Effects of Spatial Training on Mathematics in First and Sixth Grade Children

Kelly S. Mix[1], Susan C. Levine[2], Yi-Ling Cheng[3], Jerri DaSha Stockton[4], & Corinne Bower[1]

[1]University of Maryland, College Park, [2]University of Chicago, [3]Michigan State University,

[4]Michigan Public Health Institute, Okemos, Michigan

Citation: Journal of Educational Psychology: 2020, http://dx.doi.org/10.1037/edu0000494
Funding agency: Institute of Education Sciences (IES) Grant #: R305A120416
Submission date: 2019
Publication date: 7 April 2020

We thank all the children who participated in this study, as well as their parents and school district personnel, for their generous cooperation. This research was supported by a grant to Kelly S. Mix and Susan C. Levine from the Institute of Education Sciences (R305A120416). The opinions and positions expressed in this report are the authors' and do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Kelly S. Mix, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, 3304 Benjamin Building, 3942 Campus Drive, College Park, MD 20742. E-mail: kmix@umd.edu

Abstract

A pretest-training-posttest design assessed whether training to improve spatial skills also improved mathematics performance in elementary-aged children. First grade students (mean age = 7 years, *n* = 134) and sixth grade students (mean age = 12 years, *n* = 124) completed training in 1 of 2 spatial skills—spatial visualization or form perception/VSWM—or in a nonspatial control condition that featured language arts training. Spatial training led to better overall mathematics performance in both grades, and the gains were significantly greater than for language arts training. The same effects were found regardless of spatial training type, or the type of mathematics tested.

Key words: mathematics, spatial cognition, elementary students

Adults who perform better on spatial tasks also perform better on tests of mathematics (see Mix & Cheng, 2012, for a review). There is reason to believe this relation is built on shared processing between the two domains. Recent factor analyses have demonstrated that spatial skill and mathematics are separate but highly correlated domains during the elementary years, even when general cognitive ability and executive function are included (Hawes, Moss, Caswell, Seo, & Ansari, 2019; Mix et al., 2016, 2017). Similarly, longitudinal studies have shown that spatial ability at one age is a significant predictor of mathematics achievement at another (Lauer & Lourenco, 2016; Verdine et al., 2014; Wolfgang, Stannard & Jones, 2001, 2003). Further evidence comes from studies showing that similar neural circuits are activated when people process spatial and numerical information (Hubbard, Piazza, Pinel, & Dehaene, 2005; Walsh, 2003), and that performance in mathematics suffers if visuospatial processing is disrupted (e.g., Dehaene, Bossini, & Giraux, 1993; McKenzie, Bull & Gray, 2003). The nature of the shared processing driving these effects remains unknown, but some theorists have suggested the two domains are related because mathematics, along with other complex concepts, is mentally represented in a spatial format (Barsalou, 2008; Lakoff & Núñez, 2000; Lohman, 1996).

If the processes involved in spatial and mathematical thought are overlapping, it is reasonable to predict that training in one domain would lead to improvement in the other, as many have suggested (Lubinski, 2010; Levine, Foley, Lourenco, Ehrlich & Ratliff, 2016; Newcombe, 2010, 2013; Uttal et al., 2013; Verdine, Irwin, Golinkoff, & Hirsh-Pasek, 2014). Very few studies have evaluated this possibility, and so far, the results have been mixed. On one hand, several studies have shown positive effects of spatial training on numeracy and mathematics performance in both early (Cheng & Mix, 2014; Hawes, Moss, Caswell, Naqvi &

Mackinnon, 2017) and later elementary aged students (Lowrie, Logan & Ramful, 2017). In these studies, training has focused primarily on mental rotation and spatial visualization, with the amount of training varying from one 40-min session (Cheng & Mix, 2014) to several hours spread over 32 weeks (Hawes et al., 2017). Other studies, while not directly related to mathematics outcomes, have shown similar improvement among undergraduates taking science and engineering coursework following spatial visualization training (Miller & Halpern, 2013; Sorby, 2009; Sorby, Casey, Veurink & Dulaney, 2013). Yet others have reported positive effects on mathematics scores in children following mixed training that included, but was not limited to, spatial skills such as mental rotation (Nelwan & Kroesbergen, 2016). Taken together, these studies indicate a potential benefit to mathematical performance from spatial training. On the other hand, some attempts to improve mathematics performance with spatial training have failed (Cornu, Schiltz, Pazouki & Martin, 2017; Hawes, Moss, Caswell, & Poliszczuk, 2015; Rodan, Gimeno, Elosua, Montoro, & Contreras, 2019; Xu & LeFevre, 2016). In these studies, improvement in spatial skill was achieved, but there was not transfer to numeracy or mathematics. Thus, it remains an open question whether, and under what conditions, spatial training improves mathematics performance. The main aim of the present study is to address this crucial question.

Related to this, aim it is not clear why the existing studies have obtained discrepant findings. One possible explanation could be that processing differences in the mathematics content at different ages render spatial training more or less effective. We know that two of the studies for which spatial training effects were not found (Cornu et al., 2017; Xu & LeFevre, 2016) focused on younger children than had been tested in studies showing positive training effects (i.e., 3- to 5-year-olds vs. 7-to-12-year-olds). However, Hawes et al. (2015) tested 6- to

8-year-olds and similarly failed to show transfer to mathematics, so age differences may not fully explain these discrepancies. That said, this pattern is also consistent with a developmental trend in which spatial training is least effective among preschool children, has mixed effects in the early elementary grades, and is more reliably effective in late elementary grades. Because none of the existing studies have provided a direct comparison between children at various age points, it is difficult to draw firm conclusions. The present study addresses that gap by comparing children's responses to the same spatial training types in two age groups (first and sixth grade).

Secondary to the aim of determining whether spatial training improves performance in mathematics, was the aim of expanding the range of spatial skills and mathematics outcomes included so as to test specific predictions based on the particular ways in which spatial thought might be recruited to support mathematical performance. Although the existing studies differed in some ways, they shared several commonalities that limit the range of possible interpretations. For example, all of the existing studies provided training on spatial visualization and transformation (and not other spatial skills), yet some but not all studies demonstrated significant transfer to mathematics. One interpretation of this pattern may be that spatial visualization is the most potent possible spatial training, but even it does not yield consistent improvement in mathematics. Alternatively, this pattern may indicate that spatial visualization is not the most potent spatial training, and that training other spatial skills that are also highly related to mathematics performance, such as visual spatial working memory (e.g., Geary, Hoard, Byrd-Craven, Nugent & Numtee, 2007), may yield even stronger and more consistent training effects. Similarly, the existing studies have all used arithmetic, numeracy, or both as their mathematics outcomes. It is possible that certain mathematics outcomes are more sensitive to spatial training than those that have been used in existing studies, and these might yield stronger or more consistent training effects as well. For example, tasks that involve construction of a mental model (e.g., interpreting and solving word problems), or those that require attention to spatial relations in written symbols (e.g., reading algebraic equations) may be more sensitive to differences in spatial skill than tasks that require computation, particularly if this computation is based on rote procedures. An advance of the present study is that we systematically varied both spatial training type and mathematics outcomes to provide direct comparisons with the goal of obtaining a more comprehensive and nuanced account of possible spatial training effects.

The Factor Structure of Spatial Skill and Mathematics

The design of the present study was guided by recent findings related to the factor structure of spatial skill and mathematics (Mix et al., 2016, 2017). This research indicated that space and mathematics form distinct, unidimensional, but highly correlated factors. Based on this finding, one might expect that training in any spatial skill at any developmental stage should lead to improvement on any mathematics outcome. However, the training studies attempted so far suggest that this straightforward hypothesis is not the whole story. The pattern of positive effects under some conditions and a lack of transfer in others show there is no guarantee that improvement in spatial skill will transfer to mathematics, and further suggest that the underlying mechanism by which the two domains are related in development may vary, with different relations between these domains at different developmental time points.

Consistent with this notion, Mix et al. found that within the two factor structure, certain subskills showed stronger cross-domain relations than others. In particular, these studies showed an age-related difference in the specific spatial tasks that explain the most variance in mathematics scores. In multiple regressions carried out at each grade level, in which the individual spatial tasks were regressed against the mathematics factor, Mix et al. (2016) found that mental rotation and block design accounted for more variance in mathematics than any other spatial task in kindergarten. However, in sixth grade, the most predictive spatial tasks were visual-spatial working memory (VSWM) and figure copying (as measured by the Test of Visual-Motor Integration, or VMI). These age-related patterns were also reflected in the cross-domain loadings of spatial skills onto the mathematics factor in an exploratory factor analysis. That is, mental rotation and block design significantly loaded onto both the spatial and mathematics factors in kindergarten, whereas VSWM and VMI significantly loaded onto both the spatial and mathematics factors in sixth grade. Interestingly, all of these tasks were significant predictors of mathematics in third grade, and at lower and roughly equal levels, suggesting that third grade is a developmental transition period with multiple weak relations among spatial and mathematical skills instead of one dominant pattern.

Although these patterns emerged from cross-sectional evidence, they suggest an age-related shift wherein younger students' mathematical thinking relies more heavily on spatial visualization than that of older students, perhaps because so much of the mathematics young children must learn requires them to interpret new symbols and think about the transformation that these symbols require. For example, learning the meanings of single-digit numbers likely involves mappings from written or spoken numerals to groups of objects or positions on a number line. Spatial visualization may be particularly important at a time when children are actively involved in building representations of set sizes that correspond to various numerical symbols – that is, at a time when they are learning to understand and internalize mappings between number symbols and their meanings (i.e., symbol grounding) and when they are learning to interpret and solve calculations and word problems by constructing mental models of these problems.

In contrast, sixth grade students have likely moved beyond the grounding required to interpret basic mathematics symbols. Yet, spatial skill may still impact mathematics performance insomuch as it supports the ability to decode subtle differences in symbolic marks and spacing (i.e., symbol decoding). For example, when children are solving algebraic equations, they must attend to parentheses, the positions of exponents, and so forth. Indeed, much of the mathematics encountered in middle school requires complex symbol decoding as well as notationally complex multistep problem-solving, and this difference in mathematics demands might explain why there were stronger relations of VMI and VSWM to mathematics at this age. Note that although there are various ways to measure VSWM, our previous work probed memory for object locations on a grid, a skill that could be plausibly linked to tracking symbolic marks in complex algorithms, such as long division or algebra.

If these specific predictions hold true, then we should observe different training effects at different ages. For example, children in the early grades may be more responsive to training that targets spatial visualization, whereas older children may be more responsive to training that targets form perception or visual-spatial working memory (VSWM). These effects may be particularly evident on mathematics outcome measures that are sensitive to these respective spatial skills. That is, younger children trained on spatial visualization might demonstrate particularly strong effects on symbol grounding tasks such as calculation, place value, and simple word problems. However, older children are more likely to show a significant response to form perception/VSWM training compared to younger children because of their level of mathematical proficiency and the task demands of mathematics at that grade level. These responses to form perception/VSWM training may be more evident on symbol decoding tasks that likely require interpreting the spatial positions of mathematics symbols and place-keeping.

The Present Study

The present study used a pretest-training-posttest design to test the predictions outlined above. First and foremost, our aim was to test whether spatial training improves mathematics performance at these grade levels, and elucidate why previous studies have reported discrepant results. As a secondary aim, our design leverages recent insights into the factor structure of spatial skill and mathematics to expand the range of spatial skills and mathematics outcomes to test more specific predictions based on the particular ways in which spatial thought might be recruited to support mathematical performance. This is the first study to examine and compare these potential differences both within and across age groups.

We provided two kinds of spatial training (i.e., spatial visualization or form perception/VSWM) to children in two age groups (i.e., first and sixth grade), and assessed all children on a range of mathematics outcomes that tap into our hypothesized relations. We included tasks that likely require symbol grounding as well as those that likely require careful symbol decoding and place-keeping. At each grade level, we also included a control group, that engaged in language arts activities instead of spatial training, and completed all the same pre- and posttests as the training groups in order to assess whether either or both training conditions led to more mathematics learning over time than business-as-usual mathematics instruction.

Method

Participants

A total of 258 first and sixth grade children participated. An *a priori* power analysis indicated that a sample size of 222 (i.e., 111 children per grade) would be sufficient to detect a medium effect (η^2= .14, as Cheng & Mix, 2014 found) between conditions at the .90 level (Faul, Erdfelder, Buchner & Lang, 2009). First and sixth grade students were targeted as these age groups represented the two ends of the developmental trends reported by Mix et al. (2016). Note, however, that whereas Mix et al. had studied kindergarten students, we recruited first grade students because this allowed us to test a wider range of mathematics skills.

University-approved consent forms were distributed to 1,465 families whose children attended school in five different school districts in Michigan and Illinois. Schools were situated in both rural (*n* = 5) and urban (*n* = 3) areas. The median household income for these communities was $41,283 and the racial/ethnic distribution was 61.2% White, 23.7% Black, 12.5% Hispanic or Latino. Parents signed and returned 321 consent forms. Out of the 321 children who consented, 44 children were pretested but did not complete the training sessions due to excessive student absences or scheduling problems with the school or summer camp (1st, *n* = 14; 6th, *n* = 30). An additional 19 sixth grade students were excluded and replaced when it was discovered that an incorrect form of the posttest had been administered. The final sample of 258 children included 134 children in first grade (49 boys and 85 girls, mean age = 7.07, *SD* = .59) and 124 children in sixth grade (61 boys and 63 girls, mean age = 12.02, *SD* = .52). Note that only children in the final sample were included in any of the analyses. Of these, 59% of the families reported their incomes and ethnicities on an optional questionnaire attached to the consent form (*n* = 151, 1st, *n* = 70; 6th, *n* = 81). For this subsample, the median household income was between $50,000 and $74,999, and the racial/ethnic distribution was 92% White, 7% Hispanic, 5% mixed, and less than 1% Black. The socioeconomic distribution of the remaining 41% is unknown.

Children in each grade were randomly assigned to one of three conditions: (1) Spatial visualization training (first grade: *n* = 47 (19 boys and 28 girls), mean age = 7.11, *SD* =.63; sixth grade: *n* = 41 (19 boys and 22 girls), mean age = 11.95, *SD* = .49); (2) Form perception/VSWM training (first grade: *n* = 44 (16 boys and 28 girls), mean age = 7.04, *SD* = .60; sixth grade: *n* = 41 (21 boys and 20 girls), mean age = 12.09, *SD* = .51); and (3) Language arts control (first grade: *n* = 43 (14 boys and 29 girls), mean age = 7.04, *SD* = .54; sixth grade: *n* = 42 (21 boys and 21 girls), mean age = 12.03, *SD* = .51).

Procedure

Children were pretested using seven assessments that measured spatial and mathematics skills (see below for details). Testing took place in two to four 30-min sessions distributed over the course of one week. The number of test sessions depended upon the response pace and attention span of individual children. As noted below, five tests were administered in small groups (*n* = 3-6 for first grade students and *n* = 6-9 for sixth grade students) and two were administered individually. The test order was blocked and counterbalanced by individual versus group administration using a Latin square design. Furthermore, the order of the tests within each block varied randomly between pretest and posttest sessions.

Following the pretest sessions, children completed six 30-min training sessions spread over a period of 3 to 4 weeks. The pre- and posttests were administered within 2 days of the start and ending of the training sessions, respectively; however, tests were not administered on the same day as a training session. To ensure that all children received the same total duration of training, both the number and length of the training sessions were fixed. If children reached ceiling on accuracy before the six sessions were finished, they were encouraged to improve their speed on each task. The training tasks were designed to be equivalent in terms of amount of experimenter instruction. Unless children reached ceiling, there was not a reaction time (RT) requirement for either training and because the trainings were strictly spatial, they did not overlap with the mathematics outcomes. The training tasks and some of the assessments were presented using Keynote interactive slide presentations on iPads. The LED-backlit glossy widescreen iPads (version 2, 1GHz dual-core Apple A5 processor) had a diagonal screen size of 9.7-inches. The Multi-Touch display with IPS technology had a 1024-by-768-pixel resolution at 132 pixels per inch. For some training tasks (see below), feedback was also given using 3-dimensional objects.

Spatial visualization training. Training consisted of three task types: (a) Thurstone's (1974) part-whole object completion task, (b) mental rotation, and (c) tangram puzzles. The three tasks were interleaved within training sessions using a randomized block presentation. Each training task was introduced with a practice item, followed by four training trials that were ordered from easiest to hardest based on the results of previous work where possible (e.g., Mix et al., 2016). The same training trial types were repeated at each of the six sessions, but the specific objects varied.

For each trial of Thurstone's part-whole object completion task (Thurstone, 1974), a square appeared on the left side of the screen with a portion missing. Four choice shapes appeared on the right side of the screen. One choice was a shape that could be rotated to complete the square and three choices were distractors that could not be rotated to fit the empty space. Children indicated their choices by pointing. If children chose the correct shape, then a smiley face appeared on the screen with the words, “That's correct. Let's check our answers.” If children were incorrect, they were told, “That's incorrect. Let's check our answers.” To check their answers, children were given cardstock cutouts of the shapes from the training trial that could be rotated and moved into position like puzzle pieces. Children were instructed to first check the shape they chose, and then check each of the others to determine whether or not it could complete the stimulus square. The trials presented to first grade students followed

Thurstone's original procedure, in which the choice shapes were rotated clockwise 90 degrees relative to the stimulus space. To increase the difficulty of the task for sixth grade students, the choice shapes were rotated 45 degrees instead, and half of them were rotated counterclockwise relative to the target.

For mental rotation training trials, two variations of Vandenberg and Kuse's (1978) mental rotation task were used. In the first-grade version (Novack, Brooks, Kennedy, Levine, & Goldin-Meadow, 2013), small groups of children were shown four figures and asked to indicate which two were the same as the target. The two matching items could be rotated in the picture plane to overlap the target, whereas the two foils could not because they were mirror images of the target. The figures were either letters, letter-like shapes, or animals. At each session, the task was demonstrated with four practice items presented on an iPad screen. For these practice trials, children were shown animations with the correct answers rotating to match the target. On the training trials, children first responded by pointing to the two figures they thought were matching. Then they were instructed to check the accuracy of their responses by rotating into position paper circles that had the choice figures printed on them. The experimenter scaffolded point-by-point comparisons between the target and the choice stimuli. In the sixth-grade version (Neuberger, Jansen, Heil, & Quaiser-Pohl, 2011), children were shown perspective line drawings of three-dimensional block constructions, two of which could be rotated in the picture plane to match the target, based on the Shepherd-Metzler three-dimensional cube task. At the first session, the task was explained and there was a demonstration item for which children saw a perspective drawing of a three-dimensional block construction rotated into position to match the two correct choices. During the training trials, children first indicated which two choice stimuli matched the target. Then they were instructed to check the accuracy of their responses using three-dimensional cube models of the stimuli.

For the tangram puzzles, children were asked to cover a two-dimensional figure using seven geometric tiles that differed in size and shape, including two small triangles, one medium triangle, two large triangles, one square, and one parallelogram. Training sessions began with a practice trial for which the shapes of the individual tangram pieces in the solution were outlined in black within the larger stimulus figure. Children then completed four training trials (i.e., find two solutions for each of two stimulus figures) on which outlines of pieces were not provided. Children were given two minutes to cover each stimulus figure. If a child had not succeeded in finding a solution after two minutes, the experimenter provided assistance. Following joint completion of the trial, the experimenter removed the pieces covering the figure and instructed children to reproduce the same configuration independently. First grade students responded to three stimulus figures and generated two solutions for each. Sixth grade students also generated two solutions for two of the stimulus figures, but to make the task more challenging, we asked them to provide a third solution for the third stimulus shape.

Form perception/VSWM training. Training consisted of three task types: (a) VSWM (adapted from Kaufman & Kaufman, 1983); (b) Corsi Block Tapping Test (adapted from Corsi, 1972), and (c) figure copying. The three training tasks were presented in a randomized blocked order across sessions. Within each training task, there were four trials presented in increasing order of difficulty.

VSWM training trials were adapted from Kaufman and Kaufman (1983). On each trial, children were shown a 14 cm X 21.5 cm grid that was divided into squares (e.g., 3 X 3, 4 X 3, etc.), with drawings of objects displayed at random positions within the grid, but no lines indicating the divisions. On each trial, the stimulus display was left in full view for five seconds and then it was removed, at which time children indicated where the drawings had appeared by marking an "X" in the previously filled positions on a blank grid of the same size and shape. The grids for the response items were marked with lines. Stimulus displays were presented on an iPad and children made their responses in individual, paper test booklets. An experimenter showed children the correct locations on the iPad screen following each response and then made point-by-point comparisons to the children's responses. Item difficulty was manipulated by adding more divisions to the grid (up to 5 X 5) and by adding objects (up to nine). First grade students completed eight trials (four trials presented on 3 X 3 grids and four trials presented on a 4 X 3 grid), and sixth grade students completed 12 trials (four of each presented on a 4 X 3 grid, a 4 X 4 grid, and a 5 X 5 grid respectively).

For the Corsi Block Tapping Test, children were shown a sequence of blocks lighting up, and were asked to write numerals inside printed squares in a paper response booklet that represented the blocks, so as to reflect the correct sequence in which the blocks had lit up. The displays consisted of nine disconnected blocks presented on an iPad screen. The blocks lit up individually for one second each in a randomized order. For each trial, the number of blocks lighting up increased by one until the end, with nine blocks lighting up. Children were shown the correct locations on the iPad screen following each response, and the experimenter guided them to compare, block by block, the correct sequence on the screen to the written responses.

The figure copying task was adapted from the Test of Visual Motor Integration (VMI; Beery & Beery, 2010). On each trial, children saw a line drawing and their task was to copy the form in a box directly below the stimulus. In first grade, at each training session, there were three trials and in sixth grade, there were nine trials. We included more trials in the pool for sixth grade students because they tended to respond more quickly than first graders. However, the total number of sessions and the number of minutes per session were the same for both age groups. After completing each drawing, children were shown examples of “good” and “not-so-good” drawings of the form. Children were asked to look closely and identify differences between the examples. The experimenter scaffolded a comparison between the stimulus figure and the child’s drawing, and children were given instructions to help them make corrections. Figures were presented in a blocked random order at each training session. A particular set of figures was not presented more than two times in this rotation. If children drew a particular figure perfectly on the first trial, they were encouraged to reduce their response time on the second presentation. First grade students found the two-dimensional figures challenging so they did not advance to three-dimensional figures; however, sixth grade students completed training trials with both two- and three-dimensional figures.

Language activities control. Children in the control group completed three nonspatial tasks in a randomized, blocked order across sessions. The tasks included (a) crossword puzzles; (b) rhyming words; and (c) word search puzzles. All training tasks were presented on iPads using age-appropriate apps, so the particular stimuli differed between first and sixth grade, but the training tasks did not. As for the spatial training conditions, children were given feedback on the correctness of their responses and assistance generating correct responses for cases when their responses were incorrect. The duration of each session was equated to those in the spatial training conditions (i.e., six 30-min sessions).

Measures

Assessments included one nontrained spatial test (WISC-IV Block Design) and six mathematics tests (notational spacing, place value, word problems, calculation, missing terms/algebra problems, and number line estimation). The procedures for administering each measure are described below. Reliabilities for standardized tests are reported from the published test manuals, except for cases in which we created alternate forms (see below). In those cases, and for experimenter generated tests, reliabilities were estimated using Cronbach's alpha calculated from the pretest data. Posttests were administered within one week of the final training session. To reduce test-retest effects, children completed one of two test forms for each mathematics test. These forms included essentially the same items but with the specific numeric values changed (e.g., three-digit addition without carrying in both versions, but using different specific quantities in each). Note that retesting effects have been shown to occur in standardized assessments for children for a variety of test-retest intervals (Canivez & Watkins, 1998, 1999; Hausknecht, Halpert, Di Paolo, & Gerrard, 2007; Ryan, Glass, & Bartels, 2010; Tuma & Applebaum, 1980). More specifically, significant test-retest gains in spatial tasks have been noted (Uttal et al., 2013) and may be due to decreases in response times (Salthouse & Tucker-Drob, 2008); however statistically significant differences on spatial tasks on the WISC-IV such as Block Design, are not consistently evident (Canivez & Watkins, 1998, 1999). Hence, we did not use different forms of this test as pre- and posttest.

Block design (WISC-IV) (Wechsler et al., 2004). On each trial, children were shown a printed figure comprised of white and red sections, and then produced a matching figure using small cubes with red and white sides. The test was individually administered following the instructions in the WISC-IV manual. Items ranged in difficulty and children completed different numbers of items depending on their basal and ceiling performance. The reliability coefficient reported in the WISC-IV manual for the Block Design subtest is between .83 and .87 depending on age group.

Place value. Place value concepts were assessed in first grade students using a set of 20 items that required them to compare, order, and interpret multidigit numerals (e.g., "Which number has an 8 in the ones place?"), as well as match multidigit numerals to their expanded notation equivalents (342 = 300 + 40 + 2). Reliability on this experimenter-constructed measure was α =.85 for first graders. Sixth grade students completed the Rational Numbers subtest from the *Comprehensive Mathematics Abilities Test* (CMAT; Hresko, Schlieve, Herron, Swain, & Sherbenou, 2003). We considered this subtest a reasonable measure of place value understanding because more than half of the items focus on students' understanding of multidigit whole numbers and decimal place value. The CMAT is standardized for the age range 7 to 19 years of age and was administered to children in small groups. Children were asked to compare, order, and interpret written numbers, but these included a mixture of multidigit numerals, fractions, and decimals. The reliability calculated from our pretest data was α =.86.

Word problems. First grade students completed 12 word problems from the TEMA-3 (Ginsburg & Baroody, 2003) (α =.86). The test was administered in small groups (n = 3-6). Each problem was read aloud to ensure that reading ability did not influence problem solving scores. Sixth grade students completed the Problem Solving subtest from the CMAT (α =.73 calculated from our pretest data).

Number line estimation (Booth & Siegler, 2006; Siegler & Opfer, 2003). Children were given paper booklets with number lines printed on each page. The number lines were marked with a numeral at each end (e.g., 0 and 100). Children were shown a written numeral on an iPad screen and asked to mark where it would go on the number line. The particular numbers at the number line endpoints, and the range of stimulus values in between, varied by age group. Specifically, first grade students placed the numerals 4, 17, 29, 33, 48, 57, 72 and 96 on a 0-to-

100 number line, and 3, 103, 158, 240, 297, 346, 391 and 907 on a 0-to-1000 number line. Test-retest reliabilities for these two tasks, based on correlations of the pretest and posttest scores for the control group, were $r = .65$ and $r = .66$, respectively. Sixth grade students placed 1/19, 1/7, 1/4, 3/8, 4/9, 1/2, 2/3, 7/9, 5/6, and 12/13 on a 0-1 number line ($\alpha = .70$). Sixth grade students also completed a 0-to-100,000 number line task but we did not include their results due to ceiling effects (i.e., 93% of children performed with almost perfect linearity). Note that our analyses used the linearity of children's responses rather than absolute distance to the target for several reasons. Linearity captures internally consistent placements that may be otherwise incorrect (i.e., sets or responses that were linear relative to each other even if they were not correctly positioned on the number line relative to its endpoints), and thus may be more sensitive to ordering than absolute error. In past research, linearity was correlated with mathematics achievement outcomes, whereas error rates were not (Booth & Siegler, 2006). Finally, although the same developmental patterns have been observed for the two measures, absolute error rates have been affected by changes in overall accuracy (e.g., reducing overestimates across all items) rather than changes in linearity alone, and therefore absolute error rates may be less meaningful than linearity (Opfer & Siegler, 2007) .

Calculation. Children in each grade completed a group-administered test with age-appropriate multistep arithmetic problems (first grade: α= .83; sixth grade: α= .83). The 12 first grade items included two- to four-digit whole number addition and subtraction problems. The 24 sixth grade items included two- to five-digit problems, some with decimals, and sampled from all four operations (addition, subtraction, multiplication, and division).

Missing term problems/algebra. In missing term problems, children find the solution to a calculation problem where the solution is provided but one of the addends or subtrahends is missing (e.g., __ + 8 = 14). Previous research found that mental rotation training was effective at raising children's scores on such problems (Cheng & Mix, 2014). Only first grade students completed missing term problems because they are not challenging for most sixth-grade students (n = 14 items, α = .90). Instead, sixth grade students completed the Algebra subtest from the CMAT. Although more sophisticated than missing term problems, algebra items involving solving for unknowns could be considered related to the missing terms problems. The reliability reported in the CMAT manual is α =.90.

Notational spacing. First grade items were comprised of the vertical arithmetic calculation problems included on the *Test of Early Mathematics Ability-Third Edition* (TEMA-3, Ginsburg & Baroody, 2003). The 12 problems were presented one at a time on iPad screens. The spacing of the numbers was manipulated so as to vary their vertical alignment (see Appendix A in the online supplemental materials). Children were asked on each trial whether the problems were written correctly. Sixth grade items were algebra problems adapted from the stimuli used by Landy and Goldstone (2010). Children solved 25 horizontal multiplication and order of operation problems (e.g., 3 + 4 x 2), in which the spacing of symbols in the equations was manipulated to be consistent or inconsistent with the order of operations. In both grades, problems were presented individually on iPad screens using Keystone interactive software. Problems were presented in one of four random orders that varied across children, and order was counterbalanced from pre- to posttest. Across the four test forms, the specific numbers in the problems also were manipulated so that the problem structures remained parallel across forms, but not the specific computations themselves. Although the first and sixth grade measures focused on different mathematics, with different implications of shifts in spatial position, the commonality was that both assessed students' understanding of the spatial position of symbols using content that were age-appropriate. The reliabilities for these measures were α = .57 and .71 for first and sixth grades, respectively.

Results

Children's scores on each of the various outcome measures were converted into percentages and analyzed using one of several composite scores. One composite captured children's overall performance and was an average of the percent correct on all six mathematics measures. A second composite score included only tasks we hypothesized to have a strong symbol grounding component (i.e., place value, word problems, and number line estimation) and a third composite score included only tasks we hypothesized to have a strong form perception or symbol decoding component (i.e., missing terms/algebra, notational spacing, and multistep calculation).

For each mathematics outcome, we evaluated changes in children's performance from pre- to posttest using one-tailed *t* tests, as well as comparing differences in performance across training conditions using analyses of covariance (ANCOVA). We used one-tailed *t* tests because we predicted that training would improve, and not worsen performance on any of the mathematics outcomes. We used ANCOVAs because these analyses incorporate a control for pretest differences while permitting comparisons in outcomes across training groups. In this way, we adjusted for any baseline differences among students. One concern with submitting children's scores to the same omnibus analysis might be that the measures for first and sixth grade, though conceptually equivalent, were not exactly the same and thus, should not be combined into a single analysis. However, we obtained the same basic patterns of results whether we used a grade-specific ANCOVAs, so for ease of presentation, we report the results of the omnibus ANCOVAs here.

Manipulation Check

To determine whether the spatial training we provided led to a general improvement in spatial skill, we carried out an ANCOVA with grade (1^{st} vs. 6^{th}) and condition (spatial training vs. control) to analyze children's posttest performance on the Block Design subtest of the WISC-IV while controlling for pretest performance. There was a small effect of condition such that spatial training led to greater improvement in Block Design scores than language arts exercises in the control group, $F(1, 253) = 5.239$, $p = .023$, $\eta^2_p = .020$. Although there were no interactions involving grade, pairwise comparisons between children's pre- and posttest scores revealed an important difference across grades. Among first grade students, there was significant improvement in Block Design for both spatial training groups (Spatial Visualization: $t(43) = 3.24$, $p = 0.001$, $d = .49$; Form Perception: $t(43) = 4.84$, $p < 0.001$, $d = .73$) as well as the control group, $t(42) = 2.73$, $p = 0.005$, $d = .42$. This improvement in the control group may reflect the test-retest improvement that has been reported in previous research using timed spatial tasks (Salthouse & Tucker-Drob, 2008), and tempers our interpretation of the first-grade training effects. In sixth grade, only children in the two spatial training conditions improved on the WISC-IV Block Design test from pre- to posttest (Spatial Visualization: $t2(40) = 2.88$, $p = 0.003$, $d = .45$; Form Perception/VSWM: $t(40) = 3.28$, $p = 0.002$, $d = .52$; Control: $t(41) = .260$, $p = .40$, $d = .04$). Thus, the effects of spatial training on children's WISC-IV Block Design scores in sixth grade were clearcut.

Performance on a broad mathematics composite measure

The main question addressed in this study was whether we would obtain significant spatial training effects on mathematics performance given the mixed results reported in the extant literature. To evaluate this question, we first carried out an ANCOVA using children's mathematics posttest scores as the dependent variable, their mathematics pretest scores as a covariate, and grade and condition (spatial training vs. control) as the between subjects variables. Because we were interested in testing the broad effects of spatial training on mathematics, we combined the data for children in the two spatial training conditions and compared their performance to that of the control group. We also used composite mathematics scores that combined performance on all six mathematics subtests. Note that because grade was included as a between subjects factor, we used *z*-scores to equate children's scores across the two grades. Recall that our measures differed in specific mathematics content across the grades.

Children's mean pre- and posttest performance is presented in Figures 1a and 1b. The results of the ANCOVA revealed a medium effect of condition, $F\ (2, 250) = 7.80$, $p < .001$, $\eta^2_p = .059$ due to superior performance in the two spatial training conditions versus controls (Spatial Visualization vs. Controls, $p = .000$; Form Perception vs. Controls, $p = .028$) and not because of differences between the two spatial training conditions ($p = .093$). The interaction between grade and condition was not significant ($p = .864$; $\eta^2_p = .001$), indicating the same pattern of greater improvement following spatial training for both first and sixth grade. Consistent with this result, pre- to posttest comparisons showed that children in the spatial training conditions improved significantly in both grades (See Table 1). Even though first graders in the control condition also improved, their improvement was less than that of the spatial training groups, $F\ (1, 131) = 4.77$, $p = .03$, $\eta^2_p = .035$ and could have been due to learning over time and/or test-retest effects. These important findings add to the literature by demonstrating a direct, and causal effect of spatial training on mathematics performance. For performance on the individual mathematics measures as a function of training condition (see Appendix B in the online supplemental materials).

INSERT TABLE 1 ABOUT HERE

INSERT FIGURES 1a, 1b ABOUT HERE

Performance on specific mathematics outcomes: Symbol grounding versus symbol decoding composites

A secondary question was whether specific types of spatial training lead to significant improvement on specific mathematics measures. We hypothesized that spatial visualization training would improve performance on mathematics tasks that require symbol grounding (i.e., interpreting the meaning of symbols), and form perception/VSWM training would improve performance on mathematics tasks that require symbol decoding (i.e., discriminating among symbols using spatial cues or tracking steps in written problem solving). To test whether this was the case, we first analyzed children's posttest symbol decoding and symbol grounding scores using a multivariate analysis of covariance (MANCOVA) with grade (1^{st} vs. 6^{th}) and condition (spatial visualization training, form perception training, and language arts control) as between subjects factors, and children's pretest scores as covariates. As before, we used *z*-scores to equate children's scores across the two grades because the specific mathematics content included in our measures differed across the grades.

After determining that children's symbol decoding scores differed significantly from their symbol grounding scores (Wilks' Lambda = .431, ($F(8, 504) = 33.01$, $p = .000$), we used Helmert contrasts to probe for significant differences based on condition. For mathematics tasks that tapped symbol decoding, there were significant differences favoring the spatial training groups in comparison to the control group, $F(1, 253) = 12.68$, $p = .000$ but performance did not differ between the two spatial training groups, $F(1, 253) = .01$, $p = .93$. For mathematics tasks that tapped symbol grounding, none of the contrasts were significant. Furthermore, contrary to our age-related predictions, none of the interactions involving grade reached significance.

We next compared children's pre- and posttest performance on the two composite mathematics scores (symbol grounding and symbol decoding) for each of the three conditions within each grade (see Tables 2 and 3). As shown in the tables, children in both grades improved significantly on both the symbol grounding and symbol decoding composite mathematics measures, given either spatial visualization or form perception training. This pattern suggests a broad response to training that is not limited to practice on one particular spatial skill or another. Note that result is reminiscent of the factor structure revealed in previous research (e.g., Mix et al., 2016) in that the spatial measures and mathematics measures formed separate but high correlated unitary factors onto which all the measures within each domain loaded significantly.

INSERT TABLE 2A ABOUT HERE

INSERT TABLE 2B ABOUT HERE

Discussion

This study tested whether spatial training leads to improvement on a range of mathematics outcomes in first and sixth grade students. Two types of spatial training were provided in a between-subjects pretest-training-posttest design. The data were analyzed in terms of overall improvement, and using specific probes linked to age, training type, and mathematics outcome. Our main finding was that spatial training led to significant improvement in mathematics outcomes in both age groups. This was evident when the two spatial training conditions and the mathematics measures were collapsed in a general analysis, and also when improvement from pre- to posttest on more specific mathematics composite measures was considered. Thus, the present study provides reason to conclude that spatial training improves mathematics performance among elementary students.

Furthermore, this improvement was not restricted to one age group. The same basic pattern was obtained at both first and sixth grade. This finding is of interest because the pattern of successful and unsuccessful transfer to mathematics in previous studies hinted that spatial training may be less effective at younger ages. However, using similar training procedures in first grade and sixth grade, we found similar effects on mathematics performance at both grade levels. That said, it should be noted that the children in some of the previous studies were two or three years younger (i.e., preschool age) than those in our younger age group (i.e., first grade), so it remains possible that spatial training is not effective at very young ages.

We found no support for the unique predictions for symbol grounding versus symbol decoding. Children showed significant improvement on both types of mathematics outcomes with both types of spatial training. Thus, the overall picture seems to be one in which spatial skills may be interchangeable when it comes to their effects on mathematics, and likewise, mathematics skills, at least those tested, are not differentially sensitive to spatial training. This pattern makes sense given previous research demonstrating unitary factor structures for each domain that are highly correlated (Mix et al., 2016), and casts doubt on the idea that the previously reported crossloadings reflect meaningful differences.

That said, cognitive training can have subtle and complex effects (Potzko, 2017), not all of which might have been detected in the present design. For example, our sample sizes provided adequate power to detect medium effects, but if the effects involving outcome measures were specific to subgroups of children or particular outcome measures, larger samples may be needed to detect them. Furthermore, the stability of spatial training effects is difficult to assess given the extant evidence, including the present study. On one hand, the effects may be short-term and operate akin to priming or coaching effects, presenting themselves only during or shortly after training. On the other hand, they may be long term, leading to durable and perhaps increasing effects on mathematics as children are better prepared to incorporate new mathematics content using their improved spatial skills. Longitudinal research examining the time course of spatial training effects, and whether it sets children up for better learning of novel content, is needed to determine which is the case.

Finally, one might question whether the effects we obtained are specific to spatial and mathematics skill, rather than being attributable to an improvement in general cognitive ability. In previous work examining the relation of spatial skill to mathematics, general cognitive ability was controlled and the relation still held (Mix et al., 2016), suggesting that effects involving spatial skill and mathematics are not at the level of general ability. However, we did not include a measure of general cognitive ability in the present study, so we cannot completely rule out the possibility that our spatial training led to broad improvement in processes such as working memory or attention that led, in turn, to improvements on both our spatial and mathematics outcome measures.

A remaining question is why children in the present study benefitted from spatial training when other closely related studies have failed to obtain such effects (Cornu, et al., 2017; Hawes, et al., 2015; Xu & LeFevre, 2016). Recall that in these studies, improvement in spatial skill was achieved, but there was not transfer to numeracy or mathematics. As we noted previously, two of these studies focused on prekindergarten and kindergarten children (Cornu et al., 2017; Xu & LeFevre, 2016) whereas the youngest participants in the present study were in first grade. One might conjecture that differences in the mathematics content for these age groups, or differences in children's cognitive abilities at these ages might explain why spatial training was not effective in preschool and kindergarten. However, Hawes et al. (2015) tested 6- to 8-year-olds and also failed to show transfer to mathematics, so age differences cannot fully explain existing discrepancies.

An alternative explanation for the discrepant effects of spatial visualization training may be the amount and type of feedback we provided to children during the training trials compared to the studies that found null results of training on mathematics outcomes. When children were incorrect on the spatial visualization training tasks, we offered them physical models of the stimuli to rotate in and out of position and thereby check the accuracy of their responses. From an embodied cognition perspective (e.g., Barsalou, 2008; Lakoff & Núñez, 2000), this object- and movement-based feedback may have been crucial for fostering the development of strong mental rotation and spatial visualization abilities. Consistent with this, a mental transformation training study with 6-year-olds showed that having them gesture the movement of relevant pieces improved their mental transformation skills whereas watching the experimenter gesture the movement of the pieces did not (Goldin-Meadow et al., 2012). Further, training that involved having 5- to 6-year-old children move pieces or gesture the movement of pieces both improved their mental transformation skill (Levine, Goldin-Meadow, Carlson, & Hemani-Lopez, 2018). In contrast to the embodied training we provided, the training provided by Hawes et al. was presented on iPads as two-dimensional matching games. It appears that all feedback was embedded in a set of automated computer games, and was limited to only whether the answer was right or wrong. Although such training was sufficient to raise children's mental rotation scores, it might not have been extensive or detailed enough to show transfer to mathematics performance. Consistent with this possibility, a recent study found that 5-year-olds' mental rotation ability benefited more from training that involved gesturing the rotation than from training that involved rotating an image on a touchscreen (Wakefield et al., 2019).

In summary, the present results provide evidence that spatial training can have positive effects on mathematics outcomes in elementary aged students. Relations hypothesized at the intersection of training type, age, and mathematics outcomes were not supported. Thus, the overall picture seems to be that spatial training in general seems to support mathematical performance in general, as one might expect based on the correlated unitary factors for spatial and mathematical performance revealed in previous research (Mix et al., 2016).

The finding of spontaneous transfer from spatial skill to mathematics is exciting but is likely the tip of the iceberg in terms of how spatial training might be leveraged. Activating one's spatial skills prior to performing a mathematics task or improving spatial processing to some threshold may be helpful or even necessary for strong performance in mathematics, but it may not be sufficient on its own for all children, and it may not go far enough to achieve the full potential of spatial training. Children may not spontaneously recruit spatial processing into mathematics problem solving, even when they have sufficient levels of spatial skill. It is also unlikely they receive direct instruction in ways that invite them to recruit spatial processing. As suggested by Casey and Fell (2018), future research should examine not only the immediate effects that follow improvement of spatial skill, but also whether helping children purposely recruit spatial reasoning into specific mathematics tasks leads to even greater mathematics learning benefits.

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*(1), 617-645. doi:10.1146/annurev.psych.59.103006.093639

Beery, K. E., & Beery, N. A. (2010). *The Beery-Buktenica developmental test of visual-motor integration: Beery VMI with supplemental developmental tests of visual perception and motor coordination: Administration, scoring and teaching manual* (6th ed.). Minneapolis, MN: NCS Pearson.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189-201. doi:10.1037/0012-1649.41.6.189

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children—third edition. *Psychological Assessment*, *10*(3), 285-291. doi:10.1037/1040-3590.10.3.285

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children—third edition, among demographic subgroups: gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment*, *17*(4), 300-313. doi:10.1177/073428299901700401

Casey, B. & Fell, H. (2018). Spatial reasoning: A critical problem-solving tool in children's mathematics strategy tool-kit. In K. S. Mix and M. Battista (Eds.) *Visualizing Mathematics: The Role of Spatial Reasoning in Mathematical Thought* (pp. 47-75). New York: Springer.

Cheng, Y., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, *15*(1), 2-11.

Cornu, V., Schiltz, C., Pazouki, T., & Martin, R. (2017). Training early visuo-spatial abilities: a controlled classroom-based intervention study. *Applied Developmental Science, 6*(6), 1-21. doi:10.1080/10888691.2016.1276835

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain*. (vol. 34). Montreal, McGill University.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371-396. doi:10.1037//0096-3445.122.3.371

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. doi:10.3758/BRM.41.4.1149

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development, 78,* 1343–1359.

Ginsburg, H. P., & Baroody, A. J. (1983/2003). *Test of Early Mathematics Ability, Third Edition*. Austin, TX: Pro-Ed.

Goldin-Meadow, S., Levine, S. C., Zinchenko, E., Yip, T. K., Hemani, N., & Factor, L. (2012). Doing gesture promotes learning a mental transformation task better than seeing gesture. *Developmental Science*, *15*(6), 876-884.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373-385. doi.org/10.1037/0021-9010.92.2.373

Hawes, Z., Moss, J., Caswell, B., Naqvi, S., & MacKinnon, S. (2017). Enhancing children's spatial and numerical skills through a dynamic spatial approach to early geometry instruction: Effects of a 32-week intervention. *Cognition and Instruction, 35*(3), 236-264. doi:10.1080/07370008.2017.1323902

Hawes, Z., Moss, J., Caswell, B., & Poliszczuk, D. (2015). Effects of mental rotation training on children's spatial and mathematics performance: A randomized controlled study. *Trends in Neuroscience and Education, 4*(3), 60-68. doi:10.1016/j.tine.2015.05.001

Hawes, Z., Moss, J., Caswell, B., Seo, J., & Ansari, D. (2019). Relations between numerical, spatial, and executive function skills and mathemat- ics achievement: A latent-variable approach. *Cognitive Psychology, 109,* 68 –90. doi: 10.1016/j.cogpsych.2018.12.002

Hresko, W. P., Schlieve, P. L., Herron, S. R., Swain, C., & Sherbenou, R. J. (2003). *Comprehensive Mathematical Abilities Test.* Austin, TX: Pro-Ed.

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience, 6*(6), 435-448. doi:10.1038/nrn1684

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Lakoff, G., & Núñez, R. E. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York, NY, US: Basic Books.

Landy, D. & Goldstone, R. L. (2010) Proximity and precedence in arithmetic, *The Quarterly Journal of Experimental Psychology*, 63:10, 1953-1968, DOI:10.1080/17470211003787619

Lauer, J. E., & Lourenco, S. F. (2016). Spatial processing in infancy predicts both spatial and mathematical aptitude in childhood. *Psychological Science*, *27*(10), 1291-1298.

Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Sex differences in spatial cognition: advancing the conversation. *WiREs Cognitive Science*, *7*(2), 127-155. doi:10.1002/wcs.1380

Levine, S. C., Goldin-Meadow, S., Carlson, M. P., & Hemani-Lopez, N. (2018). Mental transformation skill in young children: The role of concrete and abstract motor training. *Cognitive Science, 43,* 1207–1228. http://dx.doi.org/10.1111/cogs.12603

Lohman, D. F. (1996). Spatial ability and g. In I. Dennis & P. Tapsfield (Eds.) *Human Abilities: Their Nature and Measurement* (pp. 97-116). Mahwah, N.J.: Erlbaum.

Lowrie, T., Logan, T., & Ramful, A. (2017). Visuospatial training improves elementary students' mathematics performance. *British Journal of Educational Psychology*, *87*(2), 170-186.

Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. *Personality and Individual Differences*, *49*(4), 344-351. doi:10.1016/j.paid.2010.03.022

McKenzie, B., Bull, R., & Gray, C. (2003). The effects of phonological and visual-spatial interference on children's arithmetical performance. *Educational and Child Psychology, 20*(3), 93-108.

Miller, D. I., & Halpern, D. F. (2013). Can spatial training improve long-term outcomes for gifted STEM undergraduates?. *Learning and Individual Differences*, *26*, 141-152.

Mix, K. S., & Cheng, Y. (2012). The relation between space and math: Developmental and educational implications. In J. B. Benson (Ed.), *Advances in Child Development and Behavior, Vol. 42* (pp. 197-243). San Diego, CA, US: Elsevier Academic Press.

Mix, K. S., Levine, S. C., Cheng, Y., Young, C. J., Hambrick, D. Z., & Konstantopoulos, S. (2017). The latent structure of spatial skills and mathematics: A replication of the two-factor model. *Journal of Cognition and Development,* 1-28. doi:10.1080/15248372.2017.1346658

Mix, K. S., Levine, S. C., Cheng, Y., Young, C., Hambrick, D. Z., Ping, R., & Konstantopoulos, S. (2016). Separate but correlated: The latent structure of space and mathematics across development. *Journal of Experimental Psychology: General*, *145*(9), 1206-1227. doi:10.1037/xge0000182

Nelwan, M., & Kroesbergen, E. H. (2016). Limited near and far transfer effects of jungle memory working memory training on learning mathematics in children with attentional and mathematical difficulties. *Frontiers in Psychology*, *7*, 1384. doi: 10.3389/fpsyg.2016.01384

Neuburger, S., Jansen, P., Heil, M., & Quaiser-Pohl, C. (2011). Gender differences in pre-adolescents' mental-rotation performance: Do they depend on grade and stimulus type? *Personality and Individual Differences, 50*, 1238–1242. http://dx.doi.org/10.1016/j.paid.2011.02.017

Newcombe, N. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, *34*(2), 29-35.

Newcombe, N. S. (2013). Educating to use evidence in thinking about education. *Mind, Brain, And Education*, *7*(2), 147-150.

Novack, M., Brooks, N., Kennedy, D., Levine, S., & Goldin-Meadow, S. (October, 2013). *Using Action and Gesture to Improve Mental Rotation.* Poster presented at the biennial meeting of the Cognitive Development Society, Memphis, TN.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive psychology*, *55*(3), 169-195.

Potzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin and Review, 24*(4), 1022-1031. doi: 10.3758/s13423-016-1196-1.

Rodan, A., Gimeno, P., Elosua, M. R., Montoro, P. R., & Contreras, M. J. (2019). Boys and girls gain in spatial but not in mathematical abiliytyafter mentoral rortation training in primary education. *Learning and Individual Differences, 70*, 1-11. doi.org/10.1016/j.lindif.2019.01.001.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology, 17*(1), 68-72.

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology, 22*(6), 800–811. doi:10.1037/ A0013091

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237-243. doi:10.1111/1467-9280.02438

Sorby, S. A. (2009). Educational research in developing 3-D spatial skills for engineering students. *International Journal of Science Education*, *31*(3), 459-480.

Sorby, S., Casey, B., Veurink, N., & Dulaney, A. (2013). The role of spatial training in improving spatial and calculus performance in engineering students. *Learning and Individual Differences*, *26*, 20-29.

Thurstone, T. G. (1974). *PMA readiness level*. Chicago , IL: Science Research Associates.

Tuma, J. M., & Appelbaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement, 40*(3), 671-678.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352-402.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599-604. doi:10.2466/pms.1978.47.2.599

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Filipowicz, A. T., & Chang, A. (2014). Deconstructing building blocks: Preschoolers' spatial assembly performance relates to early mathematical skills. *Child Development*, *85*(3), 1062-1076.

Verdine, B. N., Irwin, C. M., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Contributions of executive function and spatial skills to preschool mathematics achievement. *Journal of Experimental Child Psychology*, *126*, 37-51.

Wakefield, E.M., Foley, A., Ping, R., Villarreal, J., Goldin-Meadow, S., & Levine, S.C. (2019). Breaking down gesture and action in mental rotation: Understanding the components of movement that promote learning. *Developmental Psychology, 55*(5), 981-993.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483-488. doi:10.1016/j.tics.2003.09.00

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maerlender, A. (2004). *Wechsler Intelligence Scale for Children Fourth Edition-Integrated*. Psychological Corporation: San Antonio, TX

Wolfgang, C. H., Stannard, L. L., & Jones, I. (2001). Block Play Performance Among Preschoolers as a Predictor of Later School Achievement in Mathematics. *Journal of Research in Childhood Education, 15*(2), 173-180. doi:10.1080/02568540109594958

Wolfgang, C., Stannard, L., & Jones, I. (2003). Advanced constructional play with LEGOs among preschoolers as a predictor of later school achievement in mathematics. *Early Child Development and Care*, *173*(5), 467-475.

Xu, C., & LeFevre, J. (2016). Training young children on sequential relations among numbers and spatial decomposition: Differential transfer to number line and mental transformation tasks. *Developmental Psychology*, *52*(6), 854-866. doi:10.1037/dev0000124

Table 1
First grade and sixth grade performance on a broad mathematics composite measure by training condition

	First grade			Sixth grade		
	Pretest	Posttest	η_p^2	Pretest	Posttest	η_p^2
SPATIAL TRAINING	.39 (.16)	.45 * (.18)	.035	.50 (.12)	.54 * (.11)	.056
CONTROL	.42 (.11)	.44 * (.13)		.50 (.10)	.51 (.10)	

NOTE: For each Pretest and Posttest cell, we report the mean and standard deviation. Significant gains from pre- to posttest are indicated with an asterisk (one tailed, $p < .05$). Effect sizes, η_p^2, of the ANCOVAs comparing the spatial training and control groups within each grade are reported.

Table 2a
Children's symbol grounding composite scores by training condition and grade.

	First grade			Sixth grade		
	Pretest	Posttest	η_p^2	Pretest	Posttest	η_p^2
SPATIAL VISUALIZATION	.46 (.18)	.54 * (.21)	.02	.67 (.16)	.72 * (.15)	.03
FORM PERCEPTION	.40 (.21)	.46 * (.22)		.65 (.18)	.69 * (.15)	
CONTROL	.46 (.14)	.51 * (.16)		.68 (.15)	.70 (.13)	

Table 2b
Children's symbol decoding composite scores by training condition and grade.

	First grade			Sixth grade		
	Pretest	Posttest	η_p^2	Pretest	Posttest	η_p^2
SPATIAL VISUALIZATION	.34 (.15)	.41 * (.20)	.04	.33 (.10)	.36 * (.08)	.04
FORM PERCEPTION	.33 (.16)	.37 * (.16)		.35 (.10)	.37 * (.11)	
CONTROL	.35 (.12)	.35 (.13)		.32 (.08)	.32 (.09)	

NOTE: For each Pretest and Posttest cell, we report the mean and standard deviation. Significant gains from pre- to posttest are indicated with an asterisk (one tailed, $p < .05$). Effect sizes, η_p^2, of the ANCOVAs comparing the spatial training and control groups within each grade are reported.

Figure 1A: Average composite mathematics pre- and posttest performance in 1st grade.

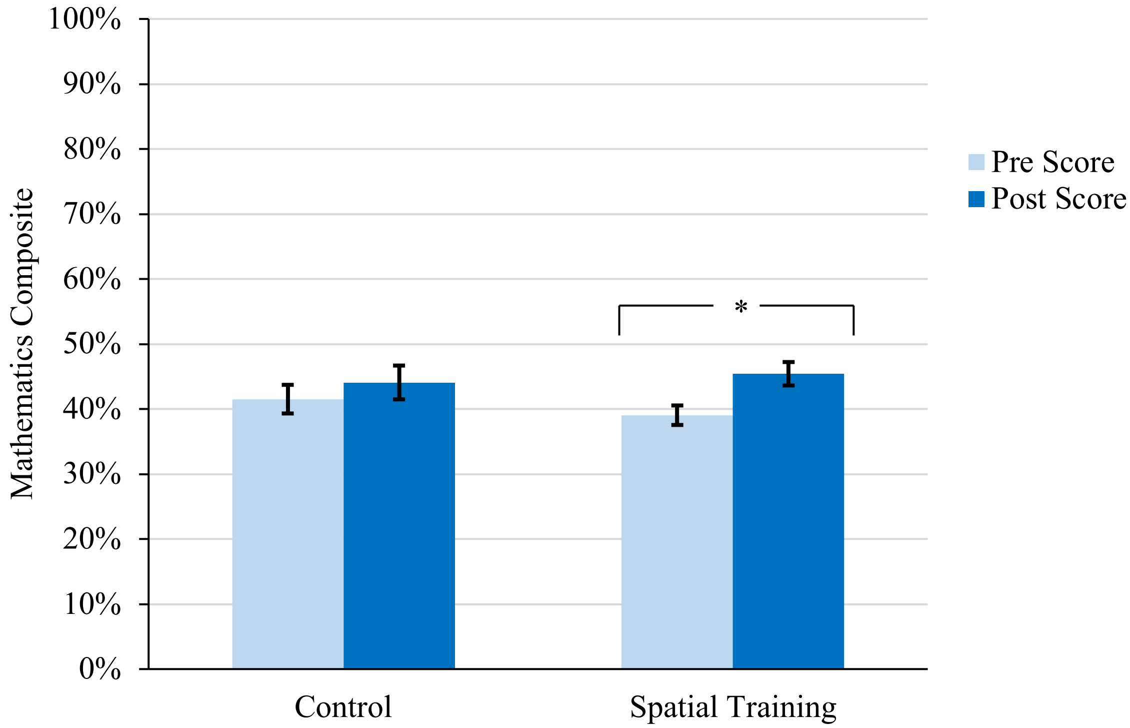

Figure 1B: Average composite mathematics pre- and posttest performance in 6th grade.

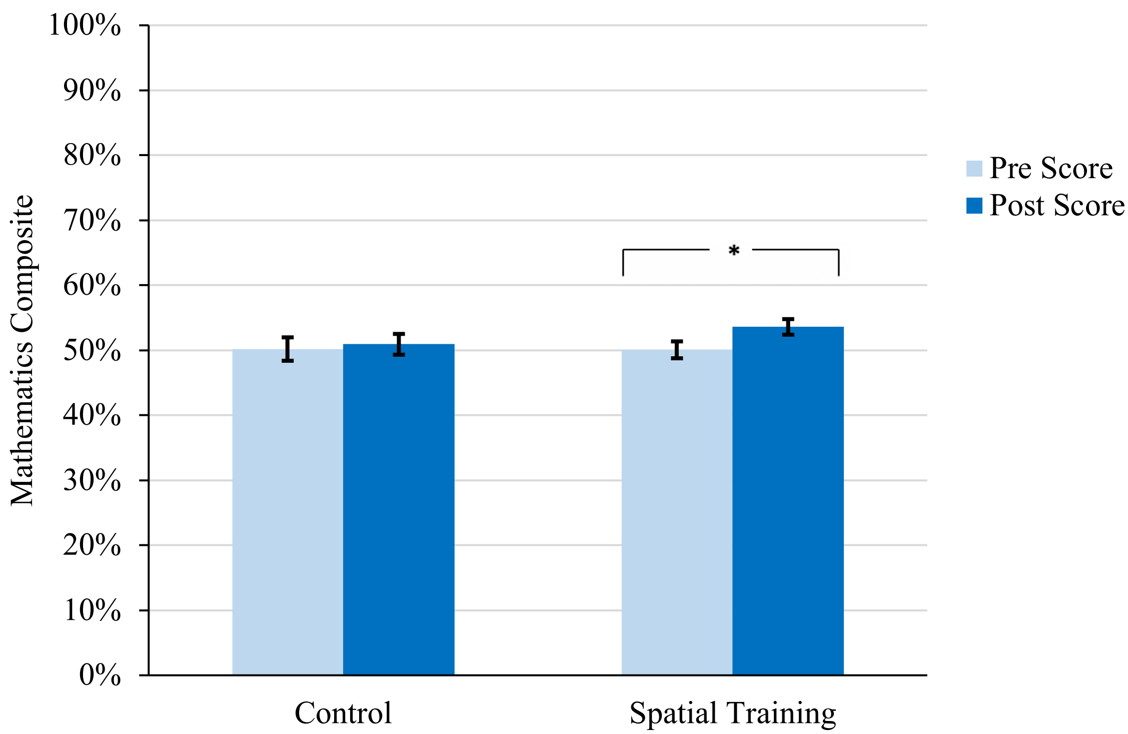

Figure 1. Mean percent correct on mathematics composite at pretest and posttest in first grade (A) and sixth grade (B) for children in the control and spatial training groups. Significant gains from pre- to posttest are indicated with an asterisk (one-tailed) p < .05. Error bars represent standard errors. See the online article for the color version of this figure.