Cognition 180 (2018) 182–190

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Learning to measure through action and gesture: Children's prior knowledge matters

Eliza L. Congdon[a,b,*], Mee-Kyoung Kwon[a,c], Susan C. Levine[a]

[a] University of Chicago, United States
[b] Bucknell University, United States
[c] Ulsan National Institute of Science and Technology, South Korea

ARTICLE INFO

Keywords:
Cognitive development
Gesture
Learning
Spatial cognition
Mathematics
Manipulatives

ABSTRACT

Learning through physical action with mathematical manipulatives is an effective way to help children acquire new ideas and concepts. Gesture is a type of physical action, but it differs from other kinds of actions in that it does not involve interacting directly with external objects. As such, gesture provides an interesting comparison to action-on-objects and allows us to identify the circumstances under which gesture versus interaction with objects (and the associated effects on the external world) may be differentially beneficial to learning. In the current study, we ask whether individual differences in first grade children's prior knowledge about a foundational mathematical concept – their understanding of linear units of measure – might interact with their ability to glean insight from action- and gesture-based instruction. We find that the children using a more rudimentary pretest strategy did not benefit from producing gestures at all, but did benefit from producing actions. In contrast, children using a more conceptually advanced, though still incorrect, strategy at pretest learned from both actions and gestures. This interaction between conceptual knowledge and movement type (action or gesture) emphasizes the importance of considering individual differences in children's prior knowledge when assessing the efficacy of movement-based instruction.

1. Introduction

We know from decades of experimental psychology research that asking children to act directly on external representations can affect their internal ideas (e.g., Wilson, 2002; Sommerville & Woodward, 2010; James, 2010; Kontra, Goldin-Meadow, & Beilock, 2012; Gerson, Beckering, & Hunnius, 2015; Levine, Goldin-Meadow, Carlson, & Hemani, 2018). In fact, children succeed in solving many problems grounded in the physical world well before they can succeed with abstract, symbolic forms of parallel problems (Bruner, Olver, & Greenfield, 1966; Piaget, 1953). These findings suggest that acting on, or manipulating, objects is a powerful way for children to learn new ideas. Gestures – a special category of action – can represent information, engage the motor system, and reference external representations in an instructional context, but unlike actions-on-objects, gestures are representational and do not create lasting change in the external environment (Novack & Goldin-Meadow, 2017). Here, we directly compare hand gestures to actions-on-objects in a linear measurement lesson with first grade children to investigate whether these different kinds of actions might differentially affect children's understanding of spatial units of measure.

Previous research has identified both benefits and drawbacks of learning through action in math contexts. Using manipulatives, objects designed to represent abstract math concepts in a tangible, physical way is one of the most common ways that action-based learning is instantiated in elementary school math lessons. For example, young children may learn to add using blocks or other sets of small objects before they are able to add Arabic numerals (e.g., Levine, Jordan, & Huttenlocher, 1992). Acting with manipulatives allows children to offload cognition onto the environment and encourages the formation of useful conceptual metaphors (Manches & O'Malley, 2012). It also directs attention to the relevant components of a complex problem (Mix, 2010) and engages young learners with limited attention spans and working memory (Petersen & McNeil, 2008). Yet some recent research cautions against action-based learning, highlighting instances where children may become distracted by irrelevant components of the manipulatives such as color or texture (Petersen & McNeil, 2008), or may see the learned actions as relevant only to a specific set of objects rather than as instantiating a broader mathematical principle (e.g., Uttal, Scudder, & DeLoache, 1997; DeLoache, 2000; Kaminski,

* Corresponding author at: Department of Psychology, O'Leary Center, Bucknell University, Lewisburg, PA 17837, United States.
E-mail address: eliza.congdon@bucknell.edu (E.L. Congdon).

https://doi.org/10.1016/j.cognition.2018.07.002
Received 16 November 2016; Received in revised form 26 June 2018; Accepted 4 July 2018
Available online 26 July 2018

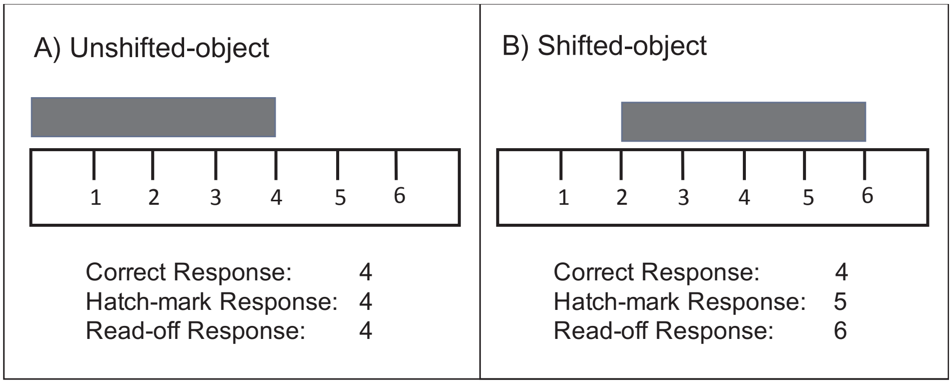

Fig. 1. Illustrations of two types of problem where a to-be-measured object is either (A) aligned with the start of the ruler or (B) shifted away from the start of the ruler. Common student responses are listed below each image.

Sloutsky, & Heckler, 2009).

Gestures differ from actions on manipulatives in that they do not require children to interact directly with physical objects and do not result in changes in the location or orientation of these objects. Importantly, research findings show that asking learners to gesture can promote learning, insight, and retention across a variety of domains including algebra, chemistry, word learning, and even moral reasoning (e.g. Wakefield & James, 2015; Macedonia, Muller, & Friederici, 2011; Ping & Goldin-Meadow, 2008; Goldin-Meadow, Cook, & Mitchell, 2009; Cook, Mitchell, & Goldin-Meadow, 2008; Brooks & Goldin-Meadow, 2015; Beaudoin-Ryan & Goldin-Meadow, 2014). Gesture may be a particularly effective way to help children focus on important relational structures or spatial features of a problem. Consistent with this possibility, children instructed in mathematical equivalence problems (e.g., $3 + 4 + 5 = \_ + 5$) learn more from a lesson that includes a gesture that highlights the two sides of the equation than from verbal instruction alone (Singer & Goldin-Meadow, 2005).

Although both action and gesture can be used as powerful learning tools, there is an open question as to *who* can best take advantage of the properties each type of tool offers. The very features that differentiate gestures from actions (i.e. the fact that they are representational, do not interact with objects, and do not affect change on the external world) may make gestures difficult to understand for some learners. In other words, some children may have trouble either mapping the abstract form of a gesture to its symbolic content, or perhaps keeping all the pieces of a problem actively in mind, which could render gesture ineffective as a teaching tool for that child. In support of this possibility, we know that very young children can understand another person's actions, like demonstrating how to twist off the top of a jar, before they can interpret a gesture that represents that action, like miming a twisting motion near the top of a jar (Novack, Goldin-Meadow, & Woodward, 2015). This evidence suggests that iconic gesture interpretation follows a later and more protracted developmental time span than action interpretation. Consequently, the meaning of iconic gestures may be unclear to some children, particularly if they are unfamiliar with the specific concept being represented by the gesture.

Very few studies have directly compared action and gesture in learning paradigms. In one study, the authors trained kindergarteners on a mental transformation task and found that learning gains in the action group happened immediately after training, while the learning gains in the gesture group occurred over a longer time course (Levine et al., 2018). In a separate study, 3rd grade children were instructed to produce a problem-solving strategy with either an action, a concrete gesture or an abstract gesture in a mathematical equivalence task (e.g., $3 + 7 + 2 = \_ + 2$) (Novack, Congdon, Hemani-Lopez, & Goldin-Meadow, 2014). While children in all groups performed similarly on a post-test, children in both of the gesture conditions performed better than other groups on a near-transfer task, and children in the abstract gesture condition performed best on a far-transfer task. The intriguing findings from these two studies suggest that the features that differentiate gesture from action may be particularly helpful for giving children a flexible, generalizable, and long-lasting understanding of the target learning concepts. Yet this leaves open the question of whether gesture is more helpful than actions-on-objects for all students, even if they have a very rudimentary understanding of a concept.

To address this question, we gave children a lesson with either action or gesture on a linear measurement task. This foundational math concept is one that many children struggle with throughout elementary school and even middle school (Lindquist & Kouba, 1989; Lehrer, Jenkins, & Osana, 1998). While traditional classroom instruction activities are largely ineffective in supporting children's understanding of spatial units, there is some recent work showing that giving children instruction that involves actions on manipulatives and evidence that their pre-existing ideas about linear measurement are wrong –'disconfirming evidence' - can improve learning outcomes (Kwon, Ping, Congdon, & Levine, submitted for publication).

Moreover, children consistently make one of two conceptually interesting errors on linear measurement problems where the to-be-measured object is not aligned with the zero-point on the ruler (shifted-object problems). See Fig. 1 for an example. In the hatch-mark counting error, children count the ruler hatch mark lines encompassing the object being measured instead of counting the intervals of space that fall between an object's left-most and right-most edges. Read-off errors consist of simply reading off the number on the ruler that aligns with the rightmost edge of the object regardless of the location of the object's left most edge on the ruler. Notably, both errors provide the *correct* answer on typical unshifted measurement problems where the object-to-be-measured is aligned with the zero point of the ruler (e.g., Blume, Galindo, & Walcott, 2007; Kamii & Clark, 1997; Lehrer et al., 1998; Solomon, Vasilyeva, Huttenlocher, & Levine, 2015).

Several findings suggest that children who primarily use the read-off strategy on shifted-object problems are further behind in their understanding of linear measurement than those who use the hatch-mark strategy. First, the read-off strategy negatively correlates with both age and socio-economic status (Solomon et al., 2015; Kwon, Levine, Ratliff, & Snyder, 2011). Second, after instruction, some students switch their strategy from read-off to hatch-mark counting, but the reverse pattern is never observed (Kwon et al., submitted for publication). Finally, at a minimum, the hatch mark strategy reflects knowledge that measurement involves counting units that are encompassed by the extent of the object, while the read-off strategy reflects no such knowledge. Taken together, these pieces of evidence suggest that children who use the read-off strategy at pre-test have lower conceptual knowledge of linear measurement than those who use the counting hatch mark strategy at pre-test.

In the current study, we begin by assessing first grade children's dominant pre-test measurement strategies. We then explore whether children at these two different levels of conceptual understanding benefit differentially from a short lesson that is accompanied by either an action (using discrete plastic unit chips superimposed on the ruler to measure length) or a gesture (counting units on the ruler with a *thumb-and-forefinger* pinching gesture). Lastly, we look for any evidence that action or gesture instruction differentially promote generalization to other unit-based tasks.

2. Method

2.1. Subjects

One hundred and twenty-two, 1st grade students (60 females; 62 males; mean age at test: 7.13 years, SD = 0.61 years) were recruited and tested in a one-on-one session with an experimenter in a quiet area of their Chicago private school. The sample had a high self-reported socio-economic status ($M = 5.7$, $SD = 0.69$, *Range* = 3–6 out of a 6-point scale, where a rating of 6 corresponded to > $100,000 family income), and was ethnically and racially diverse (10% of participants identified as Spanish/Hispanic/Latino(a); 8.3% Asian, 7.5% Black or African American, 61.6% White, and 22.5% identified with more than one racial category). The response rate for demographic information was extremely high, with 120 of 122 families choosing to report. Children whose parents signed a consent form participated in three one-on-one sessions across two weeks (Session I; Session II; Session III).

2.2. Procedure

At each of the three visits, all children received a 14-question multiple-choice paper and pencil measurement test (see Fig. 2 for a sample question). Trials were pseudo-randomized into three versions of the task, counterbalanced across visit so each child received each version only once across the three sessions. In the first four test items, the crayon image was aligned with the "0" point on the ruler ("unshifted problems"). In the 10 subsequent test items the crayons were shifted to different points on the ruler ("shifted problems"). The order of trials, unshifted items followed by shifted items, was kept constant because there was no indication of order effects in a previous study (Solomon et al., 2015). All crayons started and ended at a whole unit. The four possible answer choices reflected the correct answer, a read-off strategy answer, a hatch-mark strategy answer, and a fourth random choice that did not match any of the other three strategy-related options. This multiple-choice test was the main outcome of interest.

Because we were interested in the effect of training, we excluded children who already understood how to correctly answer these questions. Children who answered 6 or more of the 10 "shifted object" multiple-choice problems correctly on this task in Session 1 were excluded from the study and were not run in Session II or III. This criterion was based on probability values of the binomial distribution. Answering 6 or more multiple-choice questions correctly on a task with 4 options is significantly above chance ($p < .01$). Accordingly, 27

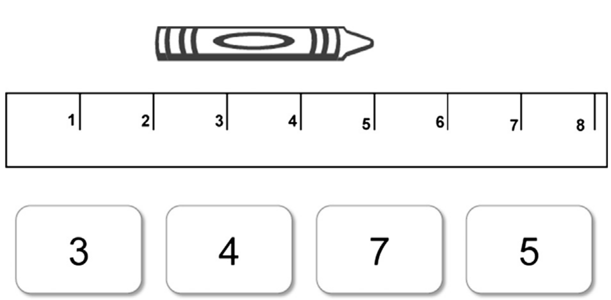

Fig. 2. A sample "shifted problem" test item from the multiple-choice measurement task.

children were excluded at Session 1, leaving 95 children in the final sample.

2.2.1. Session I

In addition to the multiple-choice crayon task (pre-test), children in the first session were given a set of tasks designed to assess their understanding of linear units and measurement tools more broadly. The first task required children to "draw a picture of a ruler that is 8-units long" on a blank piece of paper. The second required children to examine a set of four computer drawn images of rulers, which either had equal or unequal spaces and numbers or no numbers. The experimenter asked the child if each ruler, in turn, was a "useful ruler" and the child was asked to explain their answer. The purpose of this task was to assess whether or not children believe that numbers and/or evenly spaced units are crucial components of a useful measuring tool.

In the third task, participants were asked to "color a unit" on a picture of a blank ruler. The purpose of this task was to directly assess understanding of the word "unit". The fourth task was a perimeter measurement task in which children were presented with two hatch-marked shapes and asked, "How many units would it take to go all the way around the outside edge of the shape?" Finally, the children concluded the first session with a number line task in which they were asked to place 6 different numbers on 6 blank number lines, each of which were marked only with the endpoints, 0 and 100. For example, a child might be asked, "If this line goes from 0 to 100, where would 42 go?" Trials included the numbers 2, 4, 6, 18, 42, and 71 or 2, 3, 6, 25, 67, and 86. Children repeated this process on a second set of 6 number lines that were marked with the endpoints 0 and 1000 (either '4, 6, 18, 71, 230, and 780' or '2, 6, 25, 86, 390, and 810') (e.g., Siegler & Opfer, 2003). Half of the participants received the 0–100 number line first, and half received the 0–1000 number line first. Each child's set of 6 responses on each number line were coded as either logarithmic, an immature strategy in which smaller numbers are given disproportionately more "space" on the number line, or linear, a more mature strategy that allots approximately equal representation to all numbers across the number line. Children did not receive any experimenter feedback on any of the tasks in Session I.

2.2.2. Session II

In Session II, which took place at a convenient time between one and seven days after the first session (mean delay = 3.33 days, SD = 1.67 days), children were pseudo-randomly assigned to one of four training conditions, counterbalanced by both the gender of the child and their dominant answer strategy on pre-test trials (hatch-mark counting strategy or read-off strategy). The four training conditions were: unshifted unit, shifted unit, unshifted gesture and shifted gesture (Fig. 3). In each of the four conditions, an experimenter first showed children how to measure a colorful wooden stick with a 9-unit paper ruler and either discrete plastic unit chips or a thumb and forefinger gesture. In the two unshifted conditions, the wooden stick was aligned with the zero-point on the ruler and in the two shifted condition, the wooden stick was aligned with a whole number on the ruler ranging from 2 to 4. The two unit chip instruction conditions were modeled after a linear measurement training study with 2nd grade students (Kwon et al., submitted for publication). Each unit chip was exactly ¾ inches, the length of one space on the paper ruler, and was placed on the ruler in a sequential pattern as the units were counted aloud. The two gesture training conditions were developed to represent the same concepts as the unit chip instruction, using a gesture that is spontaneously produced when people talk about the size or length of small objects. The participants used this "pinching" gesture to frame the length of each individual unit of the ruler, moving sequentially from left to right as they counted aloud.

After the experimenter demonstrated how to count the units on the first trial, participants were asked to guess how long that same stick was. Then they were told to check their answer with either the unit

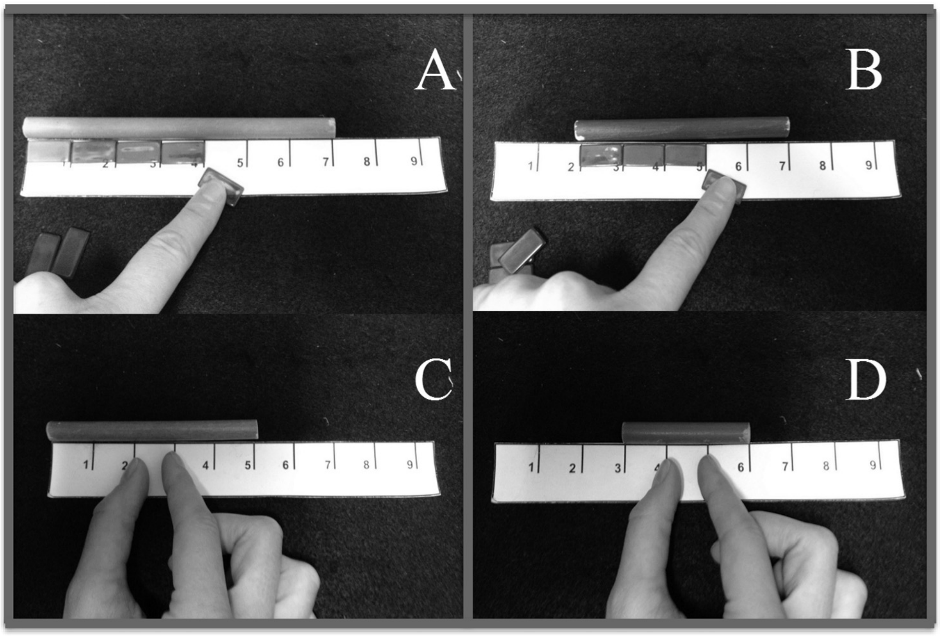

Fig. 3. A photograph of an experimenter demonstrating each of the four training conditions. A – Unshifted unit chip training condition; B – Shifted unit chip training; C – Unshifted gesture training; D – Shifted gesture training.

chips or the "gesture unit". If the child performed the movements incorrectly on the first trial, the experimenter would say, "Watch me carefully and do it just like I do" and would perform the movements one more time with their own hands and then give the child a second chance to perform the movements. On all subsequent training trials (trials 2–8), the experimenter did not provide a demonstration of counting before the child's guess and check routine. If the child performed the movement incorrectly at this point in the procedure, the experimenter would say, "Remember, do it just like I did it. Let's count together", and would move right into the "counting together" part of the training.[1] Finally, at the end of every trial, the experimenter performed the movements while counting aloud to ensure each child understood the correct answer for each trial. Following training, children immediately received a second version of the multiple-choice crayon measurement task (posttest).

2.2.3. Session III

One week after the second session (mean delay = 7.05 days, SD = 1.77 days), each participant received the multiple-choice crayon task a third time (follow-up testing). Immediately after this task, each child was given a series of generalization tasks aimed at characterizing the child's ability to transfer his or her understanding of the shifted ruler task to other tasks tapping the concept of spatial unit. In one generalization task, children were asked to measure three real-world objects with a "broken" ruler, which started with a jagged edge at the 2.5- or 3.5-unit mark. The purpose of this task was to see whether or not children would try to align the object with the broken edge of the ruler, or whether they would use the middle of the ruler to give an answer that reflected either a read-off strategy, a hatch-mark counting strategy, or the correct strategy. On a second generalization task, we asked children to use two paper clips to measure how many "paper clip units" it would take to measure a line. In addition, to assess growth across the training session, each child was again asked to color a unit on a picture of a blank ruler, complete the number line tasks, and find the perimeter of two novel but similar test items to those used on the pre-training perimeter task.

3. Results

As expected, performance on the four unshifted items on the multiple-choice crayon test was virtually perfect at all three time points for all participants (M = 3.93, SD = 0.39 at pre-test; M = 3.92, SD = 0.38 at immediate posttest; M = 3.95, SD = 0.37 at the 1-week follow-up). As such, we only carried out formal analyses on children's performance on the ten shifted-item questions.

3.1. Main outcome

The first analysis examined whether children's starting strategy (read-off vs. hatch-mark) interacted with the efficacy of the two different training conditions. Because individual children tended to get most of the problems right or most of the problems wrong at each of the three sessions, the data were non-normally distributed (Fig. 4). Accordingly, instead of performing an analysis on average scores, the data were fit with a mixed effects binomial logistic regression model that predicted correct performance on each shifted-object test item. All analyses were performed using R (R Development Core Team, 2011). In the first model, participant was a random effect, and training condition (unshifted unit, shifted unit, unshifted gesture, shifted gesture), starting strategy (read-off or hatch-mark), gender, and the two-way interaction between starting strategy and condition were used as fixed effects.

An analysis of variance of the factors in the regression model[2]

[1] Verbal reminders about performing movements correctly were administered on 3% of the total trials in the gesture training conditions and 0% of the unit chip conditions (p = .016). While these rates are quite low overall, they are consistent with the idea that the gesture may have been more difficult or non-intuitive for some children to produce. Indeed 71% of the corrections given were given to children who used the read-off strategy at pre-test.

[2] This analysis of variance of the model, also called a Wald test, is an analysis of variance of the significance of each predictor's overall coefficient, and results in the chi-square values reported here. This test, used throughout the manuscript, compares the full model to a model where each of the predictors has a coefficient that does not differ from zero. Functionally, this allows a summary of main effects of a regression model without lengthy tables of Beta values. See Bagley, White and Golomb (2001) and Peng and So (2002) for brief summaries of this regression reporting technique.

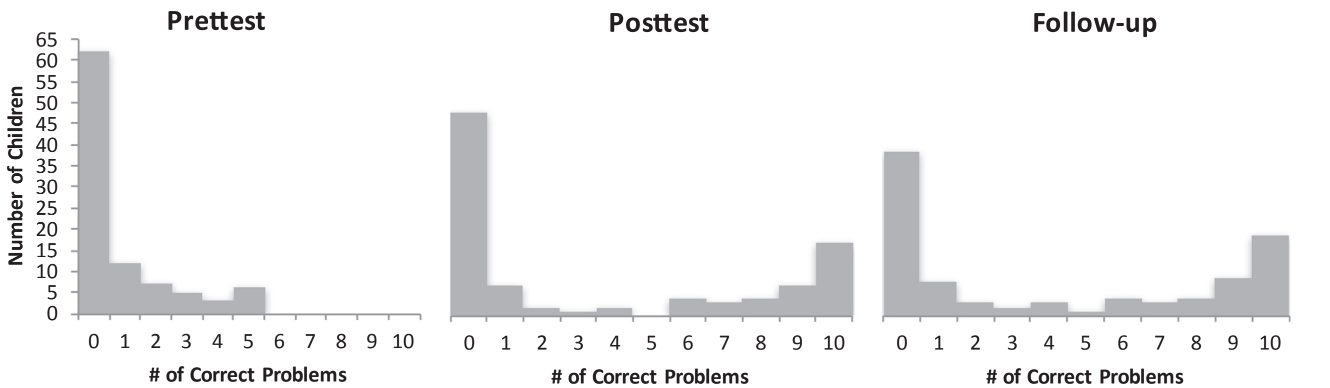

Fig. 4. Each panel shows the non-normal distribution of scores on shifted-problem test trials for all children in the final sample (N = 95), regardless of training condition, at one of the three time points. Note that children who were primarily correct at prettest (N = 27) are not represented in this figure.

showed a main effect of training condition ($X^2 = 45.60$, $p < .001$) and a main effect of starting strategy ($X^2 = 34.65$, $p < .001$). Importantly, these main effects were qualified by a significant condition by starting strategy interaction ($X^2 = 8.49$, $p < .05$). There was also a marginal effect of gender ($X^2 = 3.14$, $p = .076$), in favor of females. To better explore these results, particularly the interaction between starting strategy and training condition, we built two separate models; one for children who predominantly used the hatch-mark counting strategy at pre-test, and one for those who began by using the read-off strategy. Means and standard errors of the means for the two groups at each session are displayed in Fig. 5.

For the read-off strategy participants, an analysis of variance of a binomial logistic regression model with subject as a random effect and training condition, session, gender and the interaction between condition and session as fixed effects revealed a main effect of condition ($X^2 = 13.30$, $p < .01$), whereby children in the shifted unit condition performed better overall than each of the other three conditions. There was an expected main effect of session ($X^2 = 41.74$, $p < .001$), with a higher chance of correct responses at posttest and follow-up than at pre-test. Finally, there was a main effect of gender ($X^2 = 4.61$, $p < .05$), whereby girls outperformed boys. This gender effect was likely driving the marginal effect of gender reported in the original omnibus model, and future work should investigate why males in the read-off group were less receptive to instruction than females.

For the children who began the study by counting hatch-marks, analyses revealed a main effect of condition ($X^2 = 29.74$, $p < .001$), whereby the shifted gesture and shifted unit training conditions were more effective than the two unshifted training conditions. There was also a main effect of session ($X^2 = 170.51$, $p < .001$) whereby responses at posttest and follow-up were significantly more likely to be correct than responses at pre-test. These main effects were qualified by a significant condition by session interaction ($X^2 = 102.95$, $p < .001$), driven by the fact that there was no effect of training condition at pre-test but significant differences at posttest and follow-up. There was no effect of gender in this model ($X^2 = 0.46$, $p = .50$).

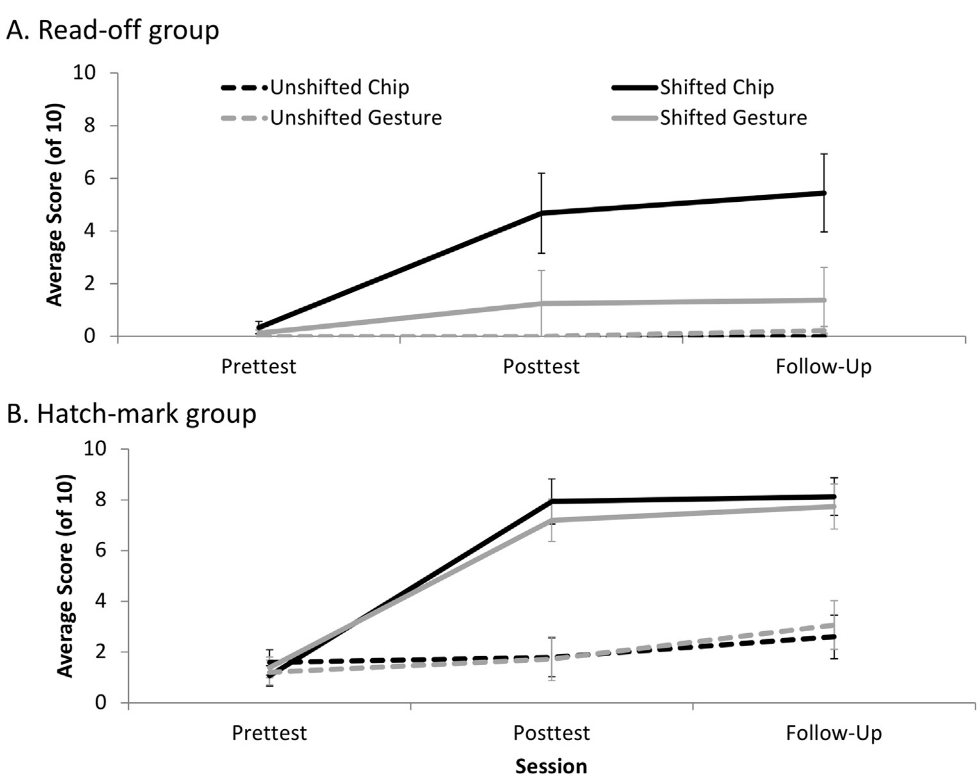

Fig. 5. Average raw performance by starting strategy (A. Read-off group; B. Hatch-mark group) and training condition across the three sessions. Error bars represent ± 1 standard error of the mean when the data are aggregated by participant. Recall that all analyses were performed on individual problems rather than on averages due to the non-normal distribution of the data.

3.2. Strategy analysis

Motivated by the low overall rates of learning in the read-off strategy group, we performed a descriptive analysis of the kinds of errors children in both groups were making before and after training to ask whether some children were showing qualitative improvements that were not captured by our main outcome (Fig. 6). This analysis showed that training led some children in the read-off group to switch their responses to the more sophisticated, yet still incorrect, hatch mark counting strategy. While the strongest effect was observed in the most successful training condition, shifted unit training, there were some children who switched their responses to the hatch-mark strategy after training in each of the other three conditions. By contrast, none of the children in the hatch-mark group increased the number of read-off strategy responses after training in any instructional condition. The overall pattern further supports the original distinction between the two groups as being at different levels of conceptual understanding. In other words, these data suggest a progression in learning from the most rudimentary strategy (read-off) to a more sophisticated but incorrect strategy (hatch-mark counting), though our data are clear that the intermediate hatch-mark counting stage is not a necessary precursor to correct performance (as some read-off strategy users did jump right to a correct strategy after our brief training, particularly those in the shifted unit training condition).

3.3. Pre-training and generalization tasks

For the two tasks administered only at pre-test, a set of simple linear regression models used performance on the pre-test task to predict an "improvement" score on the main ruler and crayon outcome task (post-test score minus pre-test score) after controlling for training condition. Children's ability to appropriately draw a ruler predicted their propensity to learn from subsequent training. This suggests that familiarity with the key features of a ruler may have been an important foundation for successful training (Table 1).

For the tasks administered at both Session I and Session III, a set of linear regression models used improvement on the main ruler and crayon outcome task (post-test score minus pre-test score); training condition; and the interaction between condition and improvement to predict the change in performance on the generalization tasks from Session I to Session III. For the last set of transfer tasks, which were administered only at Session III, the same fixed effects were used to predict performance on each transfer task. For each of these models, an analysis of variance of the regression model gives an estimate of the

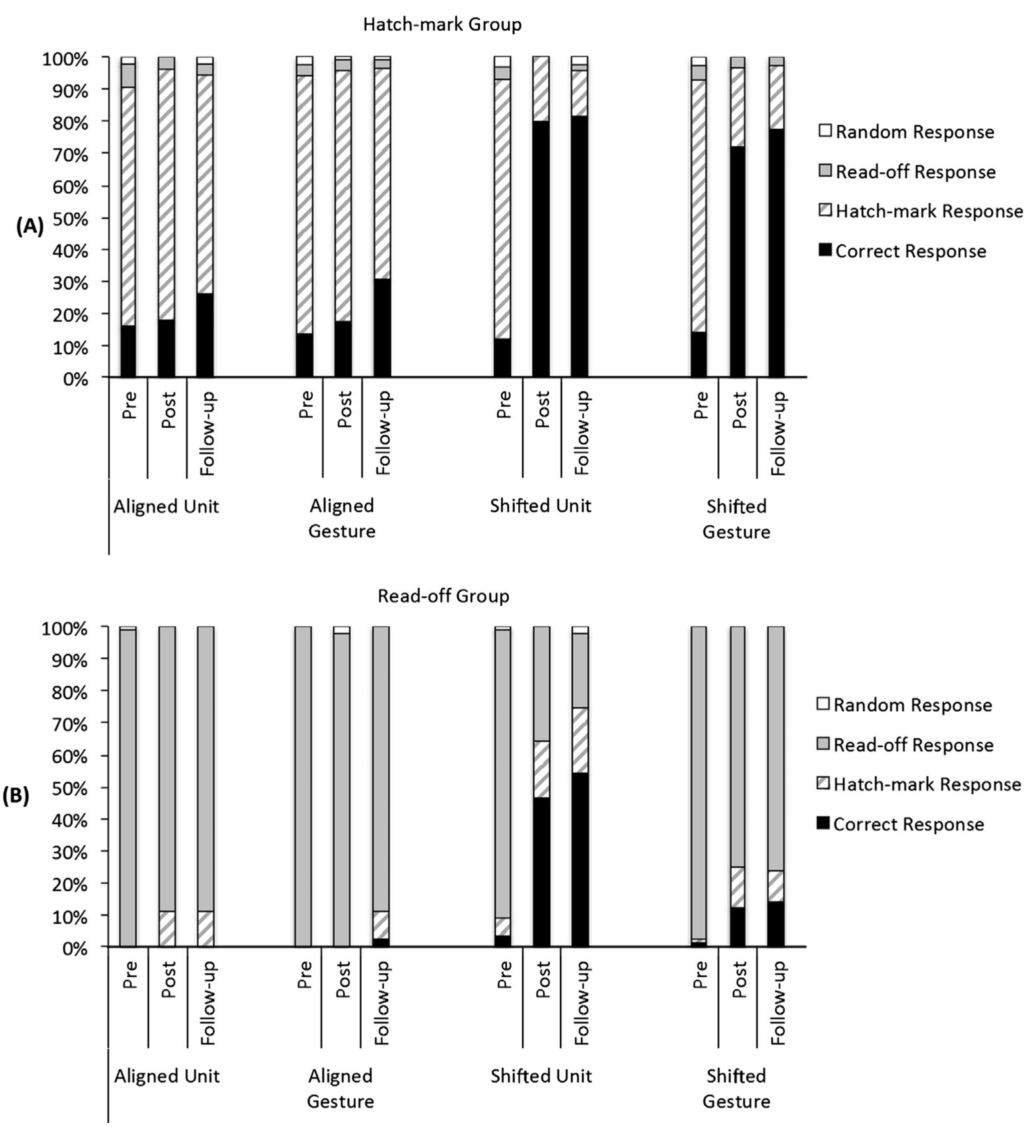

Fig. 6. Aggregate distribution of strategy use for individual trials across the entire study. After training, both correct responses (black) and hatch-mark responses (striped) increased in frequency for the read-off group [Panel B], whereas only correct responses (black) increased in frequency for the hatch-mark group [Panel A].

Table 1
Children's performance on the tasks that were administered only at pre-test.

Task	Group means (SD)	Group comparison	All participants Relation to improvement
Draw a ruler	RO = 0.51 (0.66) HM = 0.78 (0.86)	$p = .11$	$\beta = 0.95$ ($p < 0.05$)*
Useful ruler	RO = 0.79 (0.98) HM = 0.98 (0.98)	$p = .37$	$\beta = 0.45$ ($p = .31$)

Note. ** indicates that $p < .01$, * indicates that $p < .05$ and † indicates that the p-value lies between .05 and .10.

main effects and interaction terms, reported here as X^2 values (Tables 2 and 3).

Somewhat surprisingly, these analyses revealed very little evidence that training condition differentially affected transfer. The results show that while a few of the generalization tasks (Perimeter, Color a Unit, and Broken Ruler) were at least marginally related to improvement on the main outcome in general, there was no evidence of any significant effects of training condition or any interactions between improvement and training condition.

4. Discussion

The results of this study add to a growing literature that explores how the qualitative differences between actions and gestures, two similar though not identical types of movement, contribute to learning and cognition. Specifically, this study is the first to show that it is critical to consider individual differences in children's conceptual understanding of a given problem before implementing gesture-based instruction. Overall, the results raise the possibility that the properties of gesture that differentiate it from actions-on-objects, such as leaving no physical trace, providing no tactile feedback, and representing a non-present object (the unit), may make inaccessible to some learners in problem-solving contexts. By contrast, actions with appropriately designed manipulatives can be powerful drivers of learning, even for students with lower conceptual knowledge.

The results also emphasize the necessity of providing children with linear measurement instruction involving shifted-object problem types. As reported in previous work, these types of problems reveal children's misconceptions about measurement in a way that unshifted problems do not, and also support learning in a way that unshifted problems do not (Solomon et al., 2015; Kwon et al., submitted for publication). They discourage the use of simple procedural strategies and encourage the development of a more flexible and conceptually rich understanding of measurement. The kind of rapid and robust learning observed in the current study can best be explained by the idea of disconfirming evidence, or prediction error (e.g., Rescorla & Wagner, 1972; Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011). Encouraging a child to make a guess and then allowing them to discover that their answer is consistently wrong can powerfully drive conceptual change and adoption of new strategies by causing learners to question their current strategies and assumptions (Siegler & Svetina, 2006). In the training portion of the current study, all children were told to make a guess about the length of the stick before "checking their answer" with either

Table 2
Children's performance on the tasks that were administered at both pre-test and follow-up. Improvement on the perimeter task and on the color a unit task was correlated with learning, even after controlling for condition.

Task	Group means of Δ score (SD)	Group comparison	All participants Relation to improvement
Perimeter	RO = 0.29 (0.67) HM = 0.57 (0.81)	$p = .086$†	$X^2_{Improvement} = 1.93$ ($p = .06$)† $X^2_{Condition} = 2.83$ ($p = .17$) $X^2_{Interaction} = 1.58$ ($p = .41$)
Number line (1–100)	RO = 0.06 (0.54) HM = 0.10 (0.61)	$p = .72$	$X^2_{Improvement} = 0.06$ ($p = .67$) $X^2_{Condition} = 0.96$ ($p = .42$) $X^2_{Interaction} = 0.33$ ($p = .81$)
Number line (1–1000)	RO = 0.03 (0.38) HM = −0.02 (0.47)	$p = .63$	$X^2_{Improvement} = 0.06$ ($p = .59$) $X^2_{Condition} = 0.98$ ($p = .18$) $X^2_{Interaction} = 0.11$ ($p = .90$)
Color a unit	RO = 0.11 (0.47) HM = 0.27 (0.52)	$p = .16$	$X^2_{Improvement} = 1.48$ ($p < 0.05$)* $X^2_{Condition} = 1.37$ ($p = .13$) $X^2_{Interaction} = 0.95$ ($p = .27$)

Note. The number line score represents an average of three possible scores (1 = a change from logarithmic number line representation at pre-test to linear representation at posttest; 0 = no change from pre-test to posttest; or −1 = a change from linear to logarithmic). Analyses with percent absolute error and change in percent absolute error showed the same pattern of results. ** indicates that $p < .01$, * indicates that $p < .05$ and † indicates that the p-value lies between .05 and .10.

Table 3
Children's performance on the tasks that were administered only at follow-up. Children in the read-off group performed worse than children in the hatch-mark group on the broken ruler task, and performance on this task, across groups, marginally correlated with learning outcomes overall.

Task	Group means (SD)	Group comparison	All participants Relation to improvement
Broken ruler	RO = 0.63 (0.84) HM = 1.28 (0.90)	$p < 0.001$**	$X^2_{Improvement} = 2.44$ ($p = .09$)† $X^2_{Condition} = 5.01$ ($p = .12$) $X^2_{Interaction} = 2.51$ ($p = .40$)
Paperclip task	RO = 0.33 (0.48) HM = 0.50 (0.50)	$p = .124$	$X^2_{Improvement} = 0.13$ ($p = .48$) $X^2_{Condition} = 0.76$ ($p = .40$) $X^2_{Interaction} = 0.005$ ($p = .99$)

Note. ** indicates that $p < .01$, * indicates that $p < .05$ and † indicates that the p-value lies between .05 and .10.

the action or the gesture.

Although training on problems with shifted objects was necessary for improvement in the current study, it was not sufficient. In other words, not all children who recieved disconfirming evidence ended up adapting their strategies. Instead, we found an interaction between a child's starting level of conceptual knowledge and the effectiveness of gesture- and action-based instruction. First, these results support the assertion that representational gesture is more abstract than actions-on-objects, and that this distinction has context-dependent implications for cognition and learning. Second, the results add to existing literature showing that the read-off strategy is a more rudimentary procedural strategy than counting hatch marks (Kwon et al., 2011; Solomon et al., 2015; Kwon et al., submitted for publication). Not only did children in the hatch mark group learn more from training overall, but we found that after training, some children in the read-off group switched their responses to the hatch-mark counting strategy, and we never observed the opposite change in response type.

The findings raise the question about *why* children in the read-off group select such a clearly inappropriate strategy for shifted-object problems and have worse performance on the task overall. One possibility is that children who lack inhibitory control perform particularly poorly on shifted problems because at each testing point, they are exposed to four unshifted-object problems before they see the more difficult shifted problems. The read-off strategy *is* appropriate for the former problem type, and some children may simply be unable to switch strategies once they are in the rhythm of a certain response pattern, leading to read-off responses on shifted-object problems. A second possibility is that poor working memory skills are correlated with both choosing a read-off strategy and with being unable to appropriately follow and internalize the brief instruction provided during training. Future research should investigate these possibilities. Irrespective of the cause of the read-off strategy selection, the results clearly demonstrate that in the context of linear measurement, children who have a more rudimentary understanding require more concrete, tangible tools, while those with a more advanced albeit still erroneous understanding can learn from concrete actions on unit chips and from more abstract unit gestures.

Which property or properties of the gesture instruction itself are driving this condition by group interaction? One possibility is that the representational meaning of the gesture itself was opaque to the children in the read-off strategy group, and the meaning of the plastic unit chip was more obvious, or perhaps more familiar to students. This possibility is partially supported by the fact that children in the read-off group who were receiving gesture-based instruction needed more movement corrections during training than those in the hatch-mark group. Neither group needed corrections when receiving training with the plastic unit chips. In general, gesture understanding and interpretation does follow a more protracted developmental time course than does action understanding (Novack et al., 2015). Indeed, even adults require certain contextual cues to consider movements as gesture and to be able to interpret these movements appropriately (Novack, Wakefield, & Goldin-Meadow, 2016). Therefore, the failure of the children in the read-off group to learn from gesture is potentially reflective of the fact that children in the read-off group did not have the appropriate preexisting conceptual basis upon which to map the iconic measurement unit gesture. Recall that students who use the read-off strategy have not demonstrated *any* of the conceptual bases for measuring spatial extent with units. In lacking this conceptual foundation, they may have been unable to glean any novel insights from the iconic "pinching" size gesture. This may not be a problem of gesture, per se, but perhaps is a more general phenomenon that learners must have a conceptual basis upon which to map any symbolic or representational learning tool. Indeed, a prior study (Kwon et al., submitted for publication) showed that in a group of slightly older second-grade children who were persisting in using the read-off strategy on a linear measurement pretest, even training with unit chips was unsuccessful. This finding suggests that even the representational nature of the unit chips may not be immediately apparent to all learners, particularly for those children who have continued to use a rudimentary strategy as they progress later through elementary school.

A second, not mutually exclusive explanation for the difficulties associated with gesture-based training, is that gesture is cognitively demanding in the measurement context because it is iterative and does not leave a trace. In other words, it is possible that even if children in the read-off group understood that the gesture was meant to represent a small length or unit of measure, they were subsequently overwhelmed by the pragmatics of the problem; unable to keep in mind the gesture instructions, what they were supposed to be counting with the gesture, what the gesture represented, and what the final numerical answer mapped onto. In contrast, the plastic unit chips are manipulable, countable, objects that create a lasting trace in the form of a set that can be counted. Thus, it is possible that children in the read-off group, who had to make a larger conceptual leap than their peers who began with the hatch-mark counting strategy, specifically found the gesture counting, and not the unit chip counting, taxing for their working memories. Decreasing working memory load has been offered before as a potential benefit of using real-world manipulatives, because they can offload some cognitive processes (Manches & O'Malley, 2012). And while similar mechanisms have been suggested for gesture-based instruction (e.g., Morsella & Krauss, 2004; Ping & Goldin-Meadow, 2008; Cook, Yip, & Goldin-Meadow, 2012), it is possible that some familiarity with the target concept is necessary to capitalize on that feature of gesture.

While we did find some evidence of transfer in the current study, action and gesture did not *differentially* predict rates of transfer as has been reported in previous work (Novack et al., 2014). There are several potential explanations for this finding. First, there is existing research on how difficult it is for learners to apply newly acquired knowledge in novel contexts (e.g., Catrambone & Holyoak, 1989; Mix, 2010). The training we provided here was not only brief, but required children to switch between a real-world, 3D training scenario and a 2D posttest even before we assessed "transfer". Such a dimensional shift between training and testing could push the limits of flexibility in children's representational system (Barr, 2010), and the low rates of transfer on the farther generalization tasks would suggest that perhaps the tasks were not appropriately calibrated to capture meaningful differences by training condition. We must also consider the possibility that some of the analyses for the more difficult transfer tasks with smaller effect sizes were underpowered. Any future work that aims to focus purposefully on this question of transfer after measurement instruction must account for the diminishing size of participants per cell, as the sample gets split by strategy and then improvement on the main outcome measure.

It is also possible that for linear measurement, it is the learning and insight process itself that matters for success on transfer tasks and not the manner in which the insight was gained. Though there are many features that differentiate the current study from that of Novack et al., one notable difference is the type of mathematics problem being taught (linear measurement vs. mathematical equivalence). Perhaps gesture, a more abstract tool, is better suited for learning and transfer in a more abstract mathematical domain like algebraic equivalence and equation balancing than in a more spatial domain such as linear measurement. For measurement, it may be the case that either action-based or gesture-based instruction is sufficient for gaining insight and mastering this particular, highly spatial concept.

Understanding the complicated interactions between content to-be-learned and effective instruction techniques is a computationally difficult problem (Koedinger, Booth & Klahr, 2013) and there is much work to be done to discover guiding principles of when and how to implement different kinds of movement-based instruction. The current study provides a promising beginning towards this ambitious goal by highlighting two features of an instructional context that need to be considered when teaching children new ideas through hand movements –

one is the concept being taught and the other is the prior knowledge of the learner. By understanding how these factors play a role in instruction, we can support conceptual development for diverse learners in foundational mathematical domains, like measurement.

Acknowledgements

The authors would like to acknowledge Dr. Raedy Ping for her role in shaping the early stages of this project, as well as Anna Rimlinger, Maria Bavaro, Frances Carter and a wonderful team of research assistants from the Spatial Intelligence Learning Center at the University of Chicago for their help in data collection. We also thank Dr. Susan Goldin-Meadow, Dr. Amanda Woodward, and Dr. Lindsey Richland for their helpful feedback on the project during its final stages.

Funding

This research was supported by the National Science Foundation, Science of Learning Center Grants SBE-1041707 and SBE 0541957 awarded to Susan C. Levine as part of the Spatial Intelligence and Learning Center (SILC); the Benjamin Bloom Dissertation Fellowship awarded to Eliza L. Congdon; and a graduate training grant from the Institute of Education Sciences, U.S. Department of Education (R305 B090025) to the University of Chicago in support of Eliza L. Congdon.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cognition.2018.07.002.

References

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology, 54*(10), 979–985.
Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review, 30*(2), 128–154.
Beaudoin-Ryan, L., & Goldin-Meadow, S. (2014). Teaching moral reasoning through gesture. *Developmental Science, 17*, 984–990.
Blume, G. W., Galindo, E., & Walcott, C. (2007). Performance in measurement and geometry from the viewpoint of Principles and Standards for School Mathematics. In Results and interpretations of the 2003 Mathematics Assessment of the National Assessment of Educational Progress (pp. 95–138).
Brooks, N., & Goldin-Meadow, S. (2015). Moving to learn: How guiding the hands can set the stage for learning. *Cognitive Science,* 1–19.
Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1966). *Studies in cognitive growth.* New York: Wiley.
Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147.
Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*, 1047–1058.
Cook, S. W., Yip, T. K. Y., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes, 27*(4), 594–610.
DeLoache, J. S. (2000). Dual representation and children's use of scale models. *Child Development, 71*, 329–338.
Gerson, S. A., Bekkering, H., & Hunnius, S. (2015). Short-term motor training, but not observational training, alters neurocognitive mechanisms of action processing in infancy. *Journal of Cognitive Neuroscience, 27*(6), 1207–1214.
Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. (2009). Gestures gives children new ideas about math. *Psychological Science, 20*, 267–271.
Levine, S. C., Jordan, N. C., & Huttenlocher, J. (1992). Development of calculation abilities in young children. *Journal of Experimental Child Psychology, 53*, 72–93.
Levine, S., Goldin-Meadow, S., Carlson, M., & Hemani-Lopez, N. (2018). Mental transformation skill in young children: The role of concrete and abstract motor training. *Cognitive Science,* 1–22.
James, K. H. (2010). Sensori-motor experience leads to changes in visual processing in the developing brain. *Developmental Science, 13*(2), 279–288.
Kamii, C., & Clark, F. (1997). Measurement of length: The need for a better approach to teaching. *School Science and Mathematics, 97*, 116–121.
Kaminski, J. A., Sloutsky, V. M., & Heckler, A. (2009). Transfer of mathematical knowledge: The portability of generic instantiations. *Child Development Perspectives, 3*(3), 151–155.
Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science, 342*(6161), 935–937.
Kontra, C., Goldin-Meadow, S., & Beilock, S. L. (2012). Embodied learning across the life span. *Topics in Cognitive Science, 4*(4), 731–739.
Kwon, M., Levine, S. C., Ratliff, K., & Snyder, C. (2011). *The importance of alignable differences in teaching linear measurement.* Boston, MA: Cognitive Science Society.
Kwon, M., Ping, R., Congdon, E. L., Levine, S. C. (2018) Overturning children's misconceptions about ruler measurement units: The power of disconfirming evidence (submitted for publication).
Lehrer, R., Jenkins, M., & Osana, H. (1998). Longitudinal study of children's reasoning about space and geometry. In R. Lehrer, & D. Chazan (Eds.). *Designing learning environments for developing understanding of geometry and space* (pp. 137–167). Mahwah, NJ: Lawrence Erlbaum.
Lindquist, M. M., & Kouba, V. L. (1989). Measurement. In M. M. Lindquist (Ed.). *Results from the fourth mathematics assessment of the National Assessment of Educational Progress* (pp. 35–43). New York, NY: Teachers College Press.
Macedonia, M., Muller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32*, 982–998.
Manches, A., & O'Malley, C. (2012). Tangibles for learning: A representational analysis of physical manipulation. *Personal and Ubiquitous Computing, 16*, 405–419.
Mix, K. S. (2010). Spatial tools for mathematical thought. In K. S. Mix, L. B. Smith, & M. Gasser (Eds.). *In space and language*. New York: Oxford University Press.
Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *The American Journal of Psychology,* 411–424.
Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction using the hands to learn math. *Psychological Science, 25*(4), 903–910.
Novack, M., & Goldin-Meadow, S. (2017). Gesture as representational action: A paper about function. *Psychonomic Bulletin and Review, 24*, 562.
Novack, M. A., Goldin-Meadow, S., & Woodward, A. L. (2015). Learning from gesture: How early does it happen? *Cognition, 142*, 138–147.
Novack, M. A., Wakefield, E. M., & Goldin-Meadow, S. (2016). What makes a movement a gesture? *Cognition, 146*, 339–348.
Peng, C. Y., & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics, 1*(1), 31–70.
Petersen, L. A., & McNeil, N. M. (2008). *Using perceptually rich objects to help children represent number: Established knowledge counts* (pp. 1567–1572). Mahwah, NJ: Erlbaum.
Piaget, J. (1953). How children form mathematical concepts. *Scientific American, 189*(5), 74–79.
Ping, R. M., & Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology, 44*, 1277–1287.
Development Core Team, R. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.
Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLOS One, 6*(7), e22501.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory (Vol. 2, pp. 64–99).
Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237–250.
Siegler, R. S., & Svetina, M. (2006). What leads children to adopt new strategies? A microgenetic/cross-sectional study of class inclusion. *Child Development, 77*(4), 997–1015.
Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science, 16*, 85–89.
Solomon, T. L., Vasilyeva, M., Huttenlocher, J., & Levine, S. C. (2015). Minding the gap: Children's difficulty conceptualizing spatial intervals as linear measurement units. *Developmental Psychology, 51*(11), 1564.
Sommerville, J. A., & Woodward, A. L. (2010). The link between action production and action processing in infancy. In F. Grammont, D. Legrand, & P. Livet (Eds.). *Naturalizing intention in action* (pp. 67–90). Cambridge, MA: MIT Press.
Uttal, D. H., Scudder, K. V., & DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology, 18*, 37–54.
Wakefield, E. M., & James, K. H. (2015). Effects of learning with gesture on children's understanding of a new language concept. *Developmental Psychology, 5*, 1105–1114.
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review, 9*, 625–636.