

Original Articles

Spatial and mathematics skills: Similarities and differences related to age, SES, and gender[☆]

Tessa Johnson^a, Alexander P. Burgoyne^b, Kelly S. Mix^{c,*}, Christopher J. Young^d, Susan C. Levine^e

^a Department of Human Development and Quantitative Methodology, University of Maryland, United States

^b School of Psychology, Georgia Institute of Technology, United States

^c Department of Human Development and Quantitative Methodology, University of Maryland, United States

^d Consortium on School Research, University of Chicago, United States

^e Department of Psychology, University of Chicago, United States



ARTICLE INFO

Keywords:

Spatial skill
Mathematics
Sex differences
SES
Children
Development

ABSTRACT

Performance on a range of spatial and mathematics tasks was measured in a sample of 1592 students in kindergarten, third grade, and sixth grade. In a previously published analysis of these data, performance was analyzed by grade only. In the present analyses, we examined whether the relations between spatial skill and mathematics skill differed across socio-economic levels, for boys versus girls, or both. Our first aim was to test for group differences in spatial skill and mathematics skill. We found that children from higher income families showed significantly better performance on both spatial and mathematics measures, and boys outperformed girls on spatial measures in all three grades, but only outperformed girls on mathematics measures in kindergarten. Further, comparisons using factor analysis indicated that the income-related gap in mathematics performance increased across the grade levels, while the income-related gap in spatial performance remained constant. Our second aim was to test whether spatial skill mediated any of these effects, and we found that it did, either partially or fully, in all four cases. Our third aim was to test whether the “separate but correlated” two-factor latent structure previously reported for spatial skill and mathematics skill was (Mix et al., 2016; Mix et al., 2017) replicated across grade, SES, and sex. Multi-group confirmatory factor analyses conducted for each of these subgroups indicated that the same latent structure was present, despite differences in overall performance. These findings replicate and extend prior work on SES and sex differences related to spatial and mathematics skill, but provide evidence that the relations between the domains are stable and consistent across subgroups.

1. Introduction

To understand children's developmental outcomes, psychologists have long been interested in individual differences in both cognitive skills and school achievement. This research has established that children growing up in families of lower socio-economic status (SES) tend to perform worse on measures of cognitive skill, such as language, executive function, and spatial skill (Noble, McCandliss, & Farah, 2007). Low-

income children also show lower levels of academic achievement in the areas of reading and mathematics (Reardon, 2011). Individual differences in both cognitive skill and academic outcomes have also been examined in terms of sex, although as we will see, the evidence in this area is mixed (see Miller & Halpern, 2014, for a review). Although the extant literature provides evidence of individual differences in both cognitive skills and academic outcomes separately, few studies have examined how individual differences in cognitive skills and academic

[☆] This research was supported by Award R305A120416 from the U.S. Department of Education, Institute of Education Sciences to Kelly Mix and Susan Levine. 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. This research was also funded by Grant #2018-0680 from the Heising-Simons Foundation, Palo Alto, CA, USA, which supported the Development and Research in Early Math Education (DREME) Network (support to Susan C. Levine). The data are published here: <https://drum.lib.umd.edu/handle/1903/21090>.

* Corresponding author at: Department of Human Development and Quantitative Methodology, University of Maryland, 3304 Benjamin Building, College Park, MD 20742, United States.

E-mail address: kmix@umd.edu (K.S. Mix).

<https://doi.org/10.1016/j.cognition.2021.104918>

Received 23 September 2020; Received in revised form 19 September 2021; Accepted 24 September 2021

Available online 7 October 2021

0010-0277/© 2021 Elsevier B.V. All rights reserved.

outcomes are related, and whether these relations change over time. These relations are important to clarify because they are sometimes hypothesized to be causal. For example, it has been argued that one reason boys outperform girls in mathematics may be that boys have stronger spatial skills (e.g., Casey, Nuttall, & Pezaris, 1997). A first step investigating whether differences in cognitive skill underlie differences in academic outcomes is to confirm these relations in subgroups and understand how they differ across age.

The present study provides such an investigation for the relation of spatial skills and mathematics achievement via a secondary analysis of a previously published dataset (Mix et al., 2016; Mix et al., 2017). We address three specific questions. First, we ask whether performance on either spatial or mathematics tasks differed based on SES, sex, or both, in any of the three grades tested. Though this question does not pertain directly to the relation between spatial skill and mathematics, it contributes to the broader literature on individual differences in spatial skill and mathematics. Second, and more directly related to our main aim, we ask whether spatial skill mediated any observed effects of either SES or sex on mathematics. Although studies have revealed SES and gender differences in spatial and mathematics skills separately, research has not yet examined the relations between these skills for different subgroups of children, and how these relations might differ across age. Third, we ask whether the latent structure of spatial skill and mathematics differs for any of the SES- and sex-related subgroups at each of the three grade levels, which is directly relevant to the relation of spatial skill and mathematics for different subgroups. Because our dataset included multiple measures of spatial skill and mathematics at multiple age points, it is well-suited to address these critical questions.

2. SES differences

Studies have demonstrated significant relations between SES and a broad range of cognitive skills in childhood (Noble et al., 2007). Spatial skills are among those on which children from lower SES backgrounds perform worse than do children from middle or high SES backgrounds (Farah et al., 2006; Jirout & Newcombe, 2015; Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005). These SES effects have been obtained when controlling for language skill, suggesting that they are not due to differences in general cognitive ability (Noble et al., 2007).

There is also a large literature reporting SES differences in academic achievement, including mathematics (Cobb-Clark & Moschion, 2017; Denton & West, 2002; Sirin, 2005). One meta-analysis reported that on average, there is a medium effect of SES on mathematics achievement, though these effects are influenced by a number of variables that are correlated with SES, such as school location and students' race and ethnicity (Sirin, 2005). These differences appear to be weaker, though still significant, in younger versus older children, indicating a widening SES gap over time (e.g., Cobb-Clark & Moschion, 2017; Crane, 1996).

There is also evidence that SES differences can moderate other individual differences, such as sex differences in spatial skill. For example, Levine et al. (2005) found significant sex differences on spatial tasks, but only among middle- and high-income children. For children at lower SES levels, there were no significant differences in spatial performance between the sexes, with both boys and girls performing at lower levels than their higher SES peers. Note that such moderating effects are not always obtained (e.g., Wai, Lubinski, & Benbow, 2009) perhaps due to task differences (See Levine, Foley, Lourenco, Ehrlich, & Ratliff, 2016 for a discussion). Also, the origins of SES and sex differences in spatial performance are complex, with evidence pointing to both genetic and environmental factors (Hackman & Farah, 2009; Levine et al., 2016). In the present study, we ask whether the resulting spatial differences relate to mathematics performance differently at different ages, as well as relative to children's sex and SES background.

3. Sex differences

Significant sex differences have been reported in both spatial skill and mathematics, although the patterns differ somewhat for the two domains. For spatial skill, a male advantage emerges in infancy (Moore & Johnson, 2008; Quinn & Liben, 2008), is evident throughout childhood (Levine et al., 2005; Levine et al., 2016; Lippa, Collaer, & Peters, 2010; Voyer, Voyer, & Bryden, 1995), and remains evident in adulthood (Linn & Petersen, 1985; Silverman et al., 2000). Despite this apparent consistency, it should be noted that not all studies have revealed significant sex differences in spatial skill, suggesting that the effects may be task-specific (e.g., Spelke, Gilmore, & McCarthy, 2011). Further, a recent meta-analysis indicated that the magnitude of the difference appears to increase with age (Lauer, Yhang, & Lourenco, 2019). Nonetheless, the bulk of the evidence suggests a male advantage on certain spatial tasks that is evident to some extent, irrespective of age. In addition to differences in overall performance, the evidence also suggests that boys and girls differ in their approaches to spatial tasks. Specifically, males appear to use more holistic spatial strategies, such as mentally rotating a figure in its entirety, whereas females tend to use shape cues and part-by-part solutions (Geiser, Lehmann, Corth, & Eid, 2008; Heil & Jansen-Osmann, 2008; Pezaris & Casey, 1991; Wang & Carr, 2014).

The evidence for sex differences in mathematics is less straightforward (see Herts & Levine, 2020, for a review). Large scale testing and meta-analyses suggest there is a small male advantage on mathematics tests, but that the effect is most apparent among high performing students and for complex problem-solving tasks (Lindberg, Hyde, Petersen, & Linn, 2010; Reilly, Neumann, & Andrews, 2015). Also, even though girls sometimes perform worse than their male peers on high stakes standardized mathematics tests, they tend to earn higher mathematics grades in school (Duckworth & Seligman, 2006; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006; Pomerantz, Altermatt, & Saxon, 2002). Further, the male advantage in mathematics may not appear until adolescence, perhaps due to the nature of the measures used in that age range, or children's familiarity with the mathematics being tested, with males performing better than females when problems are less familiar (Ceci, Williams, & Barnett, 2009). Among younger children, most of the evidence suggests either no sex differences (Bakker, Torbeyns, Wijns, Verschaffel, & De Smedt, 2019; Hutchison, Lyons, & Ansari, 2019; Kersey, Csumitta, & Cantlon, 2019; Lachance & Mazocco, 2006; Lindberg et al., 2010; McGraw, Lubienski, & Strutchens, 2006) or a female advantage on routinized operations such as counting and computation (Duckworth & Seligman, 2006; Hutchison et al., 2019; Hyde, Fennema, & Lamon, 1990). One exception may be a small male advantage in kindergarten reported among the very highest mathematics performers (top 10%) (Penner & Paret, 2008).

4. Developmental relations between spatial skill and mathematics

Research has established that spatial skill and mathematics achievement are related over developmental time, in potentially causal ways. Longitudinal studies have shown that spatial skill is associated with achievement in science, technology, engineering, and mathematics (STEM) disciplines (Wai et al., 2009), as well as directly predicting mathematics outcomes across development (Frick, 2018; Geer, Quinn, & Ganley, 2019; Gilligan, Flouri, & Farran, 2017; Lauer & Lourenco, 2016). For example, Lauer and Lourenco (2016) demonstrated that spatial skill in infancy predicted both spatial skill and number sense at age 4 years. Mix et al., 2016; Mix et al., 2017 reported that spatial skill and mathematics were highly related in each of three age groups (kindergarten, 3rd grade, and 6th grade). Studies have also shown that training to improve spatial skill can lead to measurable improvement in mathematics (Cheng & Mix, 2014; Cheung, Sung, & Lourenco, 2019; Lowrie, Logan, & Ramful, 2017; Mix, Levine, Cheng, Stockton, & Bower,

2020). If spatial skill and mathematics are causally related, then it is possible this relation contributes to the sex and SES differences reviewed above, within each domain.

To illustrate, recall that boys and girls have exhibited different approaches to spatial tasks, with boys tending to use holistic strategies and girls tending to focus on parts. Sex differences in strategy use have also been reported for mathematics, with elementary school boys using more independently invented strategies and mental models, and girls tending to use more conventional or concrete strategies (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998). Interestingly, however, these strategy differences did not lead to performance differences—both boys and girls obtained similar scores on a mathematics test. Still, it is possible that superior spatial skill supports the use of abstract mathematics strategies, such as generating mental models, that may impact performance, particularly on challenging or novel problems that may not have been tested in prior research studies. Better spatial skill might also improve mathematics performance by helping students track their position in a complex written algorithm (Mix, 2019), which may only be apparent on complex problems for which tracking is needed. This overall line of thinking raises the issue that spatial training might close the gender gap in mathematics outcomes, though existing studies have failed to show differential improvement in spatial skill for females following training (see Newcombe, 2017, for a discussion).

Similar explanations have been posited to explain SES differences in spatial and mathematics performance. For example, one explanation for the different levels of mathematics performance for children from high versus low SES families is their lower levels of spatial visualization skill (Casey, Dearing, Vasilyeva, Ganley, & Tine, 2011). A recent investigation of student achievement in STEM education also found that spatial skill mediated SES-differences in computer science course grades (Parker et al., 2018). In line with this, investigators have asked whether spatial training might close the SES-linked STEM achievement gap (Newcombe, 2017). Others have emphasized the role of informal home experiences and parent input for specific academic content areas (Elliott & Bachman, 2018). Interestingly, there have been links reported between parent input and spatial skill (Pruden, Levine, & Huttenlocher, 2011), with parents from low SES backgrounds providing less spatial input on average (Dearing et al., 2012; Verdine, Golinkoff, Hirsh-Pasek, & Newcombe, 2017). Thus, it is possible that parent input affects mathematics achievement directly via differences in mathematics support, and indirectly via support for spatial skill development.

Only a few studies have examined these possibilities. With regard to SES, the findings have been significant but mixed. Casey et al. (2011) found that fourth grade students in more affluent families outperformed those from lower income families across various spatial and mathematics measures, yet also reported SES-linked differences in the way the two domains related. Specifically, children from higher SES families had a strong relation between spatial skill and mathematics, but this was not the case among lower SES children. In contrast, Demir, Prado, and Booth (2015) reported the opposite pattern for children of roughly the same age (9 to 12 years)—spatial regions of the brain were engaged during subtraction problem solving, in children from lower, but not higher SES families.

The findings regarding sex differences in the relation of spatial skill and mathematics are also mixed. In a study of 2nd and 4th grade students, Moè (2018) found that boys and girls showed similar correlations between spatial skill and mathematics. In contrast, a study of kindergarten students found that spatial performance was significantly related to mathematics performance in boys, but not girls. For girls, only verbal processing was predictive of mathematics scores, and for the most part, verbal processing did not predict mathematics scores for boys (Klein, Adi-Japha, & Hakak-Benizri, 2010). Taken together, these findings raise questions about whether differences in spatial skill explain differences in mathematics performance for either low-income children or girls, at least in the lower elementary school grades.

However, a different pattern emerges from evidence among

adolescents. In separate studies, Casey et al. (1997); Casey, Nuttall, and Pezaris (2001) demonstrated that the male advantage on both the SAT mathematics subtest (SAT-M) and the Third International Mathematics and Science Study (TIMSS) (eighth grade students) was mediated by level of spatial skill. Interestingly, sex differences on the SAT-M were eliminated when mental rotation skill on a 3-D task was statistically controlled (Casey, Nuttall, Pezaris, & Benbow, 1995). Fennema and Tartre (1985) also found that middle school students with stronger spatial visualization skills tended to use diagrams and drawings in problem solving more frequently, though this strategy difference did not lead to differences in problem solving performance. Whereas girls were more likely than boys to use these spatial strategies, their problem-solving accuracy was actually lower than that of boys, perhaps because their weaker spatial skills led to less effective spatial strategies. Overall, these results suggest that better spatial skills may contribute to the albeit limited male advantage on these mathematics tests, even though non-spatial mathematics strategies may be equally effective, at least for the problems tested in existing research studies.

5. The present study

The research reviewed above leads to several conclusions. First, there is strong and consistent evidence that children from higher SES families perform better on both spatial and mathematics tasks. Second, although there is strong evidence that boys outperform girls on spatial tasks, the evidence is mixed for sex-related differences in mathematics. There is abundant research showing that spatial skill and mathematics skill are related, and the strength of this relation seems stable over development. However, the specific ways that spatial skill and mathematics achievement are related may vary across age, and it is possible that similar variations exist for subgroups of children based on SES and sex. The few studies examining these potential differences have yielded mixed results.

The present study takes a fresh look at these questions using a large cross-sectional data set that includes three grade levels (kindergarten, third, sixth) and twelve measures (six spatial measures, up to seven mathematics measures, and one measure of vocabulary intended to tap general cognitive ability). Although the measures were adapted to be age-appropriate, they were conceptually constant across the grade levels. This consistency permitted us to compare performance more directly than is possible when comparing across studies using different methods and measures.

6. Method

Participants Data from 1592 children (753 boys and 839 girls) were collected in two cohorts using the same measures and age groups. The data from Cohort 1 ($n = 854$, collected in 2013–14) were submitted to an exploratory factor analysis reported previously (Mix et al., 2016). The data from Cohort 2 ($n = 738$, collected in 2014–15) were analyzed and reported as part of a multi-group confirmatory factor analysis (Mix et al., 2017). We refer readers to those articles for complete details related to grade, sex, population characteristics, exclusions, and so forth. All data were collected with informed consent from children's caregivers and the approval of the relevant institutional review boards. The sample size was determined based on a target of 220 children per age group per cohort, because a sample size of 20 children per measure (i.e., 11 measures) should ensure adequate power for latent variable analysis (MacCallum, Widaman, Zhang & Hong, 1999; Raykov & Marcoulides, 2006). We succeeded in testing more than the minimum sample both years, resulting in a far greater total number of children when the two cohorts were combined (i.e., kindergarten $n = 526$; third grade $n = 538$; sixth grade $n = 529$) and more than adequate power for the latent variable analyses reported here. Power estimates using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) also confirmed that our total sample size of 1592 children was adequate to detect a medium effect, (i.e., $f^2 =$

0.10 with 0.80 power) based on Cohen's (1992) suggested values for small, medium, and large effects ($f^2 = 0.02, 0.15, \text{ and } 0.35$) in an analysis of covariance (ANCOVA) with 36 groups (three grades by two sexes by six income levels) and one covariate.

Procedure Children completed a battery of tests that measured spatial skill, mathematics skill, and verbal skill. The tests were blocked based on whether they were individually- or group-administered but the order of tests within each block varied randomly. Further, the order of presentation for group versus individual tests was random and counterbalanced across children. The test battery was completed over three to four sessions to avoid fatigue (see for Mix et al., 2016; Mix et al., 2017 details).

Measures Reliabilities were estimated using Cronbach (1951) and were calculated from the combined dataset (Cohort 1 and Cohort 2). Most of the reliabilities approached or reached $\alpha = 0.70$, which is a generally accepted cut-off, though not a hard and fast rule (Lance, Butts, & Michels, 2006; Nunnally, 1978). In some cases (e.g., map reading) the reliabilities were lower than 0.70 which may reflect multidimensionality within the measure. Low internal consistency is known to attenuate relations among measures and as such could affect some of the relations tested here; however, some argue this risk has been overstated and measures with low reliability may still be useful if they provide meaningful content coverage (Schmitt, 1996).

Mental Rotation. (adapted from Neuburger, Jansen, Heil, & Quaiser-Pohl, 2011 and Peters et al., 1995). In the kindergarten/third grade version, small groups of children were shown four unfamiliar figures (i.e., forms based on manipulating components of capital letters) 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 task was introduced with four practice items on a laptop for which children received feedback that included animations with the correct answers rotating to match the target. The 16 test items were group-administered by presenting them in a paper booklet (kindergarten $\alpha = 0.74$; third grade $\alpha = 0.87$). The sixth-grade version was the same, except that stimuli were perspective line drawings of three-dimensional block constructions presented on paper. Children completed 12 items consisting of a target and four choice drawings, two of which could be rotated in the picture plane to match the target ($\alpha = 0.81$).

Visual Spatial Working Memory. (adapted from Kaufman & Kaufman, 1983). On each test trial, groups of children were shown a 14 cm \times 21.5 cm grid that was divided into squares (e.g., 3 \times 3, 4 \times 3, or 5 \times 5). Drawings of objects were displayed at random positions within the grid and left in full view for 5 s. Then a blank grid was displayed and children marked an X in the previously filled positions. Stimuli were presented on a laptop computer and children responded in paper test booklets. The test was introduced with two or three practice items, depending on grade, for which children received feedback and were allowed to compare their responses to the stimulus display. The test trials ($n = 15\text{--}29$, based on grade) were group-administered and began immediately after the final practice trial (kindergarten $\alpha = 0.77$; third grade $\alpha = 0.66$, and sixth grade $\alpha = 0.81$).

Test of Visual Motor Integration. (VMI, 6th ed., Beery & Beery, 2010). On each trial, children copied a line drawing of a geometric shape on a blank sheet of paper. There were 18–24 trials, depending on the age of the child, over which the figures became increasingly complex. We administered the test in small groups. The reliability of the VMI based on a split-half correlation (reported in the test manual) was 0.93.

Block Design Subtest. (WISC-IV) (Wechsler et al., 2004).

On each trial, children were shown a printed figure comprised of white and red sections, and they 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. Children completed different numbers of items depending on their basal and ceiling performance. The reliability coefficient reported in the WISC-IV manual is

between 0.83 and 0.87 depending on age.

Map Reading. (adapted from Liben & Downs, 1989). Children were shown a location on a model and then indicated where it would appear in a corresponding map. Kindergarten and third grade students completed 14 test trials in which the model was a full color 3-dimensional model town with buildings, roads, a river, and trees. Sixth grade students completed 8 trials in which the model was a full-color screenshot of a virtual model town. Children marked the corresponding location on a black and white, 2-dimensional, scale map. In the sixth grade task we also manipulated the presence of landmarks. Feedback was given on the first three test questions to ensure that children understood the task. Sixth graders completed the test in groups whereas younger children were tested individually (kindergarten $\alpha = 0.62$; third grade $\alpha = 0.69$; sixth grade $\alpha = 0.56$).

Perspective Taking (adapted from Frick, Mörhing, & Newcombe, 2014). In this individually-administered task, kindergarten and third grade children were shown a set of Play Mobil figures and asked which of four pictures was taken from each figure's perspective. The 27 test questions were preceded by 4 practice items with feedback (kindergarten $\alpha = 0.64$; third grade $\alpha = 0.87$). Sixth grade children saw six to eight figures arranged in a circle and indicated their angle of view from a particular position, by drawing an arrow toward the center object (Kozhevnikov & Hegarty, 2001). There were two practice items with feedback and 12 test items. Scores were based on the number of degrees children's responses deviated from the correct angle on each item ($\alpha = 0.83$).

Place Value. Kindergarten and third grade students completed a set of 20 items that required them to compare, order, and interpret multi-digit numerals (e.g., Which number is in the ones place?), as well as match multi-digit numerals to their expanded notation equivalents (342 = 300 + 40 + 2) (kindergarten $\alpha = 0.77$; third grade $\alpha = 0.81$). Sixth grade students completed the Rational Numbers subtest from the Comprehensive Mathematics Ability Test (CMAT) (Hresko, Schlieve, Herron, Sawain, & Sherbenou, 2003). These items similarly required them to compare, order, and interpret written numbers, but the numbers were a mixture of multi-digit whole numbers, fractions, and decimals ($\alpha = 0.83$). These measures were individually-administered in all three grades.

Word Problems. Kindergarten and third grade students were assessed using 12 word problems included in the Test of Early Mathematics Ability-Third Edition (TEMA-3, Ginsburg & Baroody, 2003) (kindergarten $\alpha = 0.73$; third grade $\alpha = 0.65$). The test was individually administered following the instructions in the test manual. Sixth grade students completed the Problem Solving subtest from the CMAT ($\alpha = 0.76$).

Calculation. We used a group-administered test consisting of age-appropriate arithmetic problems (kindergarten: $\alpha = 0.76$; third grade: $\alpha = 0.70$; sixth grade: $\alpha = 0.76$). In kindergarten, the problems were one- to four-digit whole number addition and subtraction problems. In third grade, whole number multiplication and division problems (one to three digits) were added. The sixth grade calculation test was similar but included both whole numbers and decimals.

Missing Term Problems/Algebra. In missing term problems, children find the solution to a calculation problem where the missing value is not the sum or difference (e.g., 7 + X = 15). Kindergarten and third grade students completed 8 such problems (kindergarten: $\alpha = 0.62$; third grade: $\alpha = 0.71$). To increase the difficulty level for sixth grade students, the CMAT Algebra Subtest was administered ($\alpha = 0.68$).

Number Line Estimation. (Siegler & Opfer, 2003). All children were tested in small groups ($n = 4\text{--}6$). Children were asked to mark where a written numeral would go on a number line, anchored with a written numeral at each end. The anchor points and the stimulus values varied by grade. Specifically, kindergarteners placed the numerals 4, 17, 33, 48, 57, 72 and 96 on a 0-to-100 number line (split half reliability: $r = 0.39$); third graders placed 3, 103, 158, 240, 297, 346, 391, 907 on a 0-to-1000 number line (split half reliability: $r = 0.48$); and sixth graders

placed 25,000, 61,000, 49,000, 5000, 11,000, 2000, 15,000, 73,000, 8000, 94,000 on a 0-to-100,000 number line (split half reliability: $r = 0.61$). Children's performance was evaluated based on the linearity of their placements. That is, we regressed each child's responses against the measurements for the correct placements and used the R^2 values for these regressions as their number line estimation scores in subsequent analyses.

Fraction Concepts. Fraction items were not included in the kindergarten test battery. In third grade, there were four items that tested fraction equivalence and simple calculation with common denominators ($\alpha = 0.57$). In sixth grade, children completed both a 22-item test with fraction comparisons, calculation with and without common denominators, and calculation with mixed numbers ($\alpha = 0.73$), and a version of the number line estimation task in which the anchors are 0 and 1, and the stimulus quantities are all fractions (i.e., $\frac{1}{4}$, $\frac{1}{19}$, $\frac{2}{3}$, $\frac{7}{9}$, $\frac{1}{7}$, $\frac{3}{8}$, $\frac{5}{6}$, $\frac{4}{7}$, $\frac{12}{13}$, $\frac{1}{2}$) (split half reliability: $r = 0.46$) (Siegler, Thompson, & Schneider, 2011).

Supplemental Sixth Grade Tests. To assess the breadth of mathematics skills in older children, sixth grade students completed two additional measures: CMAT Charts and Graphs ($\alpha = 0.79$) and CMAT Geometry ($\alpha = 0.66$). For Charts and Graphs, students are shown data in graphic form and asked questions that require them to interpret the information. For Geometry, they defined geometric terms, determined unknown angles, identified shapes, and so forth.

General Cognitive Ability. To estimate and control for children's general cognitive ability, children completed the Picture Vocabulary subtest from the *Woodcock-Johnson Test of Achievement-3 (WJ-3)*. Although not a comprehensive cognitive assessment, previous studies have demonstrated a strong relation between vocabulary and intelligence scores, suggesting that vocabulary is a reasonable proxy for general cognitive ability (e.g., Sattler, 2001; Woodcock, McGrew, & Mather, 2001). On each item, children were asked to point to or name a picture. The test was individually administered according to the instructions in the test manual (kindergarten: $\alpha = 0.73$; third grade: $\alpha = 0.77$; sixth grade: $\alpha = 0.74-79$).

7. Analysis approach

The present study is a series of secondary data analyses aimed at addressing three questions. First, to determine whether performance on either spatial or mathematics tasks differed based on income, sex, or both, in any of the three grades tested, we used analyses of covariance (ANCOVA). Effect sizes are reported using partial eta squared (η^2_p), where a small effect corresponds to $\eta^2_p = 0.01$, a medium effect corresponds to $\eta^2_p = 0.06$, and a large effect corresponds to $\eta^2_p = 0.14$ (Faul, Erdfelder, Lang, & Buchner, 2007). Later in the report, we repeated these analyses using potentially more sensitive factor analysis. Analyses conducted within a factor analytic model allow us to relax several important measurement assumptions—namely the assumption of no measurement error on the observed measures and the assumption that the latent factor is identically indicated by each observed measure—which thus allowed us to generate unbiased estimates of the relations among the latent factors as well as their reliability (Dunn, Baguley, & Brunson, 2013; McNeish, 2018; Raykov, 2001; Reuterberg & Gustafsson, 1992). Effect sizes in this section are reported using standardized mean differences and can be categorized as small ($\delta = 0.2$), medium ($\delta = 0.5$), or large ($\delta = 0.8$), as discussed in Cohen (1992).

Second, to determine whether spatial skill was a significant mediator, we used the PROCESS model 4 macro for SPSS (Hayes, 2017). These analyses determined whether level of spatial skill mediated the effects of either SES or sex on children's composite mathematics scores.

Third, to determine whether the latent factor model reported in previous studies (Mix et al., 2016; Mix et al., 2017) accurately characterized the data for specific SES and sex subgroups, we tested the configural invariance of the mathematics and spatial skill constructs using a multi-group confirmatory factor analysis (MG-CFA). Specifically, we

asked whether the mean structure and covariance matrices were statistically equivalent across SES and sex at each grade level (French & Finch, 2006; Sass, 2011), and we checked the validity of the hypothesized factor structure by applying it to various groups simultaneously (Byrne, Shavelson, & Muthén, 1989; Van de Schoot, Lugtig, & Hox, 2012). By evaluating the differences in factor loadings and item intercepts in the combined model (Van de Schoot et al., 2012), we were able to then compare the means and variances of the latent math and spatial factors across subgroups.

Children were grouped for sex based on family report recorded on the informed consent form. Children's sex was reported by 100% of families. To determine children's SES, we used estimates of household income¹ derived from two sources. One source was family responses to an optional demographic questionnaire attached to the informed consent. On this questionnaire, families were asked to report their household income on a 1 to 6 scale where 1 \leq \$15,000, 2 = \$15,000 to 34,999; 3 = \$35,000 to 49,999; 4 = \$50,000 to 74,999; 5 = \$75,000 to 99,999; and 6 = \$100,000 or more. Of the 1592 families, 1119 (70%) reported their income on this scale. A second source of SES information was the percentage of children at each school receiving free or reduced price lunch. By using rates of free or reduced school lunch, we could estimate SES for 100% of the children in the sample; however, these estimates were specific to schools, not specific to individual children and thus, were less precise than parent report. Preliminary analyses indicated that the two measures of SES were highly but not perfectly correlated ($r = 0.44$). However, we found that the pattern of results was similar for either coding approach in the analyses we conducted. Also, previous meta-analyses have reported similar effect sizes for both ways of estimating SES, though it was also noted that family reported SES yielded larger effects than secondary estimates (Sirin, 2005; but see Perry & McConney, 2010). Thus, we report below only the results for the analyses based on the income codes provided for a subset of children. Children for whom income data were not reported were excluded ($n = 473$). Power estimates indicated that this sample size was adequate to detect a medium effect size ($\eta^2_p = 0.06$).

8. Results

8.1. Aim 1: Group differences in mean performance

To assess differences in skill level across grade, sex, and income level, we computed composite variables by averaging z-scores (i.e., standardized scores) for the six spatial measures, and the five to nine (depending on grade) mathematics measures. Measures for which lower scores represented better performance were reverse-coded prior to computing the composite variables. Scores on these variables represent students' overall performance relative to other students within their respective grades by setting the means of each grade to approximately zero (see Table 1). Graphs depicting the spatial and mathematics composite scores of girls versus boys are presented in Figs. 1 and 2. The correlation between spatial and mathematics performance was positive and statistically significant in kindergarten ($r = 0.52$, $p < .001$), third grade ($r = 0.64$, $p < .001$), and sixth grade ($r = 0.61$, $p < .001$).

We next conducted separate three-way ANCOVAs with spatial skill and mathematics skill as dependent measures, children's WJ-3 Vocabulary subtest scores as the covariate, and sex, income, and grade as between-subjects factors. We also compared children's performance across grades by sex using two-tailed *t*-tests. In some cases, we used these *t*-tests to understand group differences in the absence of significant

¹ Unfortunately, we did not query parental education level on our questionnaire, which might have provided a more accurate estimate of SES than income levels alone. However, because income level and education level are strongly correlated (Hauser & Warren, 1997), it is unlikely that our main findings would have been significantly altered.

Table 1
Mean spatial and mathematics performance by income level and sex, within each grade level.

Variable		Kindergarten		Third grade		Sixth grade	
		Spatial	Mathematics	Spatial	Mathematics	Spatial	Mathematics
Income level	1	-0.27 (0.53)	-0.21 (0.67)	-0.42 (0.56)	-0.48 (0.62)	-0.43 (0.75)	-0.53 (0.76)
	2	-0.31 (0.58)	-0.26 (0.59)	-0.33 (0.74)	-0.23 (0.72)	-0.28 (0.68)	-0.35 (0.71)
	3	-0.09 (0.60)	-0.17 (0.64)	-0.18 (0.72)	-0.13 (0.72)	-0.05 (0.68)	-0.10 (0.66)
	4	-0.01 (0.59)	-0.02 (0.67)	-0.01 (0.74)	-0.07 (0.78)	0.09 (0.68)	0.10 (0.60)
	5	0.09 (0.62)	0.11 (0.72)	0.26 (0.60)	0.21 (0.66)	0.18 (0.70)	0.10 (0.67)
	6	0.22 (0.75)	0.22 (0.86)	0.17 (0.69)	0.17 (0.66)	0.22 (0.77)	0.42 (0.74)
Sex	Boys	0.08 (0.71)	0.10 (0.82)	0.14 (0.76)	0.06 (0.70)	0.09 (0.72)	0.01 (0.72)
	Girls	-0.08 (0.61)	-0.10 (0.65)	-0.10 (0.68)	-0.04 (0.74)	-0.08 (0.74)	-0.01 (0.76)

Note. Standard deviations are presented in parentheses.

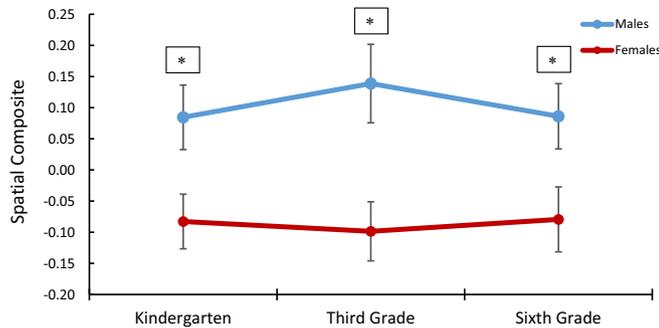


Fig. 1. Spatial Performance for Girls and Boys Within Each Grade.
Note. Error bars represent ± 1 standard error of the mean. * $p < .05$.

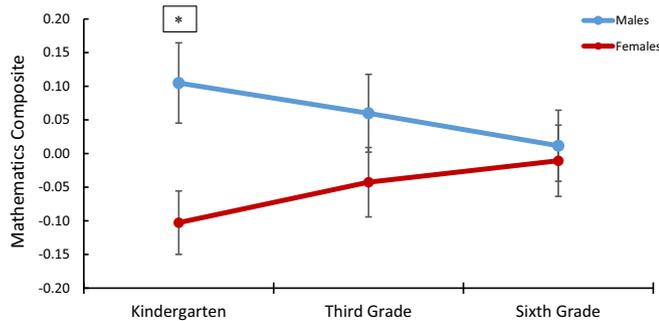


Fig. 2. Mathematics Performance for Girls and Boys Within Each Grade.
Note. Error bars represent ± 1 standard error of the mean. * $p < .05$.

interactions, following Hsu (1999).

8.1.1. Spatial skill

There were significant main effects of sex and income on spatial tasks such that boys outperformed girls ($F(1, 1082) = 16.265, p < .001, \eta^2_p = 0.015$) and children from higher income families outperformed children from lower income families ($F(5, 1082) = 11.011, p < .001, \eta^2_p = 0.048$). The main effect of grade was not significant ($F(2, 1082) = 0.012, p = .988, \eta^2_p = 0.000$), which was expected because the composite variable for spatial skill set the means for each grade to zero. Interestingly, the sex \times grade interaction was not significant ($F(2, 1082) = 0.090, p = .914, \eta^2_p = 0.000$), indicating that the magnitude of sex differences in spatial skill did not differ across grades. This pattern was supported by t -tests showing that boys significantly outperformed girls on spatial measures in all three grades (kindergarten: $t(376) = 2.470, p = .014, d = 0.25$; third grade: $t(347) = 3.066, p = .002, d = 0.33$; sixth grade: $t(390) = 2.236, p = .026, d = 0.23$). Sex differences in spatial skill remained significant in all three grades after adjusting for a false discovery rate of $Q = 0.05$ (Benjamini & Hochberg, 1995). Neither the

income \times grade interaction ($F(10, 1082) = 0.662, p = .760, \eta^2_p = 0.006$), nor the income \times sex interaction ($F(5, 1082) = 1.963, p = .082, \eta^2_p = 0.009$) was significant. Thus, the same advantage for higher income children on spatial measures was apparent across grades and for both boys and girls.

8.1.2. Mathematics skill

Children from higher income families outperformed children from lower income families ($F(5, 1082) = 11.678, p < .001, \eta^2_p = 0.051$) on mathematics tasks but there was not a significant sex difference ($F(1, 1082) = 2.936, p = .087, \eta^2_p = 0.003$). As expected, the main effect of grade was not significant ($F(2, 1082) = 0.140, p = .869, \eta^2_p = 0.000$), because the composite variable for mathematics skill set the mean to zero for each grade. Similar to the results for spatial skill, the sex \times grade interaction was not significant ($F(2, 1082) = 1.760, p = .172, \eta^2_p = 0.003$); however, t -tests revealed that boys significantly outperformed girls on mathematics measures in kindergarten ($t(376) = 2.735, p = .007, d = 0.28$), but not in either third grade ($t(347) = 1.311, p = .191, d = 0.14$) or sixth grade ($t(390) = 0.298, p = .766, d = 0.03$; see Fig. 2). The same pattern of results was obtained after adjusting for a false discovery rate of $Q = 0.05$. The income \times grade interaction did not reach significance but generated a small effect ($F(10, 1082) = 1.219, p = .274, \eta^2_p = 0.011$) that reflected a widening gap across grades. There was no interaction of sex \times income ($F(5, 1082) = 1.210, p = .302, \eta^2_p = 0.006$).

8.2. Aim 2: mediation of income and sex differences in mathematics by spatial skill

Next, we tested whether the effect of income on mathematics was mediated by spatial skill (see Fig. 3). We included participants from all three grades in this analysis because there was a significant effect of income on mathematics skill at each grade level. In the full model ($R^2 = 0.40, F(3, 1115) = 243.90, MSE = 0.33, p < .001$), which included income and spatial skill as predictors of mathematics with WJ-3 Picture Vocabulary as a covariate, the direct effect of income on mathematics was statistically significant ($\beta = 0.11, b = 0.048, SE = 0.011, t = 4.327, p < .001$), as was the effect of spatial skill on mathematics ($\beta = 0.48, b = 0.503, SE = 0.027, t = 18.695, p < .001$) and the effect of vocabulary scores on mathematics ($\beta = 0.19, b = 0.140, SE = 0.019, t = 7.384, p < .001$). The effect of income on spatial skill was also significant ($\beta = 0.19, b = 0.084, SE = 0.012, t = 6.938, p < .001$). As a test of mediation, the indirect effect of income on mathematics through spatial skill was $b = 0.042 (SE = 0.007)$ and the 95% confidence interval based on bootstrapping with 5000 random samples did not include zero (95% CI [0.029, 0.056]). The significant indirect effect coupled with the significant direct effect indicates that spatial skill partially mediated the effect of income on mathematics. Supplemental analyses revealed that children's vocabulary scores also mediated the effect of income on mathematics skill, when controlling for spatial skill (indirect effect: $b = 0.015, SE = 0.003, 95\% CI [0.009, 0.023]$). Subsequently, we tested a mediation model in which both spatial skill and vocabulary scores were simultaneously specified as mediators of the relationship between

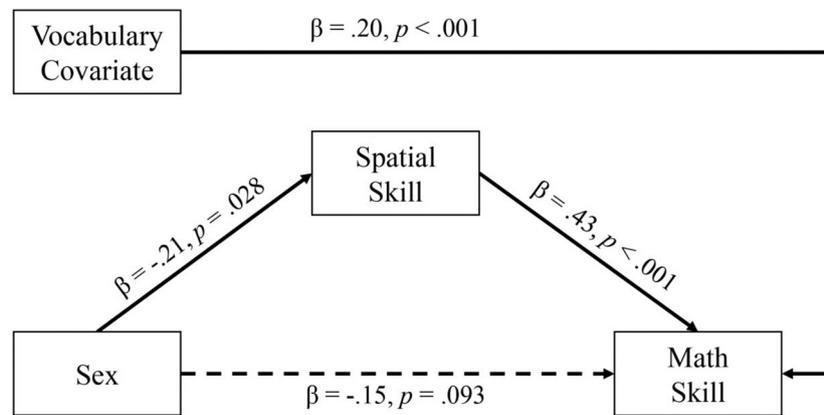


Fig. 3. Spatial Skill Fully Mediated the Effect of Sex on Math Skill in Kindergarten.

Note. The indirect effect from sex through spatial skill to math skill was significant, $b = -0.068$ ($SE = 0.032$), 95% CI = $-0.134, -0.009$.

income and mathematics skill. The results revealed that spatial skill (indirect effect: $b = 0.063$, $SE = 0.007$, 95% CI $[0.049, 0.078]$) and vocabulary scores (indirect effect: $b = 0.024$, $SE = 0.004$, 95% CI $[0.015, 0.033]$) both partially mediated the effect of income on mathematics skill, while the direct effect of income on mathematics skill still remained significant ($\beta = 0.11$, $b = 0.048$, $SE = 0.011$, $t = 4.327$, $p < .001$). This pattern of results indicates that both spatial skill and a verbal measure that is related to general cognitive ability partly explain the effect of income on mathematics skill.

We also tested whether the effect of sex on mathematics skill in kindergarten was mediated by spatial skill (see Fig. 4). We restricted this analysis to kindergarten participants because only in kindergarten was there a significant effect of sex on mathematics skill. In the full model ($R^2 = 0.31$, $F(3, 374) = 55.73$, $MSE = 0.39$, $p < .001$), which included sex and spatial skill as predictors of mathematics skill with WJ-3 Picture Vocabulary as a covariate, the direct effect of sex on mathematics was not statistically significant ($\beta = -0.15$, $b = -0.108$, $SE = 0.064$, $t = -1.68$, $p = .093$). However, there were significant effects of both spatial skill on mathematics ($\beta = 0.43$, $b = 0.487$, $SE = 0.052$, $t = 9.38$, $p < .001$), and sex on spatial skill ($\beta = -0.21$, $b = -0.140$, $SE = 0.064$, $t = -2.21$, $p = .028$). There also was a significant effect of vocabulary scores on mathematics ($\beta = 0.20$, $b = 0.152$, $SE = 0.034$, $t = 4.44$, $p < .001$). As a test of mediation, the indirect effect of sex on mathematics through spatial skill was $b = -0.068$ ($SE = 0.032$) and the 95% confidence interval based on bootstrapping with 5000 random samples did not include zero (95% CI $[-0.134, -0.009]$). The significant indirect effect coupled with the non-significant direct effect indicates that spatial skill fully mediated the effect of sex on mathematics in kindergarten.

In contrast, supplemental analyses revealed that vocabulary scores did not mediate the effect of sex on mathematics skill in kindergarten when controlling for spatial skill (indirect effect: $b = -0.004$, $SE = 0.015$, 95% CI $[-0.035, 0.026]$), indicating that spatial skill, and not a measure of general cognitive ability, accounted for the effect of sex on mathematics skill in kindergarten.

In summary, analyses related to our first aim revealed large effects of income level in all three grades, favoring children from high income families on both spatial and mathematics measures, a small effect favoring boys on the spatial measures in all grades, and a small effect favoring boys on mathematics in kindergarten only. With regard to our second aim, we found that spatial skill was a significant mediator of the observed effects of income (partial mediation) and sex (full mediation) on mathematics performance.

8.3. Aim 3: group differences in latent structure

8.3.1. Configural invariance model

In previous exploratory and confirmatory factor analyses of the

current data set (Mix et al., 2016; Mix et al., 2017), the best fitting model suggested two highly correlated factors—one comprised of spatial measures and the other comprised of mathematics measures—and the same structure was evident at all three grade levels. This pattern has been reported for a similar dataset comprised of 4- to 11-year-olds (Hawes, Moss, Caswell, Seo, & Ansari, 2019). Thus, a third aim of the present paper was to determine whether this same two-factor structure characterized the data for children from different SES groups, as well as boys versus girls.

In order to assess group differences on the latent constructs, we followed a stepwise model-building procedure (Anderson & Gerbing, 1988). We began by estimating single-factor models for each grade level individually. Once models were deemed to have adequate fit, the two-factor models were estimated at each grade level, allowing the factors to correlate with one another without imposing a causal direction. As in previously published analyses, the effect of general cognitive ability was controlled by partialling out the effect of WJ-3 Vocabulary subtest scores (Mix et al., 2016; Mix et al., 2017), and using the residualized variables in subsequent analyses. All models were estimated using Yuan-Bentler scaled maximum likelihood (MLR; Yuan & Bentler, 2000) to correct for non-normality within the observed items. Models were estimated using Mplus version 8.2 (Muthén & Muthén, 1998-2017).

Single-factor models were specified following results from previous analyses (Mix et al., 2016; Mix et al., 2017). At all three grade levels (kindergarten, third grade, & sixth grade), the spatial factor was measured by all six of the spatial tasks, and the mathematics factor was measured by all of the grade-appropriate mathematics tasks. This finding replicates the unitary within-domain structure reported by (Mix et al., 2016; Mix et al., 2017), as one might expect given that we are using the same dataset. Also as expected, model fit to the data was found to be adequate across all models, as assessed by the root mean squared error of approximation (RMSEA – at or below 0.07) and the standardized root mean square residual (SRMR – at or below 0.08) following recommendations by Hu and Bentler (1999) and Steiger (2007).

We next specified a series of two-factor configural invariance models, separately estimating parameters for each subgroup (i.e., low income vs.

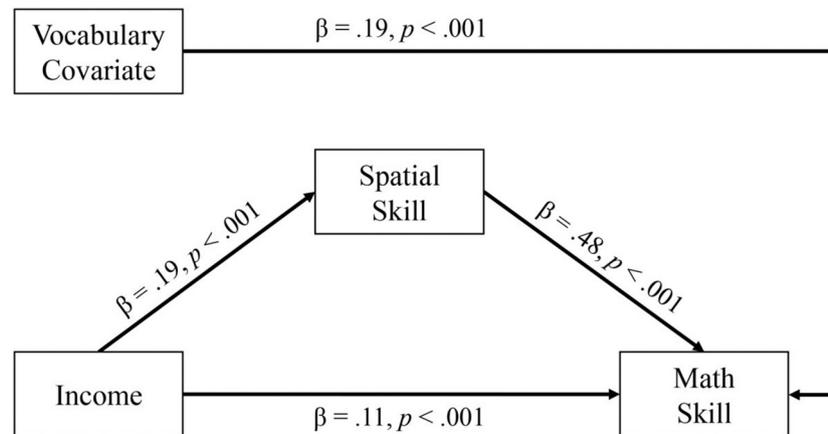


Fig. 4. Spatial Skill Partially Mediated the Effect of Income on Math Skill.

Note. The indirect effect from income through spatial skill to math skill was significant, $b = 0.042$ ($SE = 0.007$), 95% CI = [0.029, 0.056].

high income²; boys vs. girls) at each grade level, in which both the mathematics and spatial factors were estimated. Across all six models, the fit indices were within acceptable ranges, indicating that the same correlated two-factor spatial and mathematics structure previously found in the whole group analyses also held for these sex and SES subgroups (see Table 2). Thus, with regard to our third aim, we found that the latent structure of spatial skill and mathematics did not differ for any of the SES- and sex-related subgroups at any of the three grade levels.

8.3.2. Structured means models

Having established configural invariance in the two-factor multiple group models, we used structured means modeling (Thompson & Green, 2013) to revisit our first aim regarding potential group differences in mean spatial and mathematics performance within grade levels. The ANCOVA models reported above are helpful for identifying patterns of effects in the data, but they require strong assumptions about a lack of measurement error and about the equivalent measurement quality of the items used to form the mathematics and spatial composites (McNeish & Wolf, 2019). In contrast, structured means modeling allows us to estimate subgroups' measurement parameters with testable constraints, potentially reducing bias in estimates of the subgroups' means and increasing power to detect mean differences (Thompson & Green, 2013). In this way, we can ask whether subgroup differences that yielded only small or non-significant effects using ANCOVA might emerge using these less restrictive tests (Jung & Yoon, 2016; Thompson & Green, 2013).

Taking this more sensitive approach, we replicated the overall pattern of results indicated by the ANCOVA models (Table 3). That is, we found (1) sex differences in spatial skill were apparent within each grade level but only apparent for kindergarteners in mathematics skill, and (2) income differences were apparent in both spatial and mathematics skills within all grade levels. However, the effect sizes estimated in the ANCOVA models were attenuated compared to the effects

² In the ANCOVA analyses reported above, the ordered income scale (1 ≤ \$15,000, 2 = \$15,000 to 34,999; 3 = \$35,000 to 49,999; 4 = \$50,000 to 74,999; 5 = \$75,000 to 99,999; and 6 = \$100,000 or more) was used to assess the effect of income on mathematics and spatial skill. In the multiple group factor analysis models, a dichotomized version of income was used, such that individuals reporting \$50,000 or greater annual income were considered "high income" while individuals reporting under \$50,000 annual income were considered "low income." The split was based on definitions of low income, not a median split. The National Center for Children in Poverty defines low income as less than twice the poverty line. According to the NCCP, the federal poverty line in 2016 was \$24,300 (in 2020 it is \$26,200) for a family of four with two children. By that definition, low income would be \$48,600 (or \$52,400 in 2020). http://www.nccp.org/profiles/US_profile_6.html

estimated in the structured means models. For example, the results of the ANCOVA indicated that the overall effect of sex on spatial skill was small ($\eta^2_p = 0.015$, $p < .01$), while the structured means model indicated that the effect approached a medium effect size within each grade level ($\delta_{Kdg,sex} = 0.46$, $p < .01$; $\delta_{3rd,sex} = 0.48$, $p < .01$; $\delta_{6th,sex} = 0.46$, $p < .01$). One possible explanation for the discrepant effect sizes is that the ANCOVA models included both sex and income as covariates, whereas the structured means models included one covariate at a time; additionally, the income variable was dichotomized for the purposes of the structured means model but was treated as ordinal in the ANCOVA model. Regardless, given that both the ANCOVA models and the structured means models failed to detect evidence of a gap in mathematics skill between boys and girls in third and sixth grades strengthens the conclusion that sex difference in mathematics skills were not present at these grade levels.

In addition to comparing group differences in the factor means, we assessed the reliability of the structured means models using McDonald's coefficient omega (McDonald, 1978; Raykov & Shrout, 2002) in order to determine how well our two-factor models measured mathematics and spatial abilities. The multidimensional (i.e., two-factor) and unidimensional (i.e., single-factor) reliabilities ranged between 0.66 and 0.90, with the lowest reliabilities in measuring the spatial factor among three kindergarten subgroups ($\omega_{Spatial,boys} = 0.66$; $\omega_{Spatial,girls} = 0.69$; $\omega_{Spatial,low-inc} = 0.68$) and two sixth grade subgroups ($\omega_{Spatial,girls} = 0.68$; $\omega_{Spatial,high-inc} = 0.67$), and the highest reliabilities in measuring the mathematics factor among all subgroups in sixth grade ($\omega_{Math,boys} = 0.90$; $\omega_{Math,girls} = 0.87$; $\omega_{Math,low-inc} = 0.90$; $\omega_{Math,high-inc} = 0.88$). These reliabilities well exceed the minimum recommendations, establishing our structured means models as reliable measures of latent mathematics and spatial skills across grade levels and subgroups (Gagne & Hancock, 2006). Using this model, we also found evidence that the mathematics and spatial factors correlated strongly with one another within each grade level and subgroup (see Fig. 5 for a visualization using factor scores). Estimates of the interfactor correlations between mathematics and spatial skills ranged between 0.52 and 0.74 across subgroups providing strong evidence for a processing link between mathematics and spatial skills across developmental stages, SES, and sex. The full tables of results are available in the Appendix A.

8.3.3. Subgroup differences

We next used the two-group structured means models described above to compare standardized mean differences for each of the subgroups separately. Overall, the performance gaps in spatial skill shown in the ANCOVA models were replicated in the structured means models, with boys generally outperforming girls in spatial skills and high-income students consistently outperforming low-income students within each

Table 2
Configural model results for sex and income across grade levels.

Model	χ^2	df	p-value	RMSEA	(LB _{90%CI}	UB _{90%CI})	SRMR
Kindergarten							
Sex							
Full Model (n = 526)	152.71	86	<0.01	0.05	(0.04,	0.07)	0.04
Boys Only (n = 263)	84.48	43	<0.01	0.06	(0.04,	0.08)	0.04
Girls Only (n = 263)	67.95	43	0.01	0.05	(0.02,	0.07)	0.05
SES							
Full Model (n = 378)	135.06	86	<0.01	0.06	(0.04,	0.07)	0.06
High Income Only (n = 138)	64.99	43	<0.02	0.06	(0.03,	0.09)	0.06
Low Income Only (n = 240)	69.84	43	0.01	0.05	(0.03,	0.07)	0.05
Third Grade							
Sex							
Full Model (n = 537)	214.52	106	<0.01	0.06	(0.05,	0.07)	0.05
Boys Only (n = 238)	83.90	53	<0.01	0.05	(0.03,	0.07)	0.04
Girls Only (n = 299)	130.03	53	<0.01	0.07	(0.06,	0.09)	0.05
SES							
Full Model (n = 349)	199.89	106	<0.01	0.07	(0.06,	0.09)	0.05
High Income Only (n = 117)	86.45	53	<0.01	0.07	(0.04,	0.10)	0.06
Low Income Only (n = 232)	113.94	53	<0.01	0.07	(0.05,	0.09)	0.05
Sixth Grade							
Sex							
Full Model (n = 529)	297.79	178	<0.01	0.05	(0.04,	0.06)	0.05
Boys Only (n = 252)	156.48	89	<0.01	0.06	(0.04,	0.07)	0.05
Girls Only (n = 277)	141.37	89	<0.01	0.05	(0.03,	0.06)	0.04
SES							
Full Model (n = 392)	239.17	178	<0.01	0.04	(0.03,	0.06)	0.05
High Income Only (n = 152)	110.58	89	0.06	0.04	(0.00,	0.06)	0.05
Low Income Only (n = 240)	128.57	89	<0.01	0.04	(0.03,	0.06)	0.05

Note. χ^2 = Model chi-square. df = Model degrees of freedom. RMSEA = Root mean square error of approximation. LB_{90%CI} and UB_{90%CI} = Lower and upper bounds, respectively, of the 90% confidence interval for the RMSEA. SRMR = Standardized root mean square residual.

Table 3
Standardized subgroup mean differences.

Parameter	Kindergarten		Third Grade		Sixth Grade	
	Est.	(SE)	Est.	(SE)	Est.	(SE)
Mathematics Skill						
Sex (Boys – Girls)	0.35 *	(0.10)	0.16 ‡	(0.10)	0.10 ‡	(0.10)
Income (High – Low)	0.30 *	(0.11)	0.40 *	(0.13)	0.59 *	(0.12)
Spatial Skill						
Sex (Boys – Girls)	0.46 *	(0.11)	0.48*‡	(0.11)	0.46*‡	(0.10)
Income (High – Low)	0.40 *	(0.12)	0.55 *	(0.13)	0.39 *	(0.12)

Note. Est. = Point estimate. SE = Standard error. Estimates marked with an asterisk (*) are statistically significant based on an alpha level of 0.05. Within grade level, standardized effect estimates marked with a double trident (‡) differ significantly from one another. Standardized effects were compared across factors within grade level and subgroup by constructing bias-corrected bootstrapped confidence intervals for the difference.

grade level (Table 3). In general, using the structured means models, which account for measurement error and measurement non-invariance, revealed stronger effect sizes than the ANCOVA models.

First, unlike the ANCOVAs that showed mostly nonsignificant and small-to-medium performance gaps, the structured means comparisons showed larger performance gaps in the moderate-to-medium range. For spatial skill, this was the case in both subgroups, sex and income, and these gaps persisted at a similar effect size across all age groups (sex: δ_{Kdg} ; boys vs. girls = 0.46, $p < .01$; δ_{3rd} ; boys vs. girls = 0.48, $p < .01$; δ_{6th} ; boys vs. girls = 0.46, $p < .01$; income: δ_{Kdg} ; high vs. low = 0.40, $p < .01$; δ_{3rd} ; high vs. low = 0.55, $p < .01$; δ_{6th} ; high vs. low = 0.39, $p < .01$). For mathematics skill, we also found moderate-to-medium performance gaps between high-income and low-income students at all three ages. Finally, recall that although the ANCOVA had shown no significant grade x sex interaction for mathematics, pairwise comparisons suggested that boys might have outperformed girls in kindergarten only. This effect was replicated in the analyses using structured means comparisons, and, in fact, there was evidence of a moderate performance gap in mathematics performance between boys and girls in kindergarten (δ_{Kdg} , boys vs. girls =

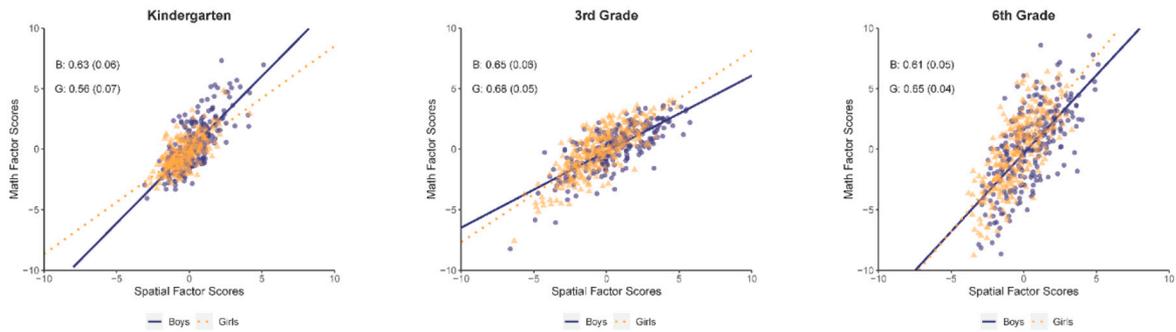
0.35, $p < .01$); however, no difference was detected in third and sixth grades (δ_{3rd} , boys vs. girls = 0.16, *n.s.*; δ_{6th} , boys vs. girls = 0.10, *n.s.*). These patterns are evident in the factor score distributions presented in Fig. 6, which depict moderate and persistent gaps between subgroup distributions of spatial performance (Panel B), while the sex subgroup distributions of mathematics performance nearly perfectly overlap by sixth grade (Panel A). Note that the present model analyzed sex and income separately due to considerations of power and model complexity, thus preventing the testing of interactions between groups. To determine whether the performance gaps between girls and boys and high-income and low-income students are additive, that is, whether low-income girls are the lowest performing group and high-income boys are the highest, further work is recommended with larger sample sizes.

Second, whereas the ANCOVA model indicated that the income by grade interaction on mathematics skill yielded a small but non-significant effect ($\eta^2_p = 0.011$, $p = .274$), the results of the structured means models suggest a mathematics performance gap between high-income and low-income students that is larger in later grades. Astonishingly, the effect size of the income level on mathematics in sixth grade was nearly twice that in kindergarten (δ_{Kdg} ; high vs. low = 0.30, $p = .01$; δ_{3rd} ; high vs. low = 0.40, $p < .01$; δ_{6th} ; high vs. low = 0.59, $p < .01$). Of course, these results are based on cross-sectional rather than longitudinal data, so must not be interpreted as increasing over time in the same children. Still, the discrepancy between older and younger students is considerable, adding to the literature focused on the importance of early intervention (e.g., Claessens, Duncan, & Engel, 2008; Raudenbush et al., 2020). In contrast, as shown in Table 3, we did not find a widening income-related gap across the grades for spatial skill.

This discrepancy may seem surprising given that spatial skill and mathematics skill are significantly related in sixth grade, and indeed, the magnitude of this relation is stable across the three grades we tested (kindergarten, third, and sixth) (Mix et al., 2016). If the two skills are highly correlated, then why does the income-gap increase in one domain and not the other?

One reason may be that spatial skill plays a larger role in mathematics learning when children are acquiring novel concepts or

(a)



(b)

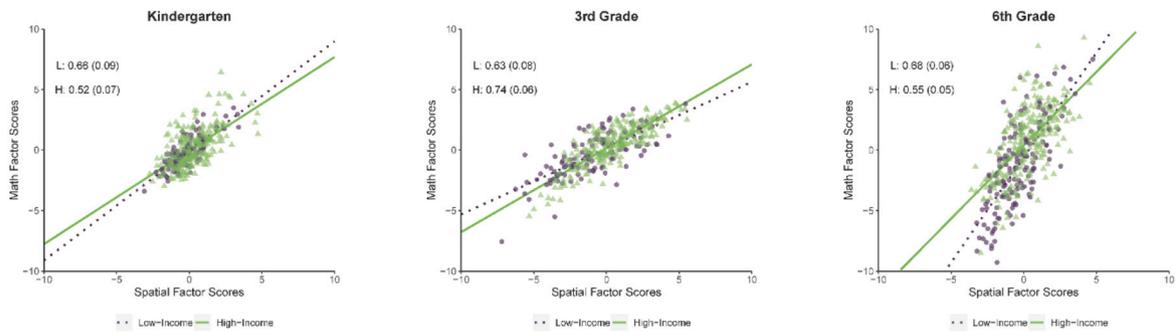
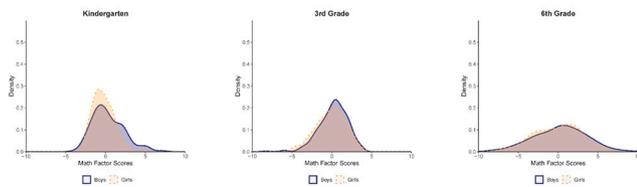


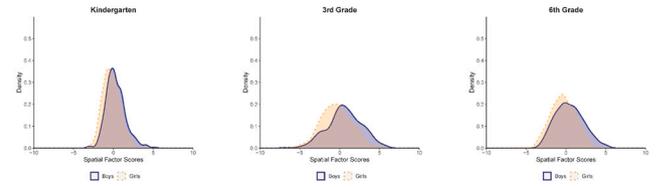
Fig. 5. Visualization of Interfactor Relations Across Grade Levels by Sex (a) and Income (b).

Note. Scatterplots are based on factor scores generated from the final partial invariance models for each grade level and grouping variable. Panel A shows the strength of interfactor (math and spatial) relations across grade levels and within sex groupings. Panel B shows the same correlations across grade and income. Figures are annotated with the estimated interfactor correlation and standard error (in parentheses) by subgroup, though there was no evidence of significant differences between inter-factor correlations. Full parameter estimates are available in Appendix A.

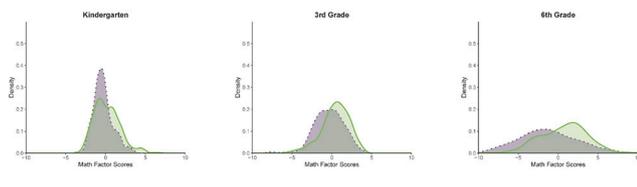
(a)



(c)



(b)



(d)

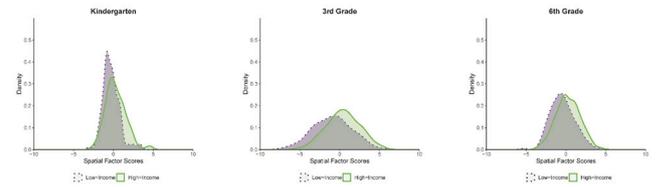


Fig. 6. Visualization of Between-Group Differences in Mathematics Skill by Sex (a) and Income (b) as well as Spatial Skill by Sex (c) and Income (d).

Note. Distributions are based on factor scores generated from the final partial invariance models for each grade level and grouping variable. Visualizations of between-group differences in each factor are broken down by sex and income groupings across grade levels.

grounding unfamiliar symbols (Mix et al., 2016; Mix et al., 2017). Perhaps these early deficits in grounding basic mathematics concepts snowball into larger differences in mathematics achievement later on, even if variation in spatial skill remains stable over time. Alternatively, the widening mathematics achievement gap may be due to other sources of variation that are linked to SES, such as verbal skills that are important for solving word problems, the quality of schooling, and/or neighborhood resources (Boonen, de Koning, Jolles, & van der Schoot, 2016; Entwistle, Alexander & Olson, 2018; McLoyd, 1998; Sirin, 2005). These explanations are not mutually exclusive.

9. Discussion

The present study offers a secondary analysis of data measuring spatial and mathematics skill in kindergarten, third grade, and sixth grade students (Mix et al., 2016; Mix et al., 2017). Our specific aims were to determine whether (1) spatial performance, mathematics performance, or both varied as a function of sex or income, (2) mathematics performance differences were mediated by spatial skill, and (3) the latent structures observed previously for children grouped by grade also held true for subgroups based on income and sex within each grade.

With regard to the first aim, we obtained strong evidence for better performance among children from higher income backgrounds on both spatial and mathematics tasks, with medium to large effect sizes for both (spatial $\eta^2_p = 0.092$; mathematics $\eta^2_p = 0.098$). This income advantage was evident in both the analyses of variance examining differences in performance levels, and in the significance tests used to evaluate differences in the factor means for these groups. These findings are robust and consistent with previous research reporting similar SES advantages for spatial and mathematics skill (Claessens et al., 2008; Jirout & Newcombe, 2015; Levine et al., 2005; Sirin, 2005).

A novel contribution of the present study is that structured means comparisons yielded evidence of increasing income subgroup differences in mathematics performance across grade levels, suggesting that the mathematics performance gap between high-income and low-income students increases with development, consistent with prior longitudinal work (Denton & West, 2002). However, this pattern differed from what we found with spatial skill income-related gaps, where the disparities among income groups did not change across development. It is interesting that mathematics and spatial skill diverge in this way, given that they remain equally correlated across the same age range—a pattern that may seem, at first, to be paradoxical.

One possible explanation could be that there is a minimum threshold for the spatial skill needed to support superior performance in mathematics, beyond which additional spatial skill does not contribute to the variance in mathematics achievement (Freer, 2017). If the spatial skill of children from low-income families falls below that threshold, they may experience worse mathematics outcomes early in development that could snowball into larger deficits under the pressure of acquiring more advanced mathematics without a strong foundation of basic skills. In contrast, children who meet the threshold for minimal spatial skill may perform better in mathematics throughout development, and never show major deficits in prerequisite skill development. In this way, it may not be necessary to have increasingly greater spatial skill in order to show increasingly greater mathematics outcomes. Alternatively, this divergence could reflect underlying educational disparities between income groups that may be overlooked in the ANCOVAs we carried out. Quality of schooling would not be expected to have as strong an effect on spatial skill development partly because it is not commonly taught (Levine et al., 2016). It is also possible that mathematics outcomes are more sensitive to differences in parenting and family resources than spatial outcomes.

As in prior research (Lauer et al., 2019; Levine et al., 2005), we obtained evidence for a significant sex difference favoring boys over girls on spatial tasks, both collapsing across the grades ($\eta^2_p = 0.018$) and when examined within grade. Also consistent with prior research

(Bakker et al., 2019; Lachance & Mazzocco, 2006), we did not observe a significant sex difference on mathematics tasks when collapsing across grades ($\eta^2_p = 0.005$). However, follow-up tests within grades revealed a male advantage for mathematics in kindergarten only. This outcome was unexpected given that when sex differences in mathematics have been reported, they tend to be restricted to older children (Lindberg et al., 2010; Reilly et al., 2015), though not always (see Penner & Paret, 2008).

A possible explanation for this early sex difference in mathematics might be that the mathematics measures we chose were unfamiliar or challenging to the specific kindergarten children we tested. For example, children in our sample may not have been exposed to missing term problems or place value questions. In support of this, an inspection of the kindergarten factor loadings indicates that the correlations between spatial skill and performance on these specific mathematics subskills were stronger for boys than for girls (See Table A1). Recall that previous research has shown a male advantage on complex or novel problem-solving tasks where invented strategies are more important (Lindberg et al., 2010; Reilly et al., 2015). In contrast, a female advantage for mathematics has been demonstrated in course grades (Duckworth & Seligman, 2006; Kenney-Benson et al., 2006; Pomerantz et al., 2002) and for routinized skills, such as counting (Hutchison et al., 2019; Hyde et al., 1990). Although we did not intentionally manipulate novelty in our measures, it is possible that for the children we tested, the balance of novel problems was greater in kindergarten than in the other grades, perhaps creating a situation where boys could outperform girls on average. This explanation is consistent with evidence showing that, among the highest performing mathematics students in kindergarten, boys outperform girls (Penner & Paret, 2008), perhaps because the mathematics content that differentiates these groups is novel and challenging.

Some may find it surprising that our results did not reveal sex differences among older children, given that previous research has shown that when there are significant sex differences in mathematics, these are more apparent in older students (e.g., Ceci et al., 2009). However, such research has focused on adolescents and adults, who are much older than even the oldest children in our sample. If we had tested older children, we might have expected to replicate those sex differences in mathematics, but the fact that we did not find such differences in third and sixth grade children seems consistent with the literature on elementary-aged children (e.g., Bakker et al., 2019; Kersey et al., 2019).

Our second aim was to test whether spatial skill mediated performance differences in mathematics for specific subgroups. Few studies have examined these relations (e.g., Casey et al., 2011) and ours is the first to demonstrate that spatial skill mediates the relation between income and mathematics (albeit partially) across three elementary grades using comparable measures. These relations between income and mathematics also were partially mediated by verbal skill, suggesting there are multiple contributors. As noted above, one contributor may be disparities in school quality, which could explain why the relations between income and mathematics are not fully mediated by spatial skill, verbal skill, or a combination of the two. Other possibilities include access to informal mathematics activities or homework support. For the portion of variance mediated by spatial skill, a potential explanation is that spatial visualization may be particularly helpful when children are attacking novel problems or choosing among potential solutions. As Mix (2019) argued, one major function of spatial processing for mathematics may be grounding mathematics symbols and operations (i.e., connecting them to a physical or imagined referent). It has also been hypothesized that spatial visualization, being relatively effortful, is recruited more often for novel or challenging content (Lowrie & Kay, 2001; Mix, 2019; Mix et al., 2016; Uttal & Cohen, 2012). If so, then we might expect income-related differences in spatial skill to mediate only part of the income-related differences in mathematics skill—that is, the part measured by novel or challenging problems. The portion of variance mediated by verbal skill could be rooted in differences in general

cognitive ability, differences in verbal strategy use, or both.

Alternatively, this mediation effect could be viewed as reflecting variation in general cognitive skill that has both a verbal and spatial component. Studies have demonstrated that fluid reasoning (e.g., matrix reasoning) and spatial skill are overlapping (Ackerman, Beier, & Boyle, 2002; Colom, Contreras, Botella, & Santacreu, 2002; Lohman, 2000), and recent work demonstrates that when fluid reasoning, verbal skill, and spatial skill are included in the same model, only fluid reasoning predicts children's subsequent mathematics outcomes (Green, Bunge, Chiongbian, Barrow, & Ferrer, 2017). Thus, it is possible our spatial measures tapped fluid reasoning and the observed effects may reflect variation in this domain general skill to some extent. Unfortunately, we were unable to test this possibility with our existing dataset because we did not include a measure of fluid intelligence, but this could be an important control in future work.

We also found that spatial skill fully mediated the male advantage we observed for kindergarteners in mathematics. This is a striking finding given that our income effects on mathematics were only partially mediated by spatial skill and seemed to have multiple origins. The finding suggests that in the case of this sex difference, the entire effect can be attributed to differences in spatial skill. In contrast to the effects involving income, this sex difference was small and limited to the youngest age group tested, so it appears to be a much narrower and potentially specific finding. As we argued in our interpretation of the sex difference itself, it seems possible that the mathematics problems we chose were novel or challenging for the kindergarten children in our sample. If that is the case, then boys may have fared better because they tend to do well when attacking novel problems and inventing solutions, whereas girls tend to do well when they are encountering more familiar mathematics content. However, this mediation finding goes further to suggest that the cognitive processing boys use to attack novel problems may be spatial, consistent with what others have hypothesized (Lowrie & Kay, 2001; Mix, 2019; Mix et al., 2016; Uttal & Cohen, 2012). These interpretations are speculative because we do not have a direct measure of novelty or the strategies used to solve problems in this dataset. Also, just as for the mediation of relations involving income, the mediation effect for sex may reflect differences in fluid reasoning to some extent. However, the proposed mechanism of spatial skill being recruited to attack novel or challenging problems seems plausible and bears testing in future research.

Although we did not test causality or training effects, our findings are consistent with the possibility that spatial training can be useful in closing performance gaps such as these. If spatial processing mediates the relations between income and mathematics outcomes, and between sex and mathematics outcomes, it stands to reason that improving spatial skill could lead to better mathematics outcomes, particularly in the more vulnerable subgroups (i.e., low-income children, young girls with respect to novel problems) (Levine et al., 2016; Mix & Cheng, 2012; Newcombe, 2017). Recall that several studies have reported positive effects of spatial training on mathematics outcomes in the elementary school age ranges (Cheng & Mix, 2014; Cheung et al., 2019; Lowrie et al., 2017). The present results suggest that these effects may be greatest among children who are not already high performing in both domains but may have little effect on those who are (i.e., high income boys). Perhaps future spatial training studies should focus on low-income students, where the potential for growth in both spatial skills and mathematics achievement may be greater. These interventions may be particularly effective early in development, before the expansion of the mathematics achievement gap observed in this study and reported previously (Denton & West, 2002; Raudenbush et al., 2020).

Our final aim was to determine whether the latent structure of spatial and mathematics skills differed depending on SES, sex, or both. Recall that in previous factor analyses of the same dataset analyzed here (CITATIONS BLINDED FOR REVIEW), spatial and mathematics performance formed unitary factors that were separate but significantly correlated. In the present analysis, we used multi-group confirmatory

factor analysis to see whether the same structure was evident in income and sex subgroups. The results confirmed that it is. Although there were differences in performance levels for these subgroups, the basic two-factor, separate but correlated structure of the relations was the same.

This finding is interesting because it was possible that for certain subgroups, spatial and mathematics performance may have related differently. For example, among high income students who tended to outperform the other subgroups on spatial tasks and mathematics tasks, it was possible spatial and mathematics performance might form a single factor, or that there were multiple factors with different combinations of spatial and mathematics measures loadings. That these other patterns were not obtained, and the two-factor, separate but correlated structure was evident across all subgroups instead, suggests that this structure is stable and deeply rooted in core processing differences between the two domains, despite their clear overlap. Defining this overlap and distinguishing it from the aspects of processing that are non-overlapping remains will help to elucidate how, specifically, spatial processes are related to mathematical thought. Getting to the bottom of this will require understanding precisely how the two domains do and do not overlap and how this partial overlap is manifested in various facets of mathematical performance.

In conclusion, the present study sought to examine subgroup differences in mathematics and spatial performance, and investigate whether the factor structure identified by Mix et al., 2016; Mix et al., 2017 replicated in income and sex subgroups. Our findings add to the literature that documents the intertwined nature of spatial and mathematical development by examining these relations in specific subgroups, and by demonstrating advantages based on income and sex, as well as evidence that spatial skill mediates these group differences. Our findings indicate that despite these group differences, the latent structure of spatial skill and mathematics skill is consistent across age, income level, and sex. Taken together, these findings suggest that interventions aimed at increasing spatial skill may have positive effects on student learning in mathematics, particularly among early elementary aged children and those from lower SES backgrounds.

Credit author statement

Johnson: Conceptualization; Methodology; Formal Analysis; Writing-Original Draft; Visualization.

Burgoyne: Conceptualization; Formal Analysis; Writing-Original Draft; Visualization.

Mix: Conceptualization; Writing-Original Draft; Data Curation; Supervision; Project Administration; Funding Acquisition.

Levine: Conceptualization; Writing-Review and Editing; Funding Acquisition.

Young: Conceptualization; Writing-Review and Editing.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2021.104918>.

References

- Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, 131(4), 567–589.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411–423.
- Bakker, M., Torbeyns, J., Wijns, N., Verschaffel, L., & De Smedt, B. (2019). Gender equality in 4-to 5-year-old preschoolers' early numerical competencies. *Developmental Science*, 22(1). e12718. 10.1111/desc.12718.
- 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 Inc.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, 289–300. Doi: 10.1111/j.2517-6161.1995.tb02031.x.
- Boonen, A. J., de Koning, B. B., Jolles, J., & van der Schoot, M. (2016). Word problem solving in contemporary mathematics education: A plea for reading comprehension

- skills training. *Frontiers in Psychology*, 7, 191. <https://doi.org/10.3389/fpsyg.2016.00191>
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issues of partial measurement invariance. *Psychological Bulletin*, 105(3), 456–466. <https://doi.org/10.1037/0033-2909.105.3.456>
- Casey, B. M., Dearing, E., Vasilyeva, M., Ganley, C. M., & Tine, M. (2011). Spatial and numerical predictors of measurement performance: The moderating effects of community income and gender. *Journal of Educational Psychology*, 103(2), 296–311. <https://doi.org/10.1037/a0022516>
- Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology*, 33(4), 669–680. <https://doi.org/10.1037/0012-1649.33.4.669>
- Casey, M. B., Nuttall, R. L., & Pezaris, E. (2001). Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on mathematics subtests using cross-national gender-based items. *Journal for Research in Mathematics Education*, 32(1), 28–57. <https://doi.org/10.2307/749620>
- Casey, M. B., Nuttall, R. L., Pezaris, E., & Benbow, C. (1995). Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on mathematics subtests using cross-national gender-based items. *Journal for Research in Mathematics Education*, 32(1), 28–57. <https://doi.org/10.2307/749620>
- Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, 135(2), 218–261.
- Cheng, Y.-L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, 15(1), 2–11. <https://doi.org/10.1080/15248372.2012.725186>
- Cheung, C.-N., Sung, J. Y., & Lourenco, S. F. (2019). Does training mental rotation transfer to gains in mathematical competence? Assessment of an at-home visuospatial intervention. *Psychological Research*, 1–18. <https://doi.org/10.1007/s00426-019-01202-5>
- Claessens, A., Duncan, G., & Engel, M. (2008). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, 28(4), 415–427. <https://doi.org/10.1016/j.econedurev.2008.09.003>
- Cobb-Clark, D. A., & Moschion, J. (2017). Gender gaps in early educational achievement. *Journal of Population Economics*, 30(4), 1093–1134. <https://doi.org/10.1007/s00148-017-0638-z>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Colom, R., Contreras, M. J., Botella, J., & Santacreu, J. (2002). Vehicles of spatial ability. *Personality and Individual Differences*, 32(5), 903–912.
- Crane, J. (1996). Effects of home environment, SES, and maternal test scores on mathematics achievement. *The Journal of Educational Research*, 89(5), 305–314. <https://doi.org/10.1080/00220671.1996.9941332>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Dearing, E., Casey, B. M., Ganley, C. M., Tillinger, M., Laski, E., & Montecillo, C. (2012). Young girls' arithmetic and spatial skills: The distal and proximal roles of family socioeconomic and home learning experiences. *Early Childhood Research Quarterly*, 27(3), 458–470. <https://doi.org/10.1016/j.ecresq.2012.01.002>
- Demir, Ö. E., Prado, J., & Booth, J. R. (2015). Parental socioeconomic status and the neural basis of arithmetic: Differential relations to verbal and visuo-spatial representations. *Developmental Science*, 18(5), 799–814. <https://doi.org/10.1111/desc.12268>
- Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade*. National Center for education statistics. US Department of Education: Office of Educational Research and Improvement.
- Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, 98(1), 198–208.
- Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the problem of internal consistency estimation. *British Journal of Psychology*, 105, 399–412. <https://doi.org/10.1111/bjop.12046>
- Elliott, L., & Bachman, H. J. (2018). SES disparities in early mathematics abilities: The contributions of parents' mathematics cognitions, practices to support mathematics, and mathematics talk. *Developmental Review*, 49, 1–15. <https://doi.org/10.1016/j.dr.2018.08.001>
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2018). *Children, schools, and inequality*. Routledge.
- Farah, M. J., Shera, D. M., Savage, J. H., Betancourt, L., Giannetta, J. M., Brodsky, N. L., & Hurt, H. (2018). Childhood poverty: Specific associations with neurocognitive development. *Brain Research*, 1110(1), 166–174. Farah, M. J., Shera, D. M., Savage, J. H., Betancourt, L., Giannetta, J. M., Brodsky, N. L., ... & Hurt, H. (2006). Childhood poverty: Specific associations with neurocognitive development. *Brain Research*, 1110(1), 166–174. <https://doi.org/10.2307/748393>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.2307/748393>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4), 1149–1160.
- Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher*, 27(5), 6–11. <https://doi.org/10.2307/1176733>
- Fennema, E., & Tartre, L. A. (1985). The use of spatial visualization in mathematics by girls and boys. *Journal for Research in Mathematics Education*, 16(3), 184–206. <https://doi.org/10.2307/748393>
- Freder, D. (2017). *The threshold hypothesis applied to spatial skill and mathematics*. Doctoral dissertation submitted to Michigan State University.
- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, 13(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3
- Frick, A. (2018). Spatial transformation abilities and their relation to later mathematics performance. *Psychological Research*, 1–20. https://doi.org/10.1207/s15328007sem1303_3
- Frick, A., Möhring, W., & Newcombe, N. S. (2014). Picturing perspectives: Development of perspective-taking abilities in 4-to 8-year-olds. *Frontiers in Psychology*, 5, 386. <https://doi.org/10.3389/fpsyg.2014.00386>
- Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41(1), 65–83.
- Geer, E. A., Quinn, J. M., & Ganley, C. M. (2019). Relations between spatial skills and mathematics performance in elementary school children: A longitudinal investigation. *Developmental Psychology*, 55(3), 637–652. <https://doi.org/10.1037/dev0000649>
- Geiser, C., Lehmann, W., Corth, M., & Eid, M. (2008). Quantitative and qualitative change in children's mental rotation performance. *Learning and Individual Differences*, 18(4), 419–429. <https://doi.org/10.1016/j.lindif.2007.09.001>
- Gilligan, K. A., Flouri, E., & Farran, E. K. (2017). The contribution of spatial ability to mathematics achievement in middle childhood. *Journal of Experimental Child Psychology*, 163, 107–125. <https://doi.org/10.1016/j.jecp.2017.04.016>
- Ginsburg, H., & Baroody, A. J. (2003). *TEMA-3: Test of early mathematics ability*. Austin, TX: Pro-Ed.
- Green, C. T., Bunge, S. A., Chiongbian, V. B., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology*, 157, 125–143. <https://doi.org/10.1016/j.jecp.2016.12.005>
- Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Science*, 13(2), 65–73. <https://doi.org/10.1016/j.tics.2008.11.003>
- Hawes, Z., Moss, J., Caswell, B., Seo, J., & Ansari, D. (2019). Relations between numerical, spatial, and executive function skills and mathematics achievement: A latent-variable approach. *Cognitive Psychology*, 109, 68–90. <https://doi.org/10.1016/j.cogpsych.2018.12.002>
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York: The Guilford Press.
- Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 61, 683–689. <https://doi.org/10.1080/17470210701822967>
- Herts, J. B., & Levine, S. C. (2020). *Gender and mathematics development*. Oxford Research Encyclopedia of Education.
- Hresko, W., Schlieve, P. L., Herron, S. R., Sawain, C., & Sherbenou, R. (2003). *Comprehensive mathematics abilities test*. Austin, TX: Pro-Ed.
- Hsu, J. C. (1999). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall/CRC.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–55. <https://doi.org/10.1080/10705519909540118>
- Hutchison, J. E., Lyons, I. M., & Ansari, D. (2019). More similar than different: Gender differences in children's basic numerical skills are the exception not the rule. *Child Development*, 90(1), e66–e79. <https://doi.org/10.1111/cdev.13044>
- Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, 107, 139–155. <https://doi.org/10.1037/0033-2909.107.2.139>
- Jirout, J. J., & Newcombe, N. S. (2015). Building blocks for developing spatial skills: Evidence from a large, representative US sample. *Psychological Science*, 26(3), 302–310. <https://doi.org/10.1177/0956797614563338>
- Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4), 567–584.
- Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman assessment battery for children* (Wiley Online Library).
- Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in mathematics performance: The role of children's approach to schoolwork. *Developmental Psychology*, 42(1), 11–26. <https://doi.org/10.1037/0012-1649.42.1.11>
- Kersey, A. J., Csumitta, K. D., & Cantlon, J. F. (2019). Gender similarities in the brain during mathematics development. *npj Sci. Learn.*, 4(19). <https://doi.org/10.1038/s41539-019-0057>
- Klein, P. S., Adi-Japha, E., & Hakak-Benizri, S. (2010). Mathematical thinking of kindergarten boys and girls: Similar achievement, different contributing processes. *Educational Studies in Mathematics*, 73(3), 233–246. <https://doi.org/10.1007/s10649-009-9216-y>
- Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29(5), 745–756. <https://doi.org/10.3758/BF03200477>
- Lachance, J. A., & Mazzocco, M. M. (2006). A longitudinal analysis of sex differences in mathematics and spatial skills in primary school age children. *Learning and Individual Differences*, 16(3), 195–216. <https://doi.org/10.1016/j.lindif.2005.12.001>

- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2), 202–220. <https://doi.org/10.1177/1094428105284919>
- 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. <https://doi.org/10.1177/09567976166655977>
- Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin*, 145(6), 537. <https://doi.org/10.1037/bul0000191>
- Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Sex differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2), 127–155. <https://doi.org/10.1002/wcs.1380>
- Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., & Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychological Science*, 16(11), 841–845. <https://doi.org/10.1037/bul0000191>
- Liben, L. S., & Downs, R. M. (1989). Understanding maps as symbols: The development of map concepts in children. *Advances in Child Development and Behavior*, 22, 145–201. [https://doi.org/10.1016/S0065-2407\(08\)60414-0](https://doi.org/10.1016/S0065-2407(08)60414-0)
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123–1135. <https://doi.org/10.1037/a0021276>
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56(6), 1479–1498. <https://doi.org/10.2307/1130467>
- Lippa, R. A., Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior*, 39(4), 990–997. <https://doi.org/10.1007/s10508-008-9460-8>
- Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 285–340). Cambridge, MA: Cambridge University Press.
- Lowrie, T., & Kay, R. (2001). Relationship between visual and nonvisual solution methods and difficulty in elementary mathematics. *The Journal of Educational Research*, 94(4), 248–255.
- Lowrie, T., Logan, T., & Ramful, A. (2017). Visuospatial training improves elementary students' mathematics performance. *British Journal of Educational Psychology*, 87(2), 170–186. doi: 10.1111/bjep.12142.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, 4(1), 84–99. <https://doi.org/10.1037/1082-989X.4.1.84>
- McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement*, 38(1), 75–79. <https://doi.org/10.1177/001316447803800111>
- McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, 129–150. <https://doi.org/10.2307/30034845>
- McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, 53(2), 185–204. <https://doi.org/10.1037/0003-066X.53.2.185>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- McNeish, D., & Wolf, M. G. (2019). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.
- Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, 18(1), 37–45. <https://doi.org/10.1016/j.tics.2013.10.011>
- Mix, K. S. (2019). Why are spatial skill and mathematics related? *Child Development Perspectives*, 13(2), 121–126. <https://doi.org/10.1111/cdep.12323>
- Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and mathematics: Developmental and educational implications. In J. B. Benson (Ed.), *Vol. 42. Advances in child development and behavior* (pp. 197–243). Waltham, MA: Academic Press.
- 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. <https://doi.org/10.1037/xge0000182>
- Mix, K. S., Levine, S. C., Cheng, Y.-L., Stockton, J. D., & Bower, C. (2020). Does spatial training improve mathematics performance? A comparison of training type, age, and mathematics outcome. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000494>. Advance online publication.
- Mix, K. S., Levine, S. C., Cheng, Y.-L., Young, C. J., Hambrick, D. Z., & Konstantopoulos, S. (2017). The latent structure of spatial skills and mathematics: Further evidence from wave 2. *Journal of Cognition and Development*, 18(4), 465–492. <https://doi.org/10.1080/15248372.2017.1346658>
- Moè, A. (2018). Mental rotation and mathematics: Gender-stereotyped beliefs and relationships in primary school children. *Learning and Individual Differences*, 61, 172–180. <https://doi.org/10.1016/j.lindif.2017.12.002>
- Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, 19(11), 1063–1066. <https://doi.org/10.1111/j.1467-9280.2008.02200.x>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- 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(8), 1238–1242. <https://doi.org/10.1016/j.paid.2011.02.017>
- Newcombe, N. (2017). *Harnessing spatial thinking to support stem learning* (OECD Library).
- Noble, K. G., McCandliss, B. D., & Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science*, 10(4), 464–480. <https://doi.org/10.1111/j.1467-7687.2007.00600.x>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Parker, M. C., Solomon, A., Pritchett, B., Illingworth, D. A., Marguileux, L. E., & Guzdial, M. (2018, August). Socioeconomic status and computer science achievement: Spatial ability as a mediating variable in a novel model of understanding. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (pp. 97–105). ACM.
- Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, 37(1), 239–253.
- Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, 112(4), 1137–1162.
- Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition*, 28(1), 39–58. <https://doi.org/10.1006/brcg.1995.1032>
- Pezaris, E., & Casey, M. B. (1991). Girls who use "masculine" problem-solving strategies on a spatial task: Proposed genetic and environmental factors. *Brain and Cognition*, 17(1), 1–22. [https://doi.org/10.1016/0278-2626\(91\)90062-D](https://doi.org/10.1016/0278-2626(91)90062-D)
- Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology*, 94(2), 396–404.
- Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science*, 14(6), 1417–1430. <https://doi.org/10.1111/j.1467-7687.2011.01088.x>
- Quinn, P. C., & Liben, L. S. (2008). A sex difference in mental rotation in young infants. *Psychological Science*, 19(11), 1067–1070. <https://doi.org/10.1111/j.1467-9280.2008.02201.x>
- Raudenbush, S. W., Hernandez, M., Goldin-Meadow, S., Carrazza, C., Foley, A., Leslie, D., ... Levine, S. C. (2020). Longitudinally adaptive assessment and instruction to increase numerical skills of preschool. *Proceedings of the National Academy of Sciences*, 117(45), 27945–27953.
- Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25(1), 69–76. <https://doi.org/10.1177/01466216010251005>
- Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, 9(2), 195–212.
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. In (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan, & R. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–116) (Russell Sage Foundation).
- Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, 107(3), 645–662. <https://doi.org/10.1037/edu0000012>
- Reuterberg, S.-E., & Gustafsson, J.-E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement*, 52, 795–811. <https://doi.org/10.1177/0013164492052004001>
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4), 347–363. <https://doi.org/10.1177/0734282911406661>
- Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M Sattler, Publisher, Inc.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353. <https://doi.org/10.1037/1040-3590.8.4.350>
- 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. <https://doi.org/10.1111/1467-9280.02438>
- Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, 62(4), 273–296. <https://doi.org/10.1016/j.cogpsych.2011.03.001>
- Silverman, I., Choi, J., Mackewn, A., Fisher, M., Moro, J., & Olshansky, E. (2000). Evolved mechanisms underlying wayfinding: Further studies on the hunter-gatherer theory of spatial sex differences. *Evolution and Human Behavior*, 21(3), 201–213. [https://doi.org/10.1016/S1090-5138\(00\)00036-2](https://doi.org/10.1016/S1090-5138(00)00036-2)
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453. <https://doi.org/10.3102/00346543075003417>
- Spelke, E. S., Gilmore, C. K., & McCarthy, S. (2011). Kindergarten children's sensitivity to geometry in maps. *Developmental Science*, 14(4), 809–821. <https://doi.org/10.1111/j.1467-7687.2010.01029.x>
- Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42(5), 893–898.
- Thompson, M. S., & Green, S. B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock, & R. O. Mueller (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. Structural equation modeling: A second course* (pp. 163–218). IAP Information Age Publishing.
- Uttal, D., & Cohen, C. A. (2012). Spatial thinking and STEM education: When, why, and how? In B. Ross (Ed.), *Vol. 57. Psychology of learning and motivation* (pp. 147–181). New York: Elsevier. <https://doi.org/10.1016/B978-0-12-394293-7.00004-2>

- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2017). Links between spatial and mathematics skills across the preschool years. *Monographs of the Society for Research in Child Development, 82*(1). <https://doi.org/10.1111/mono.12263>
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250. <https://doi.org/10.1037/0033-2909.117.2.250>
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817–835. <https://doi.org/10.1037/a0016127>
- Wang, L., & Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educational Psychologist, 49*(4), 261–282. <https://doi.org/10.1080/00461520.2014.960568>
- Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maerlender, A. (2004). *Wechsler intelligence scale for children-integrated*. San Antonio, TX: The Psychological Corporation.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside.
- Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel, & M. P. Becker (Eds.), *Sociological methodology 2000* (pp. 165–200). Washington, D.C: ASA.