underexamined dimension of learning materials. Specifically, the process of learning to effectively apply a concrete material to a task requires an initial orientation process in which the learner coordinates between features of the material and the goals of the activity. The ease of coordination can vary from almost negligible difficulty (e.g., learning to play a game with a well-designed touch interface) to overwhelming for the learner (e.g., learning to parallel park a car on a busy street).

Common sense, supported by research on cognitive load (Sweller, 1988), suggests that tools that coordinate easily with the given task should be beneficial. Yet, perhaps nonintuitively, there are benefits to more challenging, explicit coordination processes (Schwartz, Varma, & Martin, 2008). For example, Martin and Schwartz (2005) compared children learning to depict a set of fraction arithmetic problems with either a diverse set of pie pieces or a uniform set of square tile pieces. Whereas pie pieces inherently convey the meaning of the numerator and the denominator of a fraction (e.g., a semicircle is [1/2]), tile pieces represent fractions

This article was published Online First August 19, 2013.

Jonathan M. Vitale, John B. Black, and Michael I. Swart, Department of Human Development, Teachers College, Columbia University.

Jonathan M. Vitale is now at the Graduate School of Education, University of California, Berkeley.

This research was conducted as part of the first author's doctoral dissertation requirements for Teachers College, Columbia University, and was supported by a Ben D. Wood Fellowship and funding from the Institute for Learning Technologies. We would like to thank Herbert Ginsburg, Matthew Johnson, Koleen McCrink, and Sandra Okita for serving on the author's dissertation committee; Robert Siegler, Edward Hubbard, and Susan Lowes for guidance; and Genevieve Hartmann and Lisa Caswell for feedback. Finally, we wish to thank the students and dedicated staff of after-school programs conducted at PS 115, CS 154, and PS 161, particularly Grace Valera for years of support and collaboration.

Correspondence concerning this article should be addressed to Jonathan M. Vitale, 4407 Tolman Hall #1670, Graduate School of Education, University of California, Berkeley, CA 94720. E-mail: jonvitale@berkeley.edu

by relationships with sets of other tile pieces. Although both groups were able to complete the exercises, the more challenging tile condition promoted greater transfer to novel problems. In other words, for children working with tiles, the additional challenge of coordinating between the tool (tile pieces) and the task promoted a more robust conceptual representation.

More generally, in learning studies where some feature of the overall context interferes with training performance, posttest outcomes often improve. This positive association between the degree of difficulty in learning tasks and learning outcomes—that is, "desirable difficulties"—is a common finding in motor and verbal memory tasks (Bjork, 1994). For example, rather than organizing learning tasks into blocks of highly similar activities, interleaving dissimilar skills or concepts—such as multiple forms of motion in a motor learning task (T. D. Lee & Magill, 1983) or perceptual categories in a visual learning task (Dwyer, Hodder, & Honey, 2004; Lavis & Mitchell, 2006)—often leads to better long-term outcomes at a cost to short-term performance.

Additionally, desirable difficulties may facilitate development of higher level, conceptual knowledge (Bjork & Linn, 2006). In particular, appropriately placed challenges can draw a learners' attention to critical features of the materials. On the other hand, intuitive materials may foster "deceptive clarity," in which undemanding interactions are misinterpreted by learners as comprehension, dissuading further reflection (Linn, Chang, Chiu, Zhang, & McElhaney, 2011). For example, in a study of the use of outlines for text comprehension, Mannes and Kintsch (1987) found that participants more frequently solved inference problems correctly when provided with an outline that was organized inconsistently with the structure of the target text than when provided a consistent outline. In this case, the additional challenge of mapping between the outline and text promoted attention to the text's gist rather than its details.

In Mannes and Kintsch's (1987) example, the seemingly incongruent materials likely promoted a more explicit coordination process between features of the text. Likewise, in visuospatial domains, explicitly comparing and contrasting features provides learners with the opportunity to differentiate between superficial characteristics and distinguishing features (Gibson, 1969; Goldstone, Landy, & Son, 2010). Evidence from the field of grounded or embodied cognition (Barsalou, 2008) suggests that this form of perceptual learning may then facilitate development of higher level knowledge by grounding concepts in spatial features of the materials (Black, Segal, Vitale, & Fadjo, 2012; Goldstone et al., 2010).

Grounded Coordination Challenges

Although the visuospatial properties of concrete learning materials may provide an intuitive foothold for grounding concepts, these properties may unintentionally interfere with learning by reducing desirable difficulties. To address the need for both challenging activities and grounded representations, we introduce a novel instructional mechanism: *grounded coordination challenges* (GCCs). Whereas standard concrete materials are designed to elicit fluent coordination between artifact and task, GCCs are intended to interfere with fluent coordination to promote a more explicit process of mapping between critical features of materials and the context. As a result, learners will develop knowledge that is perceptually grounded in features that are applicable beyond the learned context.

More specifically, highly intuitive or familiar materials encourage learners to form implicit associations between features and their application in the given context—some of which are productive beyond the specific task and some of which are not. For example, pie pieces used by Martin and Schwartz (2005) map fraction values to specific, characteristic appearances (e.g., 1 is a circle, [1/2] is a semicircle, etc.). Although these ready-made, tacit associations make concrete materials easy to employ in a classroom—that is, less instruction is necessary—they may inhibit transfer by reinforcing misleading associations along irrelevant dimensions.

In contrast, GCCs are introduced into a task by presenting representations that either deliberately omit or exaggerate the variability of irrelevant features in a manner that learners may find initially challenging. As an example of the former, Martin and Schwartz's (2005) tile pieces did not depict their values inherently by appearance, thereby omitting a contextually useful but limiting feature of the more familiar pie pieces. In terms of the latter, varying the appearance of stimuli across nondistinguishing features is a common technique in perceptual category learning (Goldstone, 1998).

Given this overview, we make the following two hypotheses about what features are necessary to produce a successful GCC:

> *Hypothesis 1:* The task must present inherent difficulties, initially.
>
> *Hypothesis 2:* The materials must be designed to eliminate or interfere with a salient but misleading feature, while drawing attention to a distinguishing feature.

Although neither of these assertions is novel, taken together they represent a unique approach to the design of concrete material-based learning activities that deliberately eschews intuitive features. Although this approach is domain-general, we chose to apply it as a test case to a task that has garnered a great deal of attention in recent years—number line estimation. In the following we introduce the task and then describe how the GCC approach was applied in two experiments.

Number Line Estimation

Recent research suggests that the strength of an individual's number sense has far-reaching consequences on general mathematical ability (Booth & Siegler, 2006; Halberda, Mazzocco, & Feigenson, 2008; Holloway & Ansari, 2009). Interventions targeting number sense often result in rapid, robust improvements to a range of mathematical competencies (Opfer & Siegler, 2007; Thompson & Opfer, 2010), particularly with low-socioeconomic-status populations (Ramani & Siegler, 2008; Siegler & Ramani, 2008). Several of these interventions focus on strengthening children's understanding of the linear number line—a central conceptual structure in mathematical thinking (Case & Okamoto, 1996). Nonnormative mental representations of the number line are associated with younger children and children with low academic achievement. In particular, these children's estimates tend to display a characteristic logarithmic spacing, such that differences between small magnitudes are exaggerated and differences be-

tween large numbers are reduced (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Siegler & Opfer, 2003). Nonetheless, intervention may produce a qualitative shift toward a normative, linear representation in as little as one trial (Opfer & Siegler, 2007). Therefore targeting number line estimation for training may be an efficient means of producing rapid gains in a fundamental ability.

Moreover, the high malleability of children's conceptual representations affords testing of instructional approaches. For example, Thompson and Opfer (2010) applied "progressive alignment" (Kotovsky & Gentner, 1996) by providing explicit visual analogies to encourage transfer of knowledge about magnitudes from known to novel numerical scales. For example, a problem requiring the estimation of 15 on a 0–100 scale was displayed next to a problem of estimating 1,500 on a 0–10,000 scale, along with features that highlighted the congruency of the leading digits.

Likewise, with an aim at developing children's initial representation of the number line, Siegler and Ramani (Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009) successfully applied a simple numerical board game, for roughly 1 hr, to improve preschool children's estimation and arithmetic skills. In this game children counted along as they moved a token either one or two spaces—as determined by a spinner—across a demarked path, thereby promoting coordination between numerical value (the count) and spatial magnitude.

Introducing Grounded Coordination Challenges to Number Line Estimation

Siegler and Ramani (Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009) demonstrated that concrete learning tools can have a substantial impact on critical mathematical abilities. However, their specific approach—that is, to engage children in counting through all values of the numerical scale—is perhaps only applicable for early learners working within a narrow numerical range. Scaling up their approach to larger magnitudes likely requires additional measures to ensure an appropriate challenge.

Learning to estimate at large scales likely does not necessitate equal attention to each integer value within the scale. Adults often utilize landmarks, based upon common proportions (e.g., 50%), as a strategic means of estimating with large numerical scales. In the case of fractions (Siegler, Thompson, & Schneider, 2011) and angles (Vitale, Black, Carson, & Chang, 2010), older children and adults appear to use landmarks as a basis for their estimation strategy (e.g., [1/2] and 90°, respectively). Additionally, landmarks may play a role in simpler estimation tasks. For example, Siegler and Opfer (2003) found significantly less variability (i.e., greater precision) around quartile magnitudes of the 0–1,000 scale.

Even within an alternative model of magnitude representation, landmarks play a central role in mature representations of the number line. Specifically, Barth and colleagues (Barth & Paladino, 2011; Barth, Slusser, Cohen, & Paladino, 2011) fit both (seemingly) logarithmic and linear patterns of estimates by adjusting parameters of a single cyclical power function. In this model landmarks represent breakpoints between repeated cycles of the power function, such that with greater cycles the overall fit between estimated magnitudes and actual magnitudes appears increasingly linear.

Given the potentially critical role for quartile landmarks in estimation, training children to recognize and utilize these magnitudes holds a great deal of educational value. In light of our focus on concrete materials, a ruler depicting these landmark values would be a natural choice for a learning tool. Potentially, through practice applying a quartile-depicting ruler with a number line, children could learn to utilize these landmarks outside the training context.

However, Levine, Kwon, Huttenlocher, Ratliff, and Deitz (2009) observed that children often apply rulers mechanically, without deeper reflection. Specifically, when shown an image of a ruler whose 0 was unaligned with the left edge of a measured object, children often mistakenly read off the ruler's value at the right edge of the object. Once again, the intuitive affordances of the learning tool interfered with learners' deeper comprehension of the task. To counter this tendency, Levine et al. successfully trained children to apply an appropriate strategy that accounted for both edges of the measured object along the ruler.

In the experiments that follow, we apply a similar approach to Levine et al. (2009). To impede learners from performing ruler operations mechanistically, we manufactured a ruler, demarked with landmark values, but scaled 33% longer than the number line displayed on a computer screen. We expected that although this incongruent ruler would make the task initially challenging, by drawing attention to the common spatial proportions of landmarks on both representations, children would develop a conception of the number line that was grounded in more robust features of the materials.

Experiment 1

In this first study, we investigated several features of GCCs in number line estimation. As detailed above, we made two specific hypotheses regarding the necessary features of successful GCCs. The first—the task must present inherent difficulties, initially—emphasizes the role that desirable difficulties have on learning. To test this we compared the use of a standard, congruent ruler (CR)—which matched the length of on-screen number line and displayed landmark values at quartiles—to two more difficult conditions. In the no-ruler (NR) condition children performed the estimation task without the assistance of a ruler. In the incongruent ruler (IR) condition children were given a landmark-depicting ruler that did not match the length of the on-screen number line.

The second hypothesis—the materials must be designed to eliminate or interfere with a salient but misleading perceptual feature, while drawing attention to a distinguishing feature—offers a precise interpretation of which difficult features of the task are, in fact, desirable. The incongruent ruler was designed to specifically interfere with the association that children might make between the absolute position of landmark depictions on the ruler and their location on the target number line. For example, with no ruler a child might associate the value 45 with an absolute position approximately 7.5 cm from 0 (on a 30-cm line). On the other hand, given different distances between 45 and 0 on the displayed number line and the IR, a representation based on absolute distance should be less likely to emerge. To test this hypothesis, we compare the IR condition to the NR condition. Although both are considered difficult, only the IR condition fully implements our GCC approach.

In assessing the relative values of these conditions, we chose to focus on efficiency in training and transfer to similar tasks, which are important considerations in authentic learning environments. We predicted that the CR condition should elicit rapid fluency, reducing the need for extended training. In comparison, the NR condition should require more trials as learners engage in an extended trial-and-error process. The IR condition should require more training trials than CR, but less than NR.

So that we could assess learning outcomes, participants estimated on a nearly identical number line as in the training task (i.e., the "standard display"), as well as a series of spatially transformed number lines, without access to a ruler. With the standard display, we predicted that learners whose strategy relied heavily on the presence of a ruler would show decreased performance once this tool was removed. Children in the CR condition, whose ruler would likely afford rapid success, would display the least accuracy during posttest estimation, whereas children in the NR condition would be less likely to show a performance decrease because of the similarity between the learning and testing task.

Following the standard display, subjects began estimation on transformed number lines to test whether the knowledge gained over training was bound to the particular spatial characteristics of the standard number line, or whether this knowledge could be transferred to novel but related displays. Although we expected that both the IR and NR conditions would show an advantage over the CR condition (Hypothesis 1), the particular affordances of the IR condition should facilitate better transfer than the NR condition (Hypothesis 2).

To test these hypotheses, we chose to work with middle elementary school children (second to fourth grade) who would be both familiar with a wide range of numbers and capable of utilizing a reasonably complex cognitive strategy. Earlier work has demonstrated that children of this age typically are adept at estimating within the 0–100 scale but are developing mature estimation patterns of higher scales (Siegler & Booth, 2004). Although children may learn to estimate over higher scales that terminate in powers of 10 (e.g., 0–10,000) by focusing on numerical analogies between leading digits of the target scale and 0–100 (Siegler et al., 2009; Thompson & Opfer, 2010), in this study we chose to obscure these relationships by introducing the 0–180 scale, whose quartile values (45, 90, and 135) were not yet likely to be familiar or particularly meaningful to children.

Method

Participants. Participants included 80 second-, third-, and fourth-grade students. Children were gathered from two organizations within a large city whose services included daily after-school programs that extended into summer day camp during July and August. Seventy-five children were recruited from one after-school/summer school program hosted at a public school serving a predominantly low-income, Hispanic population. The remaining five children were recruited from a second day camp hosted at another public school serving a predominantly low-income, Hispanic and African American population.

The NR condition included 27 children ($M = 8.7$ years, $SD = 0.79$; 44% female, 96% Hispanic, 4% African American), the CR condition included 27 children ($M = 8.7$ years, $SD = 0.86$; 48% female, 93% Hispanic, 7% African American), and the IR condition included 26 children ($M = 8.5$ years, $SD = 0.72$; 58% female, 92% Hispanic, 8% African American). Two students who initially began the study (one NR, one IR) were unable to complete the training. In both cases the children made little progress and asked to be excused from participation.

Experimental design. Participants were assigned to a condition using a stratified random assignment procedure. Specifically, triads of children from each grade level were randomly assigned to each of three conditions, ensuring that each grade level had a roughly equal number of participants in each condition. The study was conducted without a numerical estimation pretest to avoid "proactive interference" (Opfer & Thompson, 2008), in which extended practice applying an inappropriate representation without feedback inhibits children from abandoning nonnormative representations during training. Similar posttest-only designs have been applied in prior studies of number line training (e.g., Opfer & Siegler, 2007). Children performed the learning task until they reached criterion of eight out of eight correct trials (error $< 10\%$) in a single block, and then proceeded immediately to the posttest. The training duration varied from approximately 5 to 30 min. The testing duration was approximately 10 min.

Materials and procedure.

Standardized measures. To ensure that children from each condition had similar levels of mathematical achievement, the Woodcock–Johnson III Calculation and Mathematical Fluency subtests (Woodcock, McGrew, & Mather, 2001) were administered in small groups of mixed-condition participants in a quiet room. These assessments were chosen because of their previous association with numerical estimation ability (Halberda, Mazzocco, & Feigenson, 2008).

Number line estimation game. Training was administered one to one in either a private room or a private area of a large room. The child was placed at a desk with a computer, while the experimenter sat to the child's side to provide assistance. Both testing and training was completed on a Dell laptop with a 17-in. (43.18-cm) monitor. All software was authored in the Adobe Flash CS4 environment.

At the start of training, children viewed a short animated instructional sequence, which provided the narrative context of the game—that is, fishing on a lake. Following the introduction, the child was presented with a side view of a lake, with a number line drawn across its surface (30 cm). The target magnitude was printed at the top center of the screen with the text "Catch a fish at [target] feet." During the estimation trial, contextual elements of the display faded to direct children's attention to the task (see Figure 1).

During the first trial, the experimenter engaged in condition-specific instruction. In the NR condition the children were told (approximately):

> The boat starts here at 0. The lake is 180 feet long. The fish is somewhere between 0 and 180, but you have to guess where. The words at the top of the screen give you a hint. You have to think about where that number is between 0 and 180. Can you find the fish?

Following this explanation, the children were free to estimate at a self-directed pace.

In the CR condition the children received a similar explanation; however, following initial instruction the child was presented with a ruler. In this case the number line printed on the ruler was identical in width (30 cm) to the number line displayed on-screen.

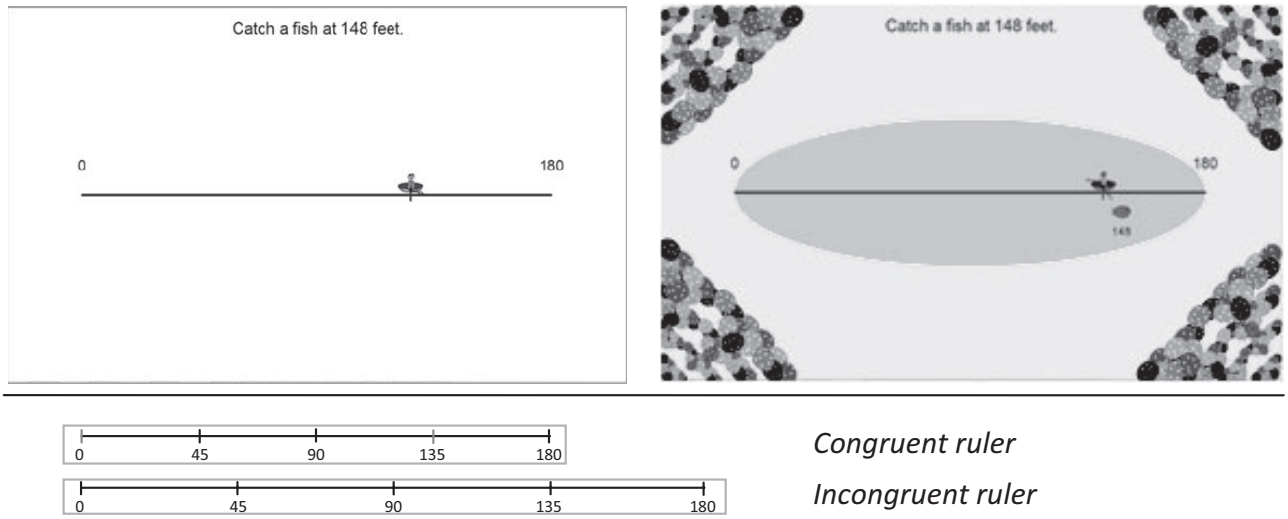

Figure 1. Estimation trial screenshots. Left side shows trial in progress with contextual features removed. Right side shows estimation trial during feedback with contextual features resumed. On the following trial, contextual elements again faded after 1 s. Below the left screenshot a rendering of the congruent ruler and incongruent ruler is displayed.

On the ruler, hatch marks were accurately placed along with associated values at 0, 45, 90, 135, and 180. The child was shown how the length of the ruler matched the length of the on-screen number line. The relationship between spatial position and numerical magnitude was explained by stating that "90 is in the middle, 45 is in the middle between 0 and 90, and 135 is in the middle between 90 and 180." The child was asked to demonstrate understanding by indicating the locations of 90, 45, and 135 on the number line displayed on-screen. If the child did not correctly align the ruler in this effort, the experimenter assisted in alignment. Following instruction, the child played the game at a self-directed pace. In some cases, when the child appeared fatigued, the experimenter offered to hold the ruler or place the ruler at the bottom of the screen, in alignment with the target number line.

In the IR condition the children received the same explanation as above and were given a ruler, which was described similarly as in the CR condition (i.e., "90 in the middle . . ."). However, in this case the ruler was constructed 33% larger than the number line on-screen (i.e., 40 cm). To alert children to this discrepancy, the experimenter told the child that the ruler was "mistakenly" made too large, and then placed the ruler on the screen to reveal the difference between lengths. The experimenter then asked the child to indicate the locations of 90, 45, and 135 on the number line displayed on-screen. If the child pointed to visibly inaccurate magnitudes, the experimenter would alert the student to the mistake (e.g., "no, 90 is not there"), and repeat the explanation of the relationship between spatial position and numerical magnitude on the ruler (i.e., "90 is in the middle . . ."). Following instruction, the child played the game at a self-directed pace. In some cases children were reminded to use the ruler during training trials.

During training the child performed a series of eight estimations in a block. Each target magnitude was sampled, at random, from one of eight subintervals of the number line (1–22, 23–45, 46–68, 68–90, 91–113, 113–135, 136–157, and 158–180). Each block consisted of one sample from each subinterval to ensure a diverse distribution of targets. During each trial the child navigated the "boat," via horizontal movement of the mouse, and pressed the mouse button to set the final estimate. If the selected magnitude was within 10% (3 cm) of the correct magnitude, the child was rewarded with an animation of a character catching the fish. If the selected magnitude fell out of the 10% margin of error, the actual location of the fish was displayed, and the child was told to click on the fish to proceed. Because of the boat's automatic placement at 0 at the start of each trial, in those cases where the (hidden) mouse cursor was not also at 0 at the start of the trial, the boat would appear to jump to a position along the number line to match the mouse cursor. Although potentially disorienting, pilot test users appeared to adapt easily to this unintentional feature of the software.

Following a block of eight trials, an animation provided a brief respite as well as summative feedback. In this animation, the child viewed the number of swimming fish that he or she "caught" in the prior block. If the child was able to catch all eight fish, training was completed. Otherwise, the child began a new block of eight trials with an altered background and type of fish. The decision to provide a relatively strict training criterion (all eight in a block) was intended to ensure familiarity with magnitudes distributed across the number line. Pilot testing of materials revealed that nearly all children were capable of achieving criterion.

Number line estimation posttest. Following successful completion of the training game, children immediately began the computerized posttest, consisting of four subtests of estimation trials over four spatially distinct number lines (see Figure 2). Each subtest consisted of 19 trials, whose target magnitudes were sampled from 16 equal subintervals of the 0–180 range. Additionally, landmark values of 45, 90, and 135 were included, resulting in the set {5, 16, 31, 36, 45, 49, 58, 70, 81, 90, 94, 106, 120, 131, 135, 140, 155, 161, 178}. The software randomly sorted this set of targets at the start of each subtest.

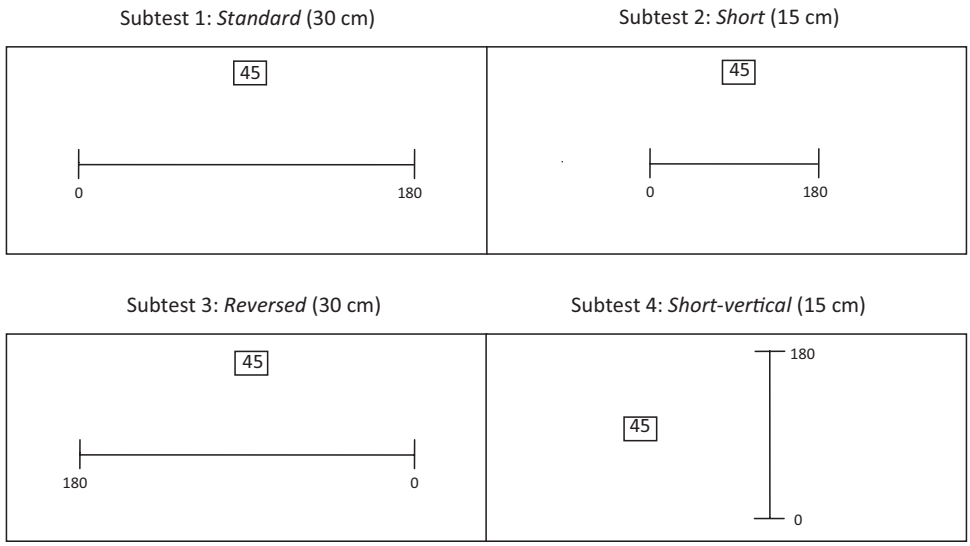

Figure 2. Subtest layouts. Each layout displays the orientation and length of each of the four postsubtests, as well as the location of the target value (45 in each case).

To initiate a trial, the participant clicked upon a green triangle located at 0. Upon clicking, the triangle was replaced with a sliding blue hatch mark that could be directed along the length of the number line by moving the mouse congruently. When the participant was ready to provide a final estimate, he or she pressed upon the mouse button until the hatch mark changed color. A 0.5-s pressing action was required, instead of a click, to avoid unintended final estimates.

If the child was satisfied with his or her final estimate, the experimenter pressed the space bar to continue to the next trial. If the child immediately recognized a mistake, the experimenter pressed the Delete key to place the trial back in the randomized queue of subtest trials. In some cases, if the child appeared to accidentally press the mouse button or seemed to be inattentive, he or she would be asked, "Is that where you wanted to put it?" If the child replied "no," the experimenter reset the trial back into the queue, otherwise the child continued to the next trial. No feedback was provided by the software. The experimenter provided only general support, such as "You're doing great, keep it up."

Verbal bisection probes. Lastly, upon completing all four subtests, the child was asked to verbally state the midpoint, first quartile, and third quartile value of the 0–180 range. The experimenter stated, "On all of those number lines 0 was on one side and 180 was on the other. What number would go right in the middle?" After answering the question, correctly or incorrectly, the experiment stated, "Imagine that we had a number line that goes from 0 on one side to 90 on the other. What number would go in the middle?" Finally, the latter question would be repeated in the context of a number line ranging from 90 to 180. Answers were recorded on a computer spreadsheet.

Results

Standardized measures. NR participants received a mean standardized score, grade-normed, of 95.7 (SD = 12.2) on Math Fluency and 101.4 (SD = 10.3) on Calculation. CR participants received a mean standardized score, grade-normed, of 97.0 on Math Fluency (SD = 11.7) and 104.7 on Calculation (SD = 9.6). IR participants received a mean standardized score, grade-normed, of 102.2 (SD = 9.6) on Math Fluency and 107.5 (SD = 9.2) on Calculation. An analysis of variance (ANOVA) did not indicate a significant difference between conditions for Math Fluency, $F(2, 77) = 1.0$, $p > .10$. However, a trend toward a significant difference in Calculation was found, $F(2, 77) = 2.4$, $p = .10$, which we applied as a covariate in several following analyses.

Number line estimation game. According to our first hypothesis, a successful application of learning materials requires some initial difficulty. Therefore the primary goal in analyzing learning task performance was to validate that the NR and IR conditions were indeed more difficult than the CR condition. As general proxies for difficulty, we used the mean total number of blocks to reach criterion and duration of training (see Figure 3).

However, to determine the locus of this difficulty, we split the training results into the first trial and all trials thereafter. The first trial included the introduction to supplementary materials, when applicable (rulers in CR and IR), and the child's attempt to coordinate these materials with the digital display. Figure 3A depicts the large differences between conditions, $F(2, 77) = 45.9$, $p < .001$, $\eta_p^2 = .54$. Differences between CR and IR conditions, $t(51) = 4.5$, $p < .001$, with nearly identical instructional scripts, suggest that the IR materials significantly increased the difficulty of initial coordination.

For all subsequent trials, differences in total duration (see Figure 3B) confound variability in difficulty (i.e., more trials-to-criterion require more time) with differences in strategy, and as such revealed no significant differences between conditions, $F(2, 77) = 2.1$, $p > .10$, $\eta_p^2 = .05$; however, a more precise indicator of general difficulty, mean total number of blocks (see Figure 3C), did reveal significant differences between conditions, $F(2, 77) = 13.5$, $p < .001$, $\eta_p^2 = .26$.

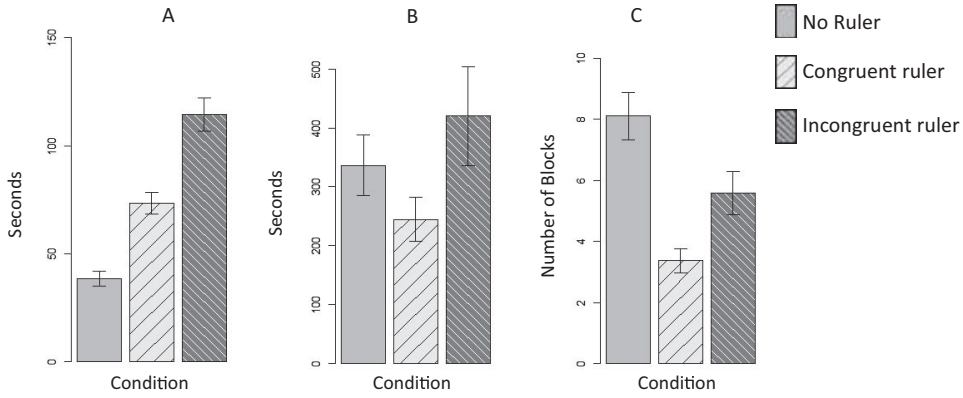

Figure 3. Summary training variables: the mean number of seconds to complete the first trial (A), the mean number of seconds to complete training starting with Trial 2 (B), and the number of blocks needed to reach criterion (C). Error bars indicate standard errors.

According to planned comparisons, CR participants required significantly fewer blocks to reach criterion than participants in the other two conditions, $F(1, 77) = 19.1$, $p < .001$, $\eta_p^2 = .25$. Likewise, IR participants required fewer total blocks than NR participants, $F(1, 77) = 7.5$, $p = .008$, $\eta_p^2 = .10$. These results suggest that the NR condition was the most challenging, whereas the CR condition was the easiest.

Although we did not measure strategy application directly, participants who engaged in a more deliberative, explicit strategy would likely spend more time on a trial. Because all participants engaged in all eight trials of each block, total durations of blocks could be compared to highlight broad differences in strategy. A summary of durations for Blocks 1–4 is displayed in Figure 4.

For the first block, starting with the second trial, a significant difference between conditions in duration emerged, $F(2, 77) = 8.1$, $p = .001$, $\eta_p^2 = .17$. Bonferroni-adjusted post hoc comparisons revealed that NR participants completed the block in less time than either CR, $t(52) = 3.4$, $p < .01$, or IR participants, $t(51) = 4.5$, $p < .001$; however, there was no difference between CR and IR participants, $t(51) = 0.2$, $p > .10$. On the second block, the general effect of condition persisted, $F(2, 69) = 7.4$, $p = .01$, $\eta_p^2 = .17$, as well as differences between NR and both CR, $t(52) = 3.3$, $p < .01$, and IR, $t(51) = 3.7$, $p < .001$. By the third and fourth blocks, these differences were no longer apparent: Block 3, $F(2, 58) = 0.9$, $p > .10$, $\eta_p^2 = .03$; Block 4, $F(2, 58) = 1.8$, $p > .10$, $\eta_p^2 = .07$. Although this shift in significance reinforces the premise that the IR condition created initial difficulties for students, this lack of an effect in latter blocks may reflect the loss of high-performing students who had reached criterion.

Number line estimation posttest. As in previous studies of number line estimation (e.g., Ramani & Siegler, 2008), we applied three measures of accuracy: linearity, slope, and mean percent absolute error (PAE). Linearity refers to the amount of variance explained by the best fitting linear function of estimated magnitudes to actual magnitudes. Slope refers to the slope of that linear regression line. Finally, PAE refers to the mean difference between estimated and actual magnitudes as a percentage of the maximum of the range (180; see Figure 5).

Figure 6 displays the relationship between condition and subtest across all three outcome measures. To separate the effects of training on displays that were spatially similar or dissimilar to the training display, we chose to separate the analysis of the first

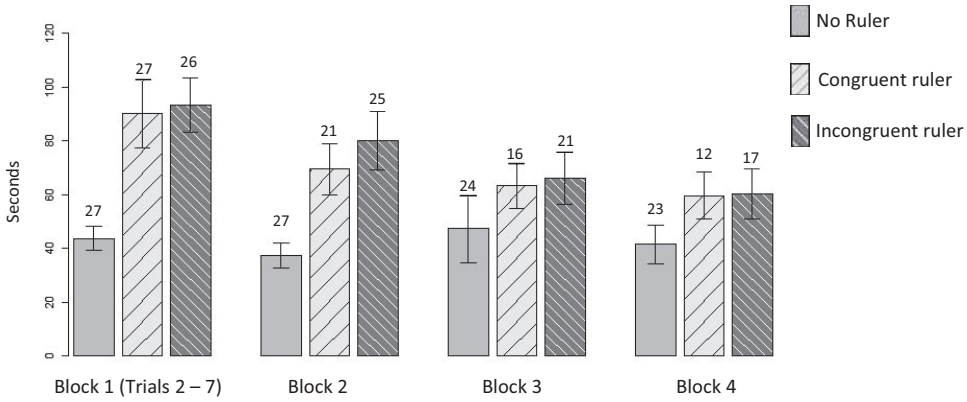

Figure 4. The mean duration (in seconds) of subject in the first four blocks in training. For the first block the first trial is not included, to eliminate the influence of instructional time. Counts above each bar represent the number of children who remained in the experiment at the given block. Error bars indicate standard errors.

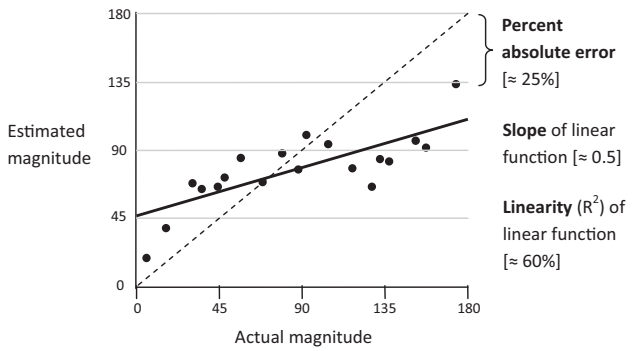

Figure 5. Example data for an individual subtest. Displays percent absolute error for a single data point, slope, and linearity of best fitting linear function.

subtest (i.e., the standard display) from the subsequent three subtests. To analyze the first subtest, for each of our three dependent measures, we performed a one-way analysis of covariance (ANCOVA) with condition as a between-subjects factor and age and Woodcock–Johnson (WJ) Calculation standardized score as covariates. Preliminary analyses did not reveal an interaction between condition and WJ Calculation for any dependent measure (mean PAE, $F(2, 74) = 0.6$; linearity, $F(2, 74) = 0.6$; slope, $F(2, 74) = 0.1$; $ps > .10$) or condition and age (mean PAE, $F(2, 74) = 0.6$; linearity, $F(2, 74) = 0.3$; slope, $F(2, 74) = 0.2$; $ps > .10$), upholding the parallel slopes assumption of ANCOVA.

Each of the three dependent measures analysis revealed significant effects of condition (mean PAE, $F(2, 75) = 6.3$, $p < .01$, $\eta_p^2 = .14$; linearity, $F(2, 75) = 7.4$, $p < .01$, $\eta_p^2 = .16$; slope, $F(2, 75) = 8.3$, $p < .01$, $\eta_p^2 = .18$), age (mean PAE, $F(1, 75) = 10.2$, $p < .01$, $\eta_p^2 = .12$; linearity, $F(1, 75) = 4.3$, $p < .05$, $\eta_p^2 = .05$; slope, $F(1, 76) = 4.6$, $p < .05$, $\eta_p^2 = .06$), and WJ Calculation (mean PAE, $F(1, 75) = 6.0$, $p < .05$, $\eta_p^2 = .07$; linearity, $F(1, 75) = 8.7$, $p < .01$, $\eta_p^2 = .10$; slope, $F(1, 75) = 10.1$, $p < .01$, $\eta_p^2 = .12$).

For the first subtest, a planned comparison between CR and the two "difficult" conditions revealed a significant difference for all three measures (mean PAE, $F(1, 75) = 10.2$, $p < .01$, $\eta_p^2 = .12$; linearity, $F(1, 75) = 14.0$, $p < .001$, $\eta_p^2 = .16$; slope, $F(1, 75) = 16.6$, $p < .001$, $\eta_p^2 = .18$). On the other hand, a comparison of IR and NR did not reveal a significant difference for any measure (mean PAE, $F(1, 75) = 1.9$, $p > .10$, $\eta_p^2 = .04$; linearity, $F(1, 75) = 0.7$, $p > .10$, $\eta_p^2 = .01$; slope, $F(1, 75) = 0.3$, $p > .10$, $\eta_p^2 = .00$).

To test the effect of condition on the latter three subtests, with spatially transformed displays, we performed repeated-measures ANCOVA with age and WJ Calculation standardized score as covariates. For each of the three dependent measures, the analysis revealed significant effects of condition (mean PAE, $F(2, 75) = 11.9$, $p < .001$, $\eta_p^2 = .24$; linearity, $F(2, 75) = 7.7$, $p < .01$, $\eta_p^2 = .17$; slope, $F(2, 75) = 7.2$, $p < .01$, $\eta_p^2 = .16$), age (mean PAE, $F(1, 75) = 17.4$, $p < .001$, $\eta_p^2 = .19$; linearity, $F(1, 75) = 8.9$, $p < .01$, $\eta_p^2 = .11$; slope, $F(1, 76) = 11.6$, $p < .01$, $\eta_p^2 = .13$), and WJ Calculation for mean PAE, $F(1, 75) = 7.5$, $p < .01$, $\eta_p^2 = .09$, and a trend toward a significant difference for the other measures (linearity, $F(1, 75) = 2.8$, $p < .10$, $\eta_p^2 = .04$; slope, $F(1, 75) = 3.4$,

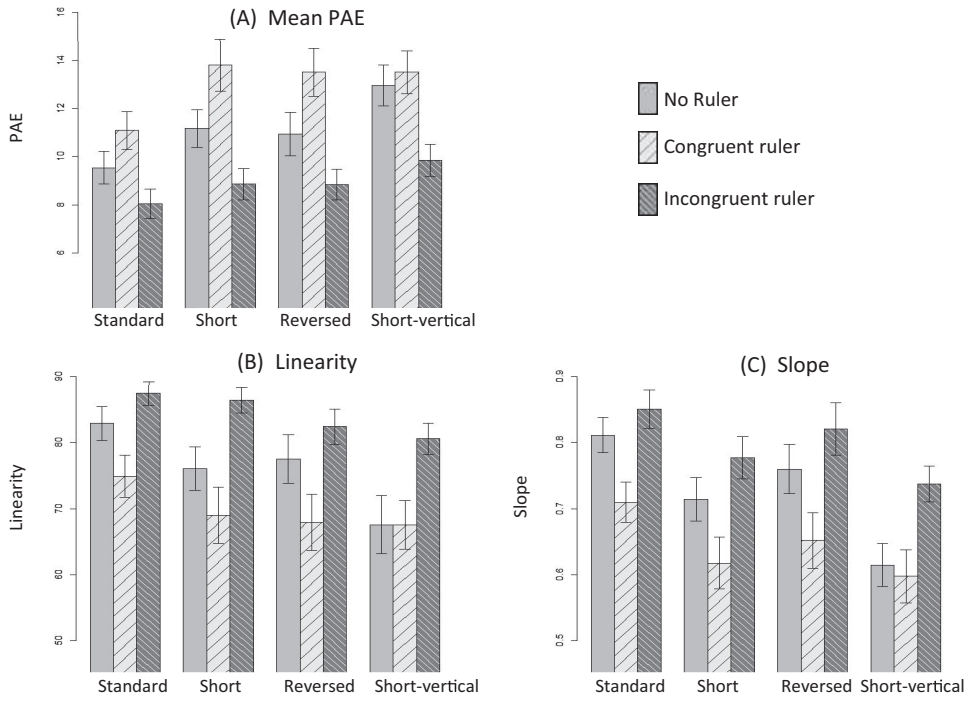

Figure 6. Posttest accuracy measures. Across all three outcome measures the congruent ruler shows the least accuracy (i.e., higher error, lower linearity, lower slope), and the incongruent ruler shows the highest accuracy. Error bars indicate standard errors. PAE = percent absolute error.

$p < .10$, $\eta_p^2 = .04$). Neither the effects of subtest, interactions between subtest and condition, nor interactions between condition and covariate measures were significant.

To test our specific hypotheses regarding the effect of condition on the spatially transformed displays, we performed the same planned comparisons as above. Our first comparison between CR and the two difficult conditions revealed a significant difference for all three measures (mean PAE, $F(1, 75) = 16.6$, $p < .001$, $\eta_p^2 = .19$; linearity, $F(1, 75) = 9.6$, $p < .01$, $\eta_p^2 = .12$; slope, $F(1, 75) = 10.6$, $p < .01$, $\eta_p^2 = .13$). Unlike the standard display, in this case a comparison of IR and NR did reveal a significant difference for two measures (mean PAE, $F(1, 75) = 5.1$, $p < .05$, $\eta_p^2 = .08$; linearity, $F(1, 75) = 4.7$, $p < .01$, $\eta_p^2 = .06$), and a trend toward a significant difference for slope, $F(1, 75) = 3.1$, $p < .10$, $\eta_p^2 = .04$. Therefore, although the effect of the type of difficulty inherent in IR and NR had little effect on the first subtest, which was nearly identical to the training display, the effect emerged with more novel displays.

Verbal bisection probes. Finally, with the bisection questions that followed estimation (see Table 1), three NR participants, 17 CR participants, and 23 IR participants correctly identified 90 as the midpoint of the 0–180 scale. Chi-square tests revealed an overall effect of condition, $\chi^2(2) = 33.3$, $p < .001$. Bonferroni-adjusted post hoc comparisons revealed a significant difference between NR and both CR, $\chi^2(2) = 13.4$, $p < .001$, and IR, $\chi^2(2) = 28.7$, $p < .001$. For the second question, one NR, seven CR, and 15 IR participants correctly identified 45 as the midpoint of the 0–90 scale. Chi-square tests revealed an overall effect of condition, $\chi^2(2) = 19.0$, $p < .001$. Post hoc comparisons revealed a significant difference between NR and IR only, $\chi^2(2) = 15.9$, $p < .001$. For the third question, no NR, five CR, and 12 IR participants correctly identified 135 as the midpoint of the 90–180 scale. Chi-square tests revealed an overall effect of condition, $\chi^2(2) = 17.0$, $p < .001$. Post hoc comparisons revealed a significant difference between NR and IR only, $\chi^2(2) = 13.6$, $p < .001$.

Discussion

Experiment 1 demonstrated that children in either difficult condition demonstrated greater learning outcomes than children whose task was more intuitive, thereby reinforcing our claim that learning tasks need to be sufficiently challenging while highlighting critical structural feature of the materials. Although this inverse relationship between learning task performance and outcomes aligns with research on desirable difficulties (Bjork, 1994), differences between the difficult conditions exhibited a more complex set of patterns. In particular, children with no ruler were less accurate than children with the incongruent ruler during both the training (i.e., required more blocks to reach criterion) and the spatially transformed postsubtests. Yet, although children with the incongruent ruler were more accurate, their pace during the initial two blocks was slower than children with no ruler.

Table 1
Frequency of Correct and Incorrect Responses to Verbal Bisection Probes (Experiment 1)

Landmark value	Correct?	No ruler	Congruent ruler	Incongruent ruler
90	Yes	3	17	23
	No	24	10	3
45	Yes	1	7	15
	No	26	20	11
135	Yes	0	5	12
	No	27	22	14

These results, as well as experimenter observation, suggest that NR and IR participants were performing fundamentally different strategies. Specifically, many children in the NR condition appeared to estimate based on recall of feedback from recent, proximal magnitudes. In part this may be due to the appearance of the boat jumping to the position of previous feedback at the onset of the next trial if the mouse remained static between trials. This strategy afforded a clear but ephemeral landmark to apply toward estimation. As evidence for this observation, we compared mean PAEs of the first trial of each block, where the most recent feedback occurred more than 10 s prior, to the average of the last seven trials of the block, where previous feedback was recent. Although IR and CR participants showed no differences between the first and seven last trials, NR participants showed significantly worse estimates for the first trial (not assuming equal variances), $t(44.3) = 2.8$, $p < .05$. The altered procedure in the posttest—that is, participants clicked on an object that always began at 0—may have disproportionally impacted NR participants who utilized previous trials' feedback during training.

Though potentially counterproductive in the long run, the use of recent feedback did promote some memory for magnitude locations, allowing NR participants to perform no worse than IR participants on the standard display subtest. However, we suspect that the recent feedback recall strategy focused attention on the absolute position of magnitudes on the number line and not their proportional location. As such, in subtests that disrupted the mapping between absolute position and magnitude, NR participants performed significantly worse than IR participants.

To some extent the performance of children with no ruler diverged from previous findings that demonstrate widespread improvements of the mental number line based on limited, targeted feedback (e.g., Opfer & Siegler, 2007). For Opfer and Siegler (2007) this rapid improvement represented a qualitative shift in representation in response to feedback at the point of largest deviation between logarithmic and linear representations. It may be the case here that the NR participants underwent a similar shift and applied the local feedback heuristic thereafter. As evidence of a rapid shift in response to feedback that maximized the difference between representations, NR participants who estimated a target magnitude between 12.5% and 25% of the scale within their first three training trials produced significantly less error on their fourth trial than participants who received feedback on this portion of the number line after their fourth trial, $t(22) = 3.0$, $p < .01$. This suggests that feedback alone did play some role in enhancing children's conception of the number line.

Yet, to the extent that NR condition was unsuccessful, in addition to the unintended consequences of feedback, it may be case that specific contextual features of the training task promoted a narrower focus and more short-sighted strategy, as has been found in several recent studies of authentic materials and narratively driven learning activities (Goldstone & Son, 2005; Son & Goldstone, 2009). For example, the fishing narrative may have encouraged a more episodic encoding of the numerical magnitudes.

Similarly, the individualized, video-game-like nature of the experience, unlike previous studies utilizing collaborative, board game settings (Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009), may have promoted a performance rather than mastery goal orientation (Dweck, 1986).

Even though these contextual features were present in the other conditions, given the presence of rulers, their role was less prominent. Specifically, because of instruction to locate landmarks prior to estimation, children applied a more time-consuming, deliberative strategy focusing on these magnitudes rather than secondary features of the context. As such, children in ruler conditions required more time to complete the first two blocks than children with no ruler. Yet, in spite of superficial similarities between strategies, children in the CR and IR conditions diverged in training and posttest accuracy. Furthermore these conditions differed markedly in the number of training trials, total duration of training, and duration of the first, instructional trial.

Although evidence suggests that the overall difficulty of the experience drove differences between these conditions, and not the number of training trials or duration per se, it may be the case that given equal time on task, similar posttest results would have emerged. Even though CR participants achieved mastery of the training task early, they may have continued to learn on subsequent trials through greater exposure. We address this concern directly in Experiment 2.

In spite of these reservations, we posit that coordination processes drove differences between conditions. Although the congruent ruler coordinated intuitively with the on-screen number line, it was precisely this ease of coordination that inhibited deeper reflection. In contrast, the IR condition elicited a more challenging process of coordinating between two mismatched representations. This additional difficulty accounts for IR participants' significantly longer first trial durations. It may even be the case that this initial coordination process was sufficient; however, greater training inaccuracy, compared to CR, suggests that additional experience was necessary to consolidate the new strategy.

In addition to being more difficult, the aim of the IR condition was to foster a proportional representation of landmarks (i.e., 90 is halfway from 0 to 180, 45 is one quarter, 135 is three quarters). An explicit understanding of these relationships could then be applied to any spatial transformation of the number line. Alternatively, the length mismatch may have simply fostered practice at mentally transforming (i.e., dilating) between representations of variable length—a skill that was directly applicable to the short display subtest and not the reversed display. However, no significant differences between the mean PAEs in these subtests, $t(25) = 0.02$, $p > .10$, suggest that IR participants did not rely on solely implicit, perceptual strategies.

Considering the possibility that IR participants did develop a proportional conception of landmarks on the number line, additional applications for transfer may be appropriate. For example, if these children were asked to estimate on a 0–90 number line, they may have recognized that 45, which was one quarter of 180, is the midpoint of 0–90, thereby discovering a significant landmark.

In beginning to address possibilities for transfer, we conducted an informal follow-up study by recruiting 36 former participants (14 NR, 11 CR, 11 IR) to retrain to criterion, with identical materials from their initial training, and then perform two estimation postsubtests on standard (30-cm) on-screen number lines. In the first, standard display, subtest participants estimated from 0 to 180, while in the second, "numerical transfer," subtest participants estimated from 0 to 90. Target magnitudes for 0–90 were derived by halving targets used for 0–180. We note that due to variability in the date of initial training, time elapsed between training sessions differed between subjects. However, this variability was not associated with condition, $F(2, 33) = 0.8$, $p > .10$, $\eta_p^2 = .05$. Likewise, although age was not controlled experimentally, no significant differences between conditions arose, $F(2, 33) = 1.8$, $p > .10$, $\eta_p^2 = .10$.

Even though this informal follow-up was not intended as a delayed posttest, some surprising patterns in training performance do suggest differences in retention by condition. As in the original experiment, large differences in blocks-to-criterion persisted, $F(2, 33) = 13.9$, $p < .001$, $\eta_p^2 = .46$. Although NR participants predictably required more blocks to reach criterion than either CR, $t(23) = 4.4$, $p < .001$, or IR participants, $t(23) = 3.5$, $p < .01$, there was no difference between the two ruler conditions, $t(20) = 1.1$, $p > .10$. Additionally, although NR participants showed no change in number of blocks to reach criterion from initial to follow-up training, $t(13) = 0.69$, $p > .10$, IR participants required significantly fewer blocks to achieve criterion at follow-up, $t(10) = 2.9$, $p < .05$. These results suggest that knowledge attained from the initial coordination process with the IR was retained between training sessions. In contrast, much of what had been learned by NR participants had to be relearned at follow-up.

Regarding follow-up posttests, no effect of condition on mean PAE emerged for the standard display, $F(2, 31) = 1.4$, $p > .10$, $\eta_p^2 = .08$, likely because of the reduced sample size. However, for the numerical transfer subtest a trend toward an effect of condition on mean PAE did emerge, $F(2, 31) = 2.6$, $p < .10$, $\eta_p^2 = .13$. Post hoc comparisons revealed a trend toward significantly lower mean PAE for IR than CR, $t(20) = 2.4$, $p < .10$. This result hints at the potential for IR materials to promote transfer to novel numerical scales by highlighting proportional relationships in a way that easily coordinated materials do not. In addition to studying the role of training duration, we formally apply this numerical transfer subtest in the second experiment.

Experiment 2

As discussed above, our interpretation of effects that contrast the two ruler conditions are attenuated by large differences in exposure to training stimuli. In this experiment we addressed this issue by constraining task duration. Task duration was chosen, instead of number of trials, because it is likely that children in the CR condition would complete the first, instructional trial more quickly than children in the IR condition, and therefore complete more trials overall in the same amount of time; thus ensuring that IR's total exposure to the training task, by either measure, did not exceed CR's.

To test our predictions efficiently, we chose to work with only second-grade students, who were less likely to have prior instruction with the 0–180 scale. Whereas in the previous experiment no significant interaction between grade and condition emerged, inspection of mean PAEs on the first subtest revealed nonsignificantly larger differences between conditions for second graders (CR: 13.1; IR: 9.1) than for fourth graders (CR: 7.2; IR: 6.0). This

suggests that second graders are equally or more sensitive to manipulation than older students.

Additionally, we examined the effect of condition on transfer to an estimation task with an alternative scale, 0–90, as introduced in the previous experiment. Because of the explicit attention to proportional features that we believe the incongruent ruler fosters, we predicted that these participants would demonstrate greater transfer than those in the CR condition.

Method

Participants. Participants included 30 second-grade students. Children were gathered from an after-school program within a large city serving primarily low-income Hispanic and African American populations. The CR condition included 15 children (M = 8.7 years, SD = 0.86; 48% female, 93% Hispanic, 7% African American). and the IR condition included 15 children (M = 8.5 years, SD = 0.72, 58% female, 92% Hispanic, 8% African American).

Materials and procedure.

Standardized measures. To ensure that children from each condition had similar levels of mathematical achievement, the Woodcock–Johnson III Calculation and Mathematical Fluency subtests were administered in small groups of mixed-condition students in a quiet room.

Number line estimation training game. Training was administered one to one in either a private room or a private area of a large room. The child was placed at a desk with a computer, while the experimenter sat to the child's side to provide assistance. The training software and physical materials (rulers) were identical to those used in Experiment 1.

However, unlike in Experiment 1, after completing the short animated instructional sequence—introducing children to the context and goals of the game—the administrator began a 15-min timer. The children then received the same condition-specific instruction as in Experiment 1 before proceeding to estimation trials. After 15 min had expired, the child was allowed to complete his or her current block of estimation trials before the learning task was halted by the administrator. This experimental feature was included to ensure that all children viewed block summary feedback prior to posttest. Although this did allow for some variability in total duration, similar block durations for IR and CR participants in Experiment 1 suggest that between-group differences would be unlikely to emerge.

Number line estimation posttest. As in Experiment 1, the first subtest of Experiment 2 contained a number line that was equivalent in length and orientation to the training number line. However, to produce a more abrupt spatial change, only the final number line display from Experiment 1 (i.e., "short vertical") was applied here to test transfer over spatial transformation. Finally, to extend the results of Experiment 1, a third subtest, spatially equivalent to the first, tested children's ability to estimate on a 0–90 scale with a target set: {3, 8, 16, 18, 23, 25, 29, 35, 42, 45, 47, 53, 60, 66, 68, 70, 78, 81, 89}. In this case, the children were explicitly alerted to the new scale and intermittently reminded throughout the subtest at the start of trials (e.g., "Remember that this line goes from 0 to 90"). The follow-up bisection questions from Experiment 1 remained the same in Experiment 2.

Results

Standardized measures. CR participants received a mean standardized score, grade-normed, of 100.9 on Math Fluency (SD = 15.2) and 101.2 on Calculation (SD = 8.4). IR participants received a mean standardized score, grade-normed, of 94.1 (SD = 12.2) on Math Fluency and 101.5 on Calculation (SD = 9.4). No significant differences between groups emerged for either subtest: Math Fluency, $t(28) = 1.3$, $p = .19$; Calculation, $t(28) = -0.10$, $p = .92$.

Number line estimation game. Although some variability between participants in total duration emerged, the average duration did not differ between conditions (CR: M = 562 s, SD = 115; IR: M = 582 s, SD = 79.2), $t(28) = 0.53$, $p = .60$. However, as expected, IR participants did spend more time on the first trial than CR participants (CR: M = 84 s, SD = 19; IR: M = 104 s, SD = 30), $t(28) = 2.25$, $p < .05$. Although this may have afforded CR participants more time to complete estimation trials, the total number of blocks performed did not differ significantly between conditions (CR: M = 7.5, SD = 2.0; IR: M = 6.7, SD = 1.5), $t(28) = 1.35$, $p = .19$.

Unlike in Experiment 1, in which some participants completed the training within a single block, in Experiment 2 all children completed at least four blocks (CR: minimum = 5, maximum = 11; IR: minimum = 4, maximum = 9). Thus, whereas in Experiment 1 we analyzed Blocks 1–4 separately because of differing number of participants in each, in Experiment 2 we analyzed the first four blocks of training with repeated-measures ANOVAs. To address accuracy, which could not be represented by total number of blocks (as in Experiment 1), we applied an ANOVA to the total number of correct estimates (i.e., error < 10%) per training block (see Figure 7), which revealed a significant effect of block, $F(3, 84) = 4.4$, $p = .007$, $\eta_p^2 = .14$; a trend toward a significant effect of condition, $F(1, 28) = 3.15$, $p = .09$, $\eta_p^2 = .10$; and no significant interaction between block and condition, $F(3, 84) = 0.73$, $p = .54$, $\eta_p^2 = .03$.

Number line estimation posttest. Figure 8 displays the relationship between condition and subtest across all three outcome measures. As in Experiment 1, in which analysis of the standard and spatially transformed subtests were performed separately, each of the three subtests for Experiment 2 were analyzed separately. Because we chose to use a single grade level, and little difference between groups emerged on standardized pretests, no covariates were included in these analyses.

For the standard display subtest, a one-way ANOVA revealed a significant difference between conditions for all three measures (mean PAE, $F(1, 28) = 8.6$, $p < .01$, $\eta_p^2 = .23$; linearity, $F(1, 28) = 5.2$, $p < .05$, $\eta_p^2 = .16$; slope, $F(1, 28) = 5.4$, $p < .05$, $\eta_p^2 = .16$). For the short-vertical display, ANOVA revealed a significant difference between conditions for two measures (mean PAE, $F(1, 28) = 6.2$, $p < .05$, $\eta_p^2 = .18$; linearity, $F(1, 28) = 7.4$, $p < .05$, $\eta_p^2 = .21$) and a trend toward a significant difference between conditions for slope, $F(1, 28) = 4.0$, $p < .10$, $\eta_p^2 = .13$. Finally, for the numerically transformed subtest, a trend toward a significant difference between conditions emerged for mean PAE, $F(1, 28) = 3.0$, $p < .10$, $\eta_p^2 = .10$, whereas no significant difference emerged between conditions for the other two measures (linearity, $F(1, 28) = 1.2$, $p > .10$, $\eta_p^2 = .04$; slope, $F(1, 28) = 0.2$, $p > .10$, $\eta_p^2 = .01$).

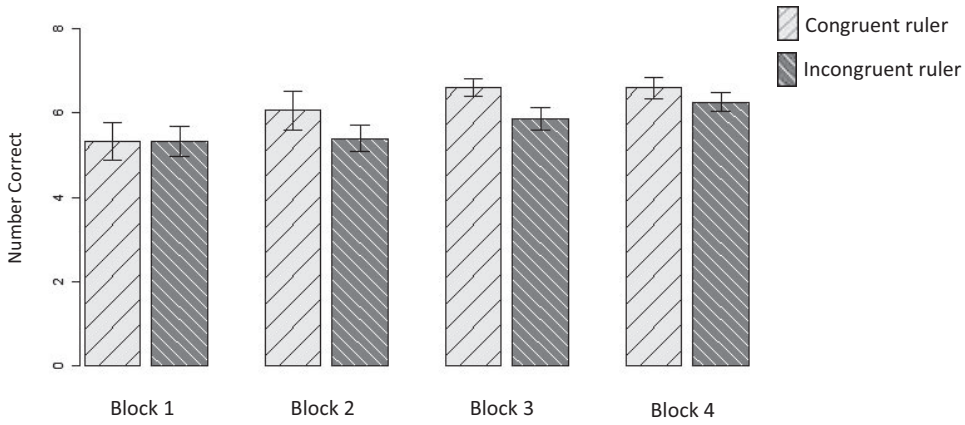

Figure 7. The mean number of correct trials in the first four blocks (i.e., "fish caught"). The number of participants from each condition is the same across these first floor blocks (15 each). Error bars indicate standard errors.

Verbal bisection probes. Finally, in Experiment 1, although children in either ruler condition were more likely to answer bisection questions correctly than NR participants, a small advantage for IR over CR trended toward significance. Yet, in Experiment 2 the groups differed nonsignificantly estimating the midpoint of 90, $\chi^2(1) = 0.17$, $p = .68$, and not at all for estimating 45 or 135 (see Table 2).

Discussion

Large differences between conditions in both the standard and spatially transformed subtests confirm that the differences between CR and IR conditions seen in Experiment 1 were not due to differences in training exposure. Rather, these effects were due to the nature of the manipulation. By impeding the physical coordination of the ruler to the on-screen number line, the additional challenge afforded by the incongruent ruler produced stronger posttest outcomes.

As in Experiment 1, the difficulty inherent in the IR condition emerged in the extended duration of the first training trial, where verbal scripts were similar but the materials differed in their ease of coordination. Additionally, a trend toward significance emerged while comparing conditions on the average number correct in the first four blocks. Although this effect did not meet our criterion for significance, a moderate effect size ($\eta_p^2 = .10$) suggests that the task was somewhat more difficult, beyond the first trial, for children in the IR condition.

In regard to the posttest, whereas the first two subtests elicited predicted results, the subtest with a novel numerical scale, 0–90, elicited mixed results. Only a trend toward significance for mean PAE emerged between conditions. Although we expected IR participants to recognize 45 as the midpoint of the scale and apply this landmark in their estimation strategy, the experimenter observed very few students doing so explicitly. Considering that participants appeared to use the 45 landmark during training, what prevented them from applying 45 as the midpoint of the 0–90 scale?

One likely possibility is that children simply misunderstood the spatial significance of these quarter landmarks. Although the spatial relevance of 90 as midpoint was clear, children may have simply viewed the 45 and 135 landmarks as "somewhere" between 90 and the respective endpoint, rather than as specifically midway. In both conditions many children persisted in estimating 45 on the 0–90 scale near the quarter point of the number line. In several cases this produced an unexpected exponential distribution of estimates (i.e., crowding of smaller magnitudes, spacing of larger magnitudes).

Additionally, whereas two thirds successfully identified 90 as the midpoint in the verbal bisection probes, less than a quarter identified 45 or 135 as midway between halves. Informally, several children with incorrect estimates could successfully recall the numbers printed on the ruler, when prompted informally following the study. Thus, it may be the case that with slightly older children, the spatial significance of these values, and their application to other scales, would be more apparent.

These results suggest that although our coordination challenge was successful in helping children understand the role of the midpoint, it was less successful in promoting an understanding of other landmarks. Nonetheless, considering the trend toward differences between conditions of moderate effect size ($\eta_p^2 = .10$) on the final subtest, the IR condition may have afforded some benefit to this novel scale.

General Discussion

Concrete materials can provide a bridge between children's intuitions, prior experiences, and complex mathematics. Yet, developing strong conceptual understanding remains a difficult process, with or without concrete materials. In contrast to some current pessimism regarding the value of concrete materials (e.g., Kaminski, Sloutsky, & Heckler, 2009), the current study suggests that by generating desirable difficulties in the context of concrete materials-based instruction, GCCs may facilitate learning. Specifically, by designing tools and activities that interfere with the intuitive but misleading application of materials, we can challenge students to make use of appropriate features and develop more robust conceptual representations.

Considering the limited resources in the classroom (i.e., time, space, and money), the design of concrete materials requires

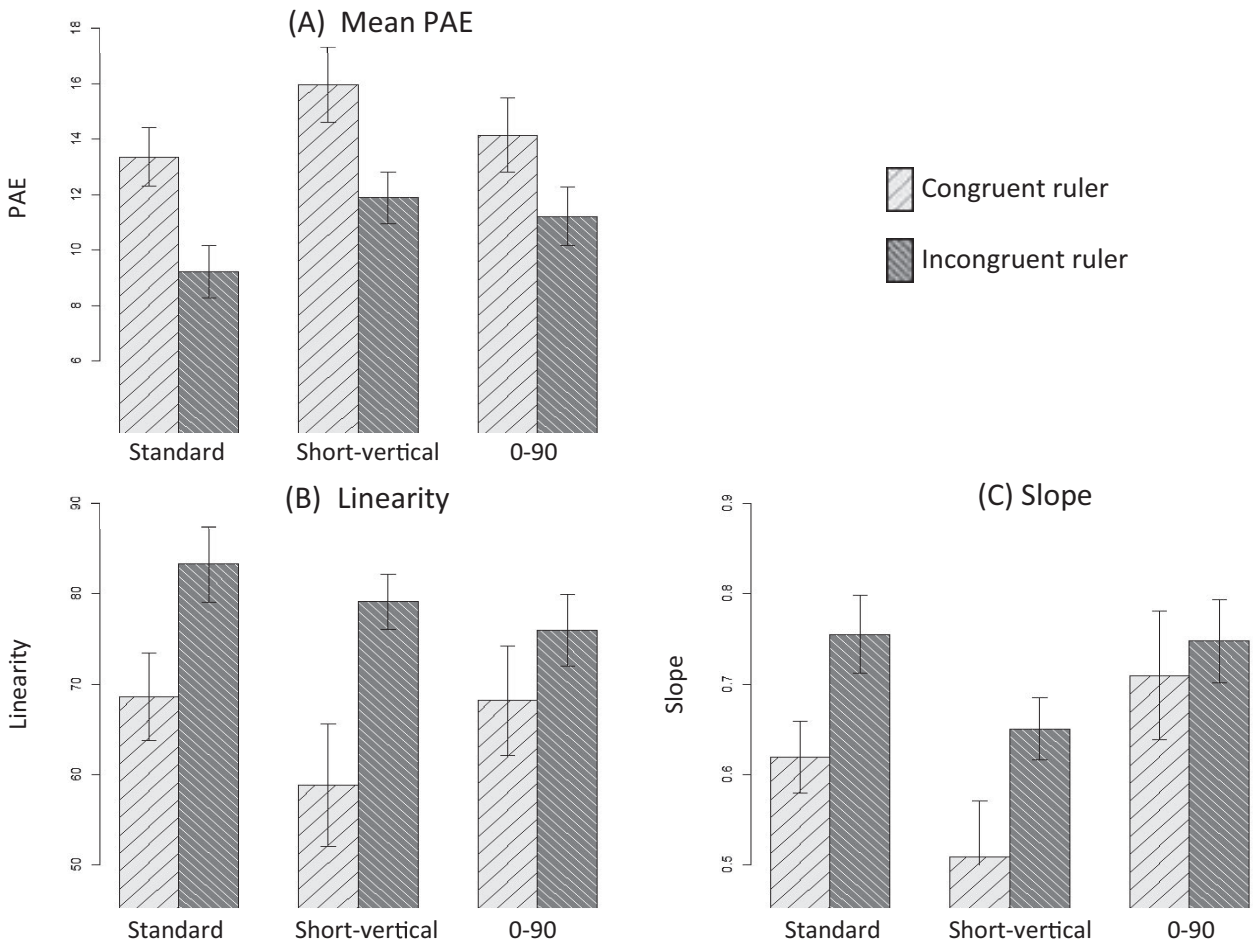

Figure 8. Across all three outcome measures, the congruent ruler shows the lower accuracy (i.e., higher error, lower linearity, lower slope), and the incongruent ruler shows the highest accuracy. Error bars indicate standard errors. PAE = percent absolute error.

careful determination of which features to target for interference and which to highlight for the learner. In the introduction we discussed why we chose to highlight proportional landmarks in the incongruent ruler. Below we discuss the degree to which this implementation of GCCs was successful. Additionally, we summarize our experience and suggest means of implementing GCCs in related activities.

Table 2
Frequency of Correct and Incorrect Responses to Verbal Bisection Probes (Experiment 2)

Landmark value	Correct?	Congruent ruler	Incongruent ruler
90	Yes	10	12
	No	5	3
45	Yes	3	3
	No	12	12
135	Yes	1	1
	No	14	14

Promoting a Proportional Representation of the Number Line

While implementing GCCs in number line estimation, two important considerations emerged: (a) what feature of a number line estimation task may be misleading, and (b) what kind of visual depiction is necessary to guide learners toward a critical feature of the representation? As we predicted, children learning to estimate on a fixed number line would attend to the absolute position of magnitudes on the number line, which caused a large drop-off in performance for NR participants upon encountering the spatially transformed displays. On the other hand, the incongruent ruler promoted attention to the proportional relationship between landmarks and endpoints of the ruler, leading to relatively stronger performance during the spatially transformed displays.

Although the results favoring IR suggest that our overall approach was successful, we must inquire about whether this approach worked broadly across all magnitudes or only those landmark magnitudes that were depicted on the ruler. Since our

goal was to promote broad change, evidence of the latter would severely limit the value of the intervention. However, in support of our intervention, even with landmark trials removed (i.e., 45, 90, 13), IR participants showed significantly lower mean PAE than NR participants, $F(1, 51) = 4.9$, $p = .03$, $\eta_p^2 = .09$, across all subtests in the first experiment.

The development of a broad shift was also reflected in the observed behaviors and spontaneous utterances of the children. In many cases (in either ruler condition) children moved their finger or the mouse directly to the midpoint of the line before proceeding to the final estimate. In other cases, children may have simply envisioned the location of landmark values as they navigated the number line. As an IR participant stated, "I just imagine where 45, 90, and 135 is." Another child stated, "Forty-five is between 0 and 90, and 90 is in the middle," as she navigated the cursor according to these landmarks. Spontaneous utterances notwithstanding, the lack of a formal protocol for probing students' strategies is a limitation of this study and deserves greater attention in future research. Additionally, future studies utilizing eye- or gesture-tracking technology could also help uncover these strategies directly.

Although it is clear that the effect of the incongruent ruler promoted robust learning, there were some limitations to this approach. In particular, even though many children explicitly utilized the midpoint as a strategic reference for values in the midrange of the scale during the posttest, fewer children did the same for the first quartile landmark, and even fewer for the third quartile landmark. To generate these secondary landmarks, a student would need to locate the midpoint first and then further divide the halves into halves. Although the centered text could have facilitated estimation of the midpoint, no analogous reference could be applied to the other landmarks. Consequently, further instruction is likely necessary to assist children in discovering the spatial significance of these landmarks. In general, although these findings reveal a limitation to our specific implementation, they suggest that coordination challenges could be and should be an ongoing process that continuously prompts children to reflect on their understanding of the materials and how they should be applied.

Practical Implications

Generating an appropriately challenging educational activity is itself a challenging activity. Although some difficulties are desirable, challenges that are not germane to the task may overwhelm or mislead the learner (Sweller, 1988). As a relevant example of this, Siegler and Ramani (2009) found that those children trained to estimate on a circular number line did not produce similar gains as those children who were trained on a linear number line. Although estimating on the circular number line was, most likely, the more difficult of the two conditions, this difficulty was irrelevant to the target knowledge. Likewise, in our study, difficulties faced in the NR condition did not facilitate learning as much as the difficulties in the IR condition that were structured to target proportional concepts.

Beyond the materials featured here, what are characteristics of activities that incorporate GCCs? Two general instructional patterns include troubleshooting and guided construction. Troubleshooting tasks require that individuals assess a problem state in a system and devise a means of overcoming this problem. Here the challenge is to coordinate between a specific error and the general state and function of the system. Troubleshooting is a complex cognitive process that requires a great deal of conceptual and procedural expertise (Jonassen, 2000), making these tasks ideal for assessment of mathematical and scientific concepts, such as children's understanding of electrical circuits (Kester, Kirschner, & van Merriënboer, 2004).

Additionally, troubleshooting tasks can provide a fertile ground for learning about a system. For example, as programmers are often too familiar with, debugging faulty software often involves lengthy interpretation of large chunks of code. Debugging exercises are often applied in computer science education and prove to be an effective form of instruction (G. C. Lee & Wu, 1999). Generally, troubleshooting exercises provide learners with a highly specific problem that may require in-depth investigation.

In our case, by informing children that the incongruent ruler was mistakenly made too large, we were, in essence, asking them to troubleshoot a problem situation. The "trouble" could be fixed by mentally bisecting the on-screen number line to generate a landmark. In pilot testing children given the incongruent ruler without a troubleshooting frame often aligned zeros of the ruler and number line—a familiar behavior with rulers—which resulted in unaligned magnitudes beyond 0. Some children then used the misaligned landmarks on the ruler to estimate on the number line. By alerting the children to a "mistake," in this study, children were able to abandon an intuitive but incorrect strategy. Whether children would spontaneously abandon the incorrect strategy, in response to incorrect trials, without the troubleshooting framing is a possible avenue of future research.

Likewise, construction activities incorporate numerous challenges that require in-depth understanding of the system. Yet, although construction activities are well supported by theory (e.g., Papert's, 1980, "constructionism"), open-ended approaches to design and construction activities are often ineffective in producing efficient learning outcomes (Mayer, 2004) compared to more guided approaches. In many cases the myriad challenges facing a learner in an open construction activity simply overwhelm the learner's cognitive resources, allowing little room for deeper conceptual processing (Kirschner, Sweller, & Clark, 2006).

In the face of such difficulty, children may resort to the construction of highly prototypical artifacts, which, though valid, exhibit misleading superficial characteristics. For example, children attend to irrelevant features when identifying shapes. When asked to identify rectangles, young children may overlook squares and misidentify elongated (nonrectangular) parallelograms (Clements, Swaminathan, Hannibal, & Sarama, 1999). Asking children to construct rectangles (e.g., by drawing) is likely to elicit production of highly prototypical, elongated rectangles, further reinforcing this "skinny" conception.

Guided construction activities, on the other hand, may be an effective means of instruction by directing learners to the critical features of target concept. Task constraints can be utilized to promote the construction of novel and diverse artifacts, which share subtle distinguishing features. In the case of rectangle construction, a GCC could be implemented by interfering with children's ability to produce rectangles with prototypical aspect ratios (Vitale, Black, & Swart, 2011). For example, children could be asked to produce multiple rectangles where one pair of sides was

constrained to a specific length. Additionally, in the case of number line estimation, asking children to construct their rulers may confer benefits that exceed the incongruent ruler (e.g., comprehension of secondary landmarks).

In general, GCCs represent a balance between discovery-based approaches and direct instruction. Although the instructional designer chooses normative representations (e.g., proportional landmarks), the learner plays a central and active role in mapping out the representational structure of this concept. Although this approach may require more effort during the learning task—though less so than a pure discovery approach—the work detailed here suggests that the rewards in learning are well worth the effort. Furthermore, by successfully overcoming challenges, rather than performing repetitive trial and error or following a predetermined, rote procedure, students may be more engaged in the learning task.

In conclusion, concrete learning materials are valuable. However, the inherent accessibility of these tools offers mixed blessings. GCCs require designers of concrete materials and related activities to balance incorporation of features that ground knowledge intuitively with features that ensure sufficiently difficult to elicit deliberative, reflective thinking. With further research we hope to delineate ways to navigate this balance in the design of educational activities.

References

Ball, D. L. (1992). Magical hopes: Manipulatives and the reform of math education. *American Educator, 16,* 14–19.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617–645. doi:10.1146/annurev.psych.59.103006.093639

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science, 14,* 125–135. doi:10.1111/j.1467-7687.2010.00962.x

Barth, H. C., Slusser, E., Cohen, D., & Paladino, A. M. (2011). A sense of proportion: Commentary on Opfer, Siegler and Young. *Developmental Science, 14,* 1205–1206. doi:10.1111/j.1467-7687.2011.01081.x

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology, 46,* 545–551. doi:10.1037/a0017887

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A., & Linn, M. C. (2006). The science of learning and the learning of science: Introducing desirable difficulties. *APS Observer, 19,* 29–39.

Black, J. B., Segal, A., Vitale, J. M., & Fadjo, C. L. (2012). Embodied cognition and learning environment design. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments* (2nd ed., pp. 198–223). New York, NY: Routledge.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42,* 189–201. doi:10.1037/0012-1649.41.6.189

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

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). doi:10.2307/1166077

Clements, D. H. (2000). "Concrete" manipulatives, concrete ideas. *Contemporary Issues in Early Childhood, 1,* 45–60. doi:10.2304/ciec.2000.1.1.7

Clements, D. H., Swaminathan, S., Hannibal, M. A. Z., & Sarama, J. (1999). Young children's concept of shape. *Journal for Research in Mathematics Education, 30,* 192–212. doi:10.2307/749610

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040–1048. doi:10.1037/0003-066X.41.10.1040

Dwyer, D. M., Hodder, K. I., & Honey, R. C. (2004). Perceptual learning in humans: Roles of preexposure schedule, feedback, and discrimination assay. *Quarterly Journal of Experimental Psychology: Section B*. Comparative and Physiological Psychology, *57,* 245–259. doi:10.1080/02724990344000114

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York, NY: Appleton-Century-Crofts.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585–612. doi:10.1146/annurev.psych.49.1.585

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science, 2,* 265–284. doi:10.1111/j.1756-8765.2009.01055.x

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences, 14,* 69–110. doi:10.1207/s15327809jls1401_4

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with math achievement. *Nature, 455,* 665–668. doi:10.1038/nature07246

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103,* 17–29. doi:10.1016/j.jecp.2008.04.001

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48,* 63–85. doi:10.1007/BF02300500

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science, 320,* 454–455. doi:10.1126/science.1154659

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. (2009). Transfer of mathematical knowledge: The portability of generic instantiations. *Child Development Perspectives, 3,* 151–155. doi:10.1111/j.1750-8606.2009.00096.x

Kester, L., Kirschner, P. A., & van Merriënboer, J. J. G. (2004). Information presentation and troubleshooting in electrical circuits. *International Journal of Science Education, 26,* 239–256. doi:10.1080/69032000072809

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86. doi:10.1207/s15326985ep4102_1

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797–2822. doi:10.2307/1131753

Lavis, Y., & Mitchell, C. (2006). Effects of preexposure on stimulus discrimination: An investigation of the mechanisms responsible for human perceptual learning. *Quarterly Journal of Experimental Psychology, 59,* 2083–2101. doi:10.1080/17470210600705198

Lee, G. C., & Wu, J. C. (1999). Debug It: A debugging practicing system. *Computers & Education, 32,* 165–179. doi:10.1016/S0360-1315(98)00063-3

Lee, T. D., & Magill, R. A. (1983). The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 730–746. doi:10.1037/0278-7393.9.4.730

Levine, S. C., Kwon, M., Huttenlocher, J., Ratliff, K., & Deitz, K. (2009). Children's understanding of ruler measurement and units of measure: A training study. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2391–2395). Austin, TX: Cognitive Science Society.

Linn, M. C., Chang, H.-Y., Chiu, J. L., Zhang, Z. H., & McElhaney, K. (2011). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 235–258). New York, NY: Psychology Press.

Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text organization. *Cognition and Instruction, 4,* 91–115. doi:10.1207/s1532690xci0402_2

Martin, T., & Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science, 29,* 587–625. doi:10.1207/s15516709cog0000_15

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59,* 14–19. doi:10.1037/0003-066X.59.1.14

McNeil, N. M., & Uttal, D. H. (2009). Rethinking the use of concrete materials in learning: Perspectives from development and education. *Child Development Perspectives, 3,* 137–139. doi:10.1111/j.1750-8606.2009.00093.x

Mix, K. S. (2009). Spatial tools for mathematical thought. In K. S. Mix, L. B. Smith, & M. Gasser (Eds.), *Spatial foundations of language and cognition* (pp. 41–66). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199553242.003.0003

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55,* 169–195. doi:10.1016/j.cogpsych.2006.09.002

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development, 79,* 788–804. doi:10.1111/j.1467-8624.2008.01158.x

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY: Basic Books.

Piaget, J. (1954). *The construction of reality in the child* (M. Cook, Trans). New York, NY: Basic Books. doi:10.1037/11168-000

Piaget, J. (1970). *Science of education and the psychology of the child* (D. Coltman, Trans.) New York, NY: Orion Press.

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development, 79,* 375–394. doi:10.1111/j.1467-8624.2007.01131.x

Schwartz, D. L., Varma, S., & Martin, L. (2008). Dynamic transfer and innovation. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 479–506). New York, NY: Routledge.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75,* 428–444. doi:10.1111/j.1467-8624.2004.00684.x

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14,* 237–250. doi:10.1111/1467-9280.02438

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science, 11,* 655–661. doi:10.1111/j.1467-7687.2008.00714.x

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology, 101,* 545–560. doi:10.1037/a0014239

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273–296. doi:10.1016/j.cogpsych.2011.03.001

Son, J. Y., & Goldstone, R. L. (2009). Contextualization in perspective. *Cognition and Instruction, 27,* 51–89. doi:10.1080/07370000802584539

Sowell, E. J. (1989). Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education, 20,* 498–505. doi:10.2307/749423

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285. doi:10.1207/s15516709cog1202_4

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development, 81,* 1768–1786. doi:10.1111/j.1467-8624.2010.01509.x

Vitale, J. M., Black, J. B., Carson, E., & Chang, C. (2010). Development in the estimation of degree measure: Integrating analog and discrete representations. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2242–2247). Austin, TX: Cognitive Science Society.

Vitale, J. M., Black, J. B., & Swart, M. I. (2011). Promoting development of geometry concepts: Interfacing multiple embodied representations with a computer game. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2383–2388). Austin, TX: Cognitive Science Society.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Received November 14, 2012
Revision received July 11, 2013
Accepted July 15, 2013 ■