The Latent Structure of Spatial Skills and Mathematics:

A Replication of the Two-Factor Model

Kelly S. Mix
University of Maryland

Christopher J. Young
University of Chicago

Susan C. Levine
University of Chicago

David Z. Hambrick
Michigan State University

Yi-Ling Cheng
Michigan State University

Spyros Konstantopoulos
Michigan State University

Citation: Journal of Cognition and Development: 2017, v18, n4, pp. 465-492
Funding agency: Institute of Education Sciences (IES)
Grant #: R305A120416
Submission date: 2017
Publication date: 28 Jun 2017

Abstract

In a previous study, Mix et al. (2016) reported that spatial skill and mathematics were composed of 2 highly correlated, domain-specific factors, with a few cross-domain loadings. The overall structure was consistent across grade (kindergarten, 3rd grade, 6th grade), but the cross-domain loadings varied with age. The present study sought to replicate these patterns. Using the data from Wave 1 (n = 854) and data collected from a 2nd sample of kindergarten (n= 251), 3rd-grade (n = 247), and 6th-grade students (n = 241) with the same measures as in Wave 1, we carried out a multigroup confirmatory factor analysis to compare the 2 waves. We also completed several analyses of the Wave 2 data alone. The overall pattern obtained in Wave 1 — 2 highly correlated domain-specific factors — was clearly replicated in Wave 2. However, more subtle effects involving cross-domain loading were only partially replicated and generally appear fragile and context-specific. In Wave 2, we also included 2 new measures (i.e., proportion matching and fraction identification) that were analyzed in a separate model. Including these new measures did not change the overall pattern of factors and domain-specific factor loadings but did alter some of the cross-domain loadings.

The Latent Structure of Spatial Skills and Mathematics:

A Replication of the Two-Factor Model

Previous research has demonstrated a strong relation between spatial skill and mathematics. Those with better spatial skills perform better in mathematics and also go on to longer, more successful careers in science, technology, engineering, and mathematics fields (Casey, Nuttall, Pezaris, & Benbow, 1995; Geary, Saults, Liu, & Hoard, 2000; Laski et al., 2013; Lubinski & Benbow, 1992; Thompson, Nuerk, Moeller, & Cohen Kadosh, 2013). Neural and behavioral studies have indicated these correlations are based on shared processing (Hubbard, Piazza, Pine, & Dehaene, 2005; McKenzie, Bull, & Gray, 2003; Walsh, 2003); however, the nature of this shared processing and its developmental course are largely unknown. Indeed, recent work has indicated these relations shift depending on how spatial skill and mathematics are measured (Caviola, Mammarella, Cornoldi, & Lucangeli, 2012; Robert & LeFevre, 2013; Trbovich & LeFevre, 2003), suggesting that the shared processing may be more specific than is currently understood.

To investigate this shared processing and whether it changes over development, we conducted a cross-sectional study in which performance on a range of spatial and mathematics tasks was analyzed together using exploratory structural equation modeling (ESEM) and multiple regression (Mix et al., 2016). The three age groups targeted were kindergarten, third-grade, and sixth-grade students. The study revealed that various measures of spatial and mathematics skill formed two separate, but highly correlated domain-specific factors (see Table 1). The factor structure and correlations were consistent across age. Although all the tasks within each domain loaded significantly onto their respective factors, there also were significant cross-domain factor loadings—tasks that significantly loaded onto both the spatial and mathematics factors—and these cross-domain loadings differed across age. In kindergarten,

mental rotation and block design significantly cross-loaded onto the mathematics factor. In sixth grade, visual-spatial working memory (VSWM) and figure copying (Test of Visual-Motor Integration [VMI]) significantly cross-loaded onto the mathematics factor, and place value and algebra significantly cross-loaded onto the spatial factor. In third grade, there were no significant cross-domain loadings in the factor analysis; however, the multiple regressions revealed small but significant effects involving most of the same tasks that cross-loaded in kindergarten and sixth grade (i.e., mental rotation, VSWM, and VMI), suggesting that third grade is a transition period with numerous weak relations.

INSERT TABLE 1 HERE

We interpreted these results in terms of three broad mechanisms by which spatial processing might relate to mathematics: a) spatial visualization, b) form perception, and c) spatial scaling (Mix et al., 2016). As the following brief review will show, the previous study showed evidence for the first two mechanisms but not the third. The aim of the present study is to replicate and extend these findings with a second wave of data collection. For the replication, we carried out a parallel study using the same measures and including children drawn from many of the same schools used previously. There is a general need for replication in psychological research, as highlighted by recent reports (e.g., Pashler & Wagenmakers, 2012). This is particularly true in the case of spatial skill and mathematics, as the latent structure of each domain has been long debated with many conflicting claims (see Mix & Cheng, 2012, for a review). It was also important to replicate our specific design because large-scale studies such as this study are expensive and difficult to duplicate, and the structures we probe have substantial theoretical and practical implications (Newcombe, 2010; Verdine, Golinkoff, Hirsh-Pasek, & Newcombe, 2017). For the extension, we added two new measures (proportion matching and

fraction identification) that allowed us to follow up on questions raised in the original study. We report the replication and the extension as two separate studies. To distinguish the data collected in both the original study and present study, we will refer to them henceforth as Wave 1 and Wave 2, respectively.

Spatial visualization

Spatial visualization is the ability to imagine and mentally manipulate figures or objects in space. It could play a role in mathematics by helping children spatially ground concepts or represent a problem. Consistent with the notion that people ground symbolic and abstract thought in bodily movement through space (e.g., Barsalou, 2008; Lakoff & Núñez, 2000), we predicted particularly strong connections between spatial tasks that simulate movement and require dynamic visualizations of relative position, such as perspective taking, block design, and mental rotation, and mathematics tasks with relatively complex conceptualization requirements, such as interpreting word problems, comprehending place value, or fraction concepts. This prediction was borne out in Wave 1. First, the mathematics tasks with the strongest relations to spatial skill (i.e., place value, word problems, fractions, algebra) are known to have the highest representational demands. Second, the spatial tasks with the strongest relations to mathematics (i.e., block design, mental rotation, perspective taking) also have strong spatial visualization components. Thus, there were several indications that spatial visualization is a major source of shared variance between the two domains.

Form perception

Form perception is the ability to recognize shapes and tell them apart, distinguish shapes from their backgrounds, and flexibly shift focus between an object and its parts. This skill could relate to the symbol-reading demands of mathematics. When children read mathematical

symbols, they must make fine spatial discriminations, such as detecting the difference between a plus sign (+) and a minus sign (-), or noticing that 126 is different from 162 because the positions of "6" and "2" have shifted. Adults are sensitive to these relations, and their performance can be disrupted by subtle spatial shifts (Landy & Goldstone, 2007). In principle, shared processing based on form perception could be evident in mathematics tasks that require careful attention to symbolic notation, such as multistep calculation, missing-term problems, algebra, and interpreting charts and graphs, and spatial tasks that involve reproducing spatial locations and forms, such as VSWM, map reading, and figure copying.

Indeed, our results in Wave 1 showed form perception and mathematics were strongly related, but only in the oldest age group we studied (i.e., sixth grade). Specifically, only VSWM and figure copying (VMI) cross-loaded significantly onto the mathematics factor. Similar, albeit weaker, relations were evident in the third-grade regression analyses, but they were not exclusive as in sixth grade (i.e., several spatial visualization tasks also were significantly correlated with mathematics in third grade, as was the case for kindergarteners). We hypothesized that relations involving symbol reading and form perception might emerge after a procedure has become conceptually grounded and automatic. Prior to this, spatial visualization may play a larger role in the grounding process. Consistently, the relations between spatial visualization and mathematics were more evident in younger children, and relations involving form perception were more evident in older children. Moreover, figure copying (VMI) was significantly related to familiar mathematics content in kindergarten, and spatial visualization was significantly related to novel content in sixth grade, suggesting a developmental pattern that is recapitulated as children consolidate concepts and procedures and go on to learn new concepts and procedures.

Spatial scaling

A third possible type of shared processing could involve spatial scaling—the ability to distinguish absolute and relative distances and recognize equivalence across different spatial scales. This skill has a theoretical link to numeracy and symbol grounding in mathematics (Newcombe, Levine, & Mix, 2015), so it is a strong candidate for cross-domain overlap with mathematics. We thus expected to find strong connections between spatial tasks that require attention to relative distance or scaling, such as finding corresponding locations across representations of space at different scales (e.g., Möhring, Newcombe, & Frick, 2014), and mathematics tasks that focus on number meaning, such as number line estimation, as some have already shown (e.g., Slusser, Santiago, & Barth, 2013; Ye et al., 2016). However, these relations were not obtained in Wave 1. Number line estimation loaded significantly onto the mathematics factor, and map reading loaded significantly onto the spatial factor, but neither exhibited significant cross-domain relations. One concern could be that the task we used to assess spatial scaling was not a pure test. We used a map-reading task that involved matching maps to three-dimensional models across differences in scale but also involved remembering locations and mentally rotating the map relative to the model. In the present study, we followed up on this surprising null finding with a more direct measure of spatial scaling.

New versus familiar content

In addition to the evidence related to these three potential shared processes, we also observed an interesting developmental pattern in Wave 1. We found that fraction understanding was particularly related to spatial skill in third graders but not sixth graders, whereas missing terms/algebra were related to spatial skill in sixth graders but not third graders. This pattern suggested that spatial skill may be recruited for new or challenging content. To examine this pattern more directly, we coded all the items used in our mathematics measures based on the

Common Core State Standards for Mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). We considered items that appeared in the CCSS-M at or below the students' grade level to be "familiar" and items that appeared above the students' grade level to be "novel." In regressions between these content categories and the individual spatial measures, it appeared that spatial visualization was more related to novel content and form perception was more related to familiar content, as noted earlier. These results suggest we should find that performance on very simple fraction tasks are not strongly related to spatial skill in kindergarten, just as performance on very simple components of algebra (i.e., missing-term problems) was not related to spatial skill in third grade. However, we could not confirm this pattern in Wave 1 because kindergarteners did not complete fraction items. In Wave 2, these measures were included.

We also added these more simplified fraction-matching items to the third-grade battery to see whether it strengthened the relation to spatial skill. In Wave 1, there was no evidence of cross-domain loading from fractions to the spatial factor, but the relation between fraction performance and the spatial factor was significant in a multiple regression analysis. It seemed possible that by adding the fraction-matching items to the third-grade battery, this relation would be more evident, as it has been in previous studies focusing more directly on the relations between spatial skill and fraction understanding (Möhring, Newcombe, Levine, & Frick, 2016; Ye et al., 2016).

In summary, Wave 1 yielded several key findings. One was that spatial and mathematics measures formed separate, but highly correlated factors. Another was that some measures loaded onto both factors and the specifics of these cross-domain loadings differed across age. Finally, there was evidence that spatial visualization was particularly related to novel content and

form perception was particularly related to familiar content. The aim of the present study was to replicate and extend these findings. In Study 1, we used the same measures and age groups as before, but with a new sample of children to verify that the latent factor model we obtained previously is accurate and replicable. In Study 2, we analyzed these data along with additional data from the two new measures (proportion matching and fraction matching) to provide stronger tests of the process-driven explanatory mechanisms we posited previously.

Study 1

Replication of Mix et al. (2016)

Method

Participants. A total of 1,592 children participated in two waves. The data from Wave 1 (n = 854, collected in 2013-2014) were used in an exploratory factor analysis (EFA) reported previously (Mix et al., 2016). We refer readers to that article for complete details related to grade, sex, exclusions, and so forth. The main aim of the present study was to confirm the patterns we obtained in Wave 1, so we collected a second wave of data using the same tasks and age groups. In Wave 2 (2014-2015), a total of 738 children participated. The sample was drawn from 29 schools serving a range of rural, suburban, and urban communities in the Midwestern United States (11 communities). The average free/reduced-price lunch rate across the 11 communities in Wave 2 was 46.20% (range = 0% - 99%). Of these schools, 17 were the same as in Wave 1, and of the 11 communities, 7 were the same as in Wave 1. In terms of individual students, 58% of the Wave 2 students came from the same schools as in Wave 1. Thus, the populations were overlapping and diverse but not identical.

The study and its consent forms were approved by the institutional review boards (IRBs) at both universities. Midway through data collection, the study was deemed exempt in accord

with federal regulations of projects exempt from institutional review board (IRB) review. Children's parents were contacted through their schools, and only children whose parents signed an IRB-approved consent form were tested. Children from all 29 schools gave consent and were tested. The average response rate across schools was 20.4%. Of 4,276 children contacted, a total of 873 gave consent, and of these children, 134 were excluded because either a) tests were missing due to student absences, children who declined to participate, or schools that declined to participate after consents were turned in (n = 123); or b) tests were administered incorrectly or not recorded due to experimenter error (n =11). The final sample of 739 children was divided into three age groups: kindergarteners (n = 251, 132 boys; M_{age} = 6;0, SD = 4.32 months), third graders (n = 247, 96 boys; M_{age} = 9;1, SD = 4.56 months) and sixth graders (n = 241, 121 boys; M_{age} =11;10, SD = 5.28 months).

Procedure. Just as in Wave 1, children completed a battery of tests that measured spatial ability, mathematics, and verbal skill. All the specific measures used in Wave 1 were included in Wave 2. However, we included two additional measures not used before: Miura et al.'s (1999) fraction identification items were added in kindergarten and third grade, and Boyer and Levine's (2012) proportion-matching task was added in all three grades. (See Experiment 2 for a full description of these new measures).

Children were tested in three 1-hr sessions during the course of 2 weeks. Some tests were group administered (i.e., n = 4-6 for kindergarteners and third graders, and n = 20-30, or whole classes, for sixth graders). The other tests were administered individually. These details are provided in the "Measures" section. The tests were blocked based on whether they were individually administered or group-administered, but the order of tests within each block varied randomly. Also, the order of presentation for group versus individual tests was random and

counterbalanced across children. Children received a decorative folder as a reward for their participation.

Measures. We next describe each measure. Reliabilities were estimated using Cronbach's α (1951) and were calculated from the combined data set (Wave 1 and Wave 2) unless otherwise noted. Most of the reliabilities approached or reached $\alpha = .70$, which is the generally accepted cutoff, though not a hard and fast rule (Lance et al., 2006; Nunnally, 1978). In some cases (e.g., map reading) the reliabilities were less than .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 have argued 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; Peters et al., 1995).* In the kindergarten/third-grade version, small groups of children were shown 4 unfamiliar figures (i.e., forms based on manipulating components of capital letters) and were asked to indicate which 2 were the same as the target. The 2 matching items could be rotated in the picture plane to overlap the target, whereas the 2 foils could not be rotated because they were mirror images of the target. The task was introduced with 4 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 presented in a paper booklet (kindergarten, $\alpha = .74$; third grade, $\alpha = .87$). 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 4 choice drawings, 2 of which could be rotated in the picture plane to match the target ($\alpha = .81$).

Visual-spatial working memory *(adapted from Kaufman & Kaufman, 1983).* On each test trial, groups of children were shown a 14-cm x 21.5-cm grid that was divided into squares (e.g., 3 x 3, 4 x 3, or 5 x 5). Drawings of objects were displayed at random positions within the grid and were 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 = 15-29, based on grade) began immediately after the final practice trial (kindergarten, α = .77; third grade, α = .66, and sixth grade, α = .81).

Test of Visual-Motor Integration *(sixth edition; Beery & Beery, 2010).* On each trial, children copied a line drawing of a geometric shape on a blank sheet of paper. There were 18 to 24 trials, depending on the age of the child, during 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 .93.

Block Design *(Wechsler Intelligence Scale for Children-fourth edition; Wechsler et al., 2004).* On each trial, children were shown a printed figure composed 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 Wechsler Intelligence Scale for Children-Fourth Edition (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 .83 to .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 on a corresponding map. Kindergarten and third-grade students completed 14 test trials in which the model was a full-color three-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, two-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, α =.62; third grade, α =.69; sixth grade, α =.56).

Perspective taking. Kindergarten and third-grade children saw a set of Play Mobil figures and were asked to indicate which of four pictures was taken from each figure's perspective (Frick, Mohring, & Newcombe, 2014). The 27 test questions were preceded by 4 practice items with feedback (kindergarten, α = .64; third grade, α =.87). Sixth-grade children saw six to eight objects 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 2 practice items with feedback and 12 test items. Responses were scored based on the number of degrees they deviated from the correct angle on each item (α = .83).

Place value. Younger children completed a set of 20 items that required them to compare, order, and interpret multidigit numerals (e.g., "Which number is in the ones place?"), as well as match multidigit numerals to their expanded notation equivalents (342 = 300 + 40 + 2; kindergarten, α = .77; third grade, α = .81). Sixth-grade students completed the Rational Numbers subtest from the Comprehensive Mathematics Ability Test (CMAT; Hresko, Schlieve, Herron, Swain, & Sherbenou, 2003). These items similarly required them to compare, order, and

interpret written numbers, but the numbers were a mixture of multidigit whole numbers, fractions, and decimals (α =.83).

Word problems. Kindergarten and third-grade students were assessed using the 12 word problems from the Test of Early Mathematics Ability-Third Edition (Ginsburg & Baroody, 2003; kindergarten, α =.73; third grade, α =.65). The test was individually administered following the instructions in the test manual. Sixth-grade students completed the Problem Solving subtest from the CMAT (α =.76).

Calculation. We used a group-administered test consisting of age-appropriate arithmetic problems (kindergarten, α = .76; third grade, α = .70; sixth grade, α = .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-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., X + 9 = 12). Kindergarten and third-grade students completed eight such problems (kindergarten, α = .62; third grade, α = .71). Sixth-grade students completed the CMAT Algebra subtest (α = .68).

Number line estimation (Siegler & Opfer, 2003). All children were tested in small groups (n = 4-6). Given a stimulus card with a written numeral, they were asked to mark where it would go on a number line with a 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 = .39); third graders placed 3, 103, 158, 240, 297, 346, 391, and 907 on a 0-to-1,000 number line (split half reliability, r = .48); and sixth graders placed 25,000, 61,000, 49,000, 5,000, 11,000, 2,000, 15,000, 73,000, 8,000, and 94,000

on a 0-to-100,000 number line (split half reliability, r = .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 for Wave 1 (but see Experiment 2 for information about this measure in Wave 2). In third grade, we included 4 items that tested fraction equivalence and simple calculation with common denominators (α =.57). Sixth-grade students completed a 22-item test with fraction comparisons - calculation with and without common denominators, and calculation with mixed numbers (α =.73) - and a version of the number line estimation task wherein the anchors are 0 and 1 and the stimulus quantities are all fractions (i.e., 1/4, 1/19, 2/3, 7/9, 1/7, 3/8, 5/6, 4/7, 12/13, 1/2; split half reliability, r = .46; Siegler, Thompson, & Schneider, 2011).

Supplemental sixth-grade tests. The breadth of mathematics skills increases in middle school, so we assessed sixth graders' performance on two additional measures: CMAT Charts and Graphs (α =.79) and CMAT Geometry (α =.66). For Charts and Graphs, students were shown data in graphic form and were asked questions that require them to interpret the information. For Geometry, they identified shapes, defined geometric terms, solved equations, and so forth.

General cognitive ability. To estimate and control for children's general cognitive ability, we used 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 name a picture (e.g., "What kind of insect is this?"). The test was individually administered according to the instructions in the test manual (kindergarten, α = .73; third grade, α = .77; sixth grade, α= .74-79).

Results and discussion

Our main aim was to determine whether the latent factor model we reported in our previous study (Mix et al., 2016) is a plausible characterization of the Wave 2 data, as well as the data set as a whole. Toward that end, we report three analyses. First, we compared the test scores and demographic information for the two waves to confirm that the samples were comparable. Second, we carried out a multigroup confirmatory factor analysis (MGCFA; French & Finch, 2006; Sass, 2011) to determine whether the covariance matrices were statistically equivalent across waves and to check the validity of our hypothesized factor structure by applying it to both waves simultaneously (Byrne, Shavelson, & Muthén, 1989; van de Schoot, Lugtig, & Hox, 2012). Third, we carried out an EFA using only the Wave 2 data.

Descriptive statistics comparing Wave 1 and Wave 2. Table 2 presents the means and standard deviations of children's scores on each measure, by grade (kindergarten, third grade, and sixth grade) and testing wave (Wave 1 and Wave 2), as well as their mean ages. We used two-tailed *t* tests to determine whether these values differed. The critical alpha level was set at p = .004 (kindergarten and third grade) and p = .003 (sixth grade) to control for multiple comparisons. The performance of the two samples was mostly comparable, with only a few significant differences in test scores (see Table 3). There also was a significant age difference between the sixth-grade samples, such that the children in Wave 2 were 1 month older on average ($M_{\text{WAVE 1}}$ = 140.4 months, $M_{\text{WAVE 2}}$ = 141.6 months), t (527) = 2.55, p = .01.

INSERT TABLE 2 HERE

Multigroup confirmatory factor analysis. To determine whether the same latent structures were present in both Wave 1 and Wave 2 data, we used a MGCFA (also known as a measurement invariance modeling). In this approach, a series of models is applied to data from different samples simultaneously to see whether the same patterns are obtained in both groups. The models are applied progressively from unconstrained to increasingly constrained models. By comparing unconstrained and constrained models in a stepwise fashion, we can determine how the models for multiple data sets correspond and how they diverge (Muthén & Asparouhov, 2013).

As a first step, we performed an omnibus between-groups test to see whether the covariance matrices themselves differed from Wave 1 to Wave 2 (e.g., Vandenberg & Lance, 2000). It would be unusual, though not impossible, for different samples to have statistically equivalent covariance matrices. Not surprisingly, the χ^2 tests comparing the two waves revealed significant differences at each grade level: kindergarten, $\Delta\chi^2$ (77) = 161.01, $p < .0001$; third grade, $\Delta\chi^2$ (90) = 135.44, $p = .001$; sixth grade, $\Delta\chi^2$ (135) = 228.92, $p < .0001$. This outcome indicates that the covariance matrices were not equivalent, but it does not indicate to what degree or in what specific ways they differed. Possible differences include the structure of the factors and their loadings (i.e., configural variance), the magnitude of the factor loadings, or slope (i.e., metric variance), and the test scores themselves, or intercepts (i.e., scalar variance). For the purpose of replicating the results for Wave 1, the most important of these differences is configural variance, as it refers to the pattern of factors and factor loadings.

Thus, to determine how the samples differed, we evaluated the fit of three models designed to isolate and test each of these parameters separately. We used three goodness-of-fit indices to determine whether each successive model had an acceptable fit: a) the root mean

square error of approximation (RMSEA), b) the Comparative Fit Index (CFI), and c) the standardized root mean residual (SRMR). We considered models with RMSEA values < .08, CFI values > .95, and SRMR values < .08 to have acceptable fit (Hu & Bentler, 1999; Kline, 2005; Raykov & Marcoulides, 2006). We also used Satorra-Bentler Scaled X^2 tests (Satorra & Bentler, 2001) to compare the fit between the more constrained models and their less constrained predecessors. Significant X^2 tests would indicate the two groups are not invariant (i.e., they differ) along the dimension that was constrained.

All analyses controlled for differences in general cognitive ability by specifying models that used the residualized covariance matrix after partialing out children's WJ-3 Vocabulary scores. As noted earlier, vocabulary is highly correlated with overall intelligence. Also, the reported analyses all used maximum likelihood estimation with robust standard errors (i.e., MLR) to guard against non-normal distributions in the explanatory variables. MLR uses Huber sandwich estimation to provide standard errors that are robust against specification errors due to non-normal distribution (Freedman, 2006; Muthén & Muthén, 2012)—an approach that has proven successful in simulation studies with distributions ranging in skewedness from -2 to 2 (Chou & Bentler, 1995). An examination of the distributions of scores used in the present study confirmed that all fell within this range, except for the 0-to-100,000 number line estimation task in sixth grade, which was slightly skewed (-2.16). We repeated the analyses after correcting for this skew using a Box-Cox transformation (Osborne, 2010), but we obtained the same pattern of results.

In Model 1, we tested configural invariance by holding the factor structure constant but freeing the other parameter estimates. That is, we specified both the number of factors (two) and which variables loaded onto each factor, based on the results of Wave 1. This step is essentially

the same as fitting our hypothesized factor structure to both waves simultaneously. The results are presented in Table 3. The fit of the model was good (kindergarten, RMSEA = .05, 90% CI [.03, .06], CFI = .97, SRMR = .03; third grade, RMSEA = .06, 90% CI [.05, .07], CFI = .96, SRMR = .04; sixth grade, RMSEA = .04, 90% CI [.03, .05], CFI = .98, SRMR = .03). These numbers indicate that configural invariance was achieved. That is, there were two separate but correlated factors (one for space and one for mathematics) in both waves. Because the model was simultaneously fit to both waves, the goodness-of-fit statistics also apply to both waves, so as an added check, we applied the model to the Wave 2 data separately. We again found acceptable fit (kindergarten, RMSEA = .05, 90% CI [.02, .07], CFI = .97, SRMR = .04; third grade, RMSEA = .07, 90% CI [.6, .09], CFI = .94, SRMR = .04; sixth grade, RMSEA = .05, 90% CI [.03, .06], CFI = .97, SRMR = .04). As in Wave 1, the spatial and mathematics factors were highly correlated (kindergarten, r = .67; third grade, r = .73; sixth grade, r = .55). These findings demonstrate that the two-factor model obtained in the exploratory analysis of Wave 1 is plausible for the Wave 2 data and thus is replicated.

INSERT TABLE 3 HERE

Further, the within-domain factor loadings also replicated as all the spatial tasks loaded significantly onto the spatial factor and all the mathematics tasks loaded significantly onto the mathematics factor. However, not all the cross-domain loadings observed in Wave 1 were evident in Wave 2. Specifically, although VSWM significantly cross-loaded onto the mathematics factor in sixth grade, neither mental rotation nor block design cross-loaded in kindergarten and neither algebra, place value, or VMI cross-loaded in sixth grade. In a few cases (algebra and VMI in sixth grade), the magnitudes of the loadings were comparable across waves, but the loadings did not reach significance in Wave 2 because the standard errors were greater.

Thus, even though the overall model fit both waves of data well, some subtle effects were not the same.

The finding of configural invariance in Model 1 also indicated that the difference in covariance matrices we reported earlier was due to one of the other parameters (i.e., metric or scalar invariance). These differences are of less theoretical interest, but we tested two additional models to specify the sources of disparity. Model 2 probed for metric invariance (i.e., differences in the magnitude of the factor loadings) by specifying equal factor loadings while allowing the intercepts and residuals to vary freely. The fit for the constrained model was good for all three grades (kindergarten, RMSEA = .05, 90% CI [.04, .07], CFI = .96, SRMR = .08; third grade, RMSEA = .06, 90% CI [.05, .07], CFI = .95, SRMR = .06; sixth grade, RMSEA = .04, 90% CI [.03, .05], CFI = .98, SRMR = .05). However, the X^2 tests comparing the fit of Model 1 to Model 2 were significant in both kindergarten, $\Delta\chi^2$ (11) =30.92, p = .001, and third grade, $\Delta\chi^2(10)$= 26.07, p =. 01, suggesting that the magnitude of the factor loadings was not exactly the same in both waves. In sixth grade, the difference approached but did not reach significance, $\Delta\chi^2$ (17) = 26.65, p = .063.

To determine the source of the differences in kindergarten and third grade, we identified the parameters with the highest modification index values (MIV) and freed them, one by one, from largest to smallest until we obtained partial invariance for the two groups. MIVs estimate the influence of specific parameters, and in a well-fitting model, they should generally be low. However, freeing those with higher values (i.e., those at or above the critical values of 4) can significantly improve model fit and may indicate where, specifically, the two samples differ (Brown, 2006). For kindergarten students, partial metric invariance was achieved when the variable VMI was freed (MIV = 21.00_{WAVE1} and 20.98_{WAVE2}), $\Delta\chi^2$ (10) = 10.24, p = .42. For

third-grade students, invariance was achieved by freeing the factor loadings on VMI (MIV = 6.82 $_{WAVE1}$ and 6.82 $_{WAVE2}$) and missing-term problems (MIV = 6.20 $_{WAVE1}$ and 6.20 $_{WAVE2}$), $\Delta\chi^2$ (8) =13.19, p = .11.

These results indicate that differences in the factor loadings for these measures (VMI in kindergarten and VMI and missing terms in third grade) can explain the failure to achieve full metric invariance in Model 2. Note, however, that these differences reflect quantitative, but not qualitative, differences. For example, VMI loaded significantly onto the third grade spatial factor in both waves. The difference highlighted by this analysis is merely a larger loading in Wave 2. An interesting side effect, however, was that specifying these loadings as equal altered some of the cross-domain loadings. Specifically, in kindergarten, mental rotation cross-loaded onto mathematics in both waves, but block design still did not cross-loaded in either wave (see Table 4). This finding provides further evidence that the cross-domain loadings in these models may be context-dependent and unstable.

The third and final model was used to evaluate scalar invariance, or the equality of absolute level of performance. To test for this invariance, we constrained both the intercepts and factor loadings and then compared the fit of this new model to Model 2. As before, if the fit is not worsened by this constraint, it suggests the performance levels are equal across groups. This was the case in kindergarten, $\Delta\chi^2$ (8) = 14.10, p = .08, suggesting that the latent structures were very similar. In third grade, the difference was significant, $\Delta\chi^2$(8) = 21.62, p = .006, but partial scalar invariance was obtained after relaxing the scores for map reading (MIV = 10.64 $_{WAVE1}$ and 10.64 $_{WAVE2}$), $\Delta\chi^2$(6) = 5.15, p = .52. In sixth grade, the difference between models also was significant, $\Delta\chi^2$ (13) = 24.12, p = .03, but partial scalar invariance was obtained after relaxing the scores for word problems (MIV = 6.51 $_{WAVE1}$ and 6.51 $_{WAVE2}$), $\Delta\chi^2$ (12) = 17.59, p = .13.

These differences do not change the overall pattern in terms of replicating the latent structure. They simply highlight tests for which children in one wave performed worse than children in the other. The factor loadings for this final model are presented in Table 4.

INSERT TABLE 4 HERE

Overall, the MGCFA provided strong evidence that the two-factor model reported previously for Wave 1 (Mix et al., 2016) was replicated in Wave 2. The same configural model fit both data sets well, and aptly characterized the factor structure and loadings, at least within domains. The previous claim that spatial skill and mathematics form separate, but highly correlated, factors was clearly supported. The few differences obtained between the two waves were related to the size, but not significance, of the within-domain factor loadings for specific tasks and differences in overall performance on few tasks. Notably, however, the significant cross-domain loadings obtained for Wave 1 were only partially replicated in Wave 2. In the final model, achieved after relaxing the few measures that were discrepant across waves, most of the previously reported cross-domain loadings were obtained, which was an encouraging result. Yet, even after achieving invariant models, the cross-domain loading of block design in kindergarten was not replicated in Wave 2. Also, the cross-domain loading for place value in sixth grade was only marginally significant in the final model. These findings suggest caution in interpreting these specific effects. Also, the fact that cross-domain loadings sometimes shifted when other variables were relaxed reminds us that the outcomes of factor analysis are sensitive to all the measures included. In sum, the cross-domain loadings, even those that were replicated, appear much more fragile and context-specific than the within-domain loadings, and thus, they may be difficult to interpret.

Exploratory factor analysis for Wave 2. The MGCFA demonstrated that the same

model could plausibly apply to both Wave 1 and Wave 2, and that with a few exceptions, it resulted in the same factor loadings in both data sets. Does this mean that if we simply carried out an unconstrained ESEM on the Wave 2 data, we would obtain the same results reported for Wave 1 (Mix et al., 2016)? One might argue this approach would be the most straightforward way to replicate our previous finding. Yet, EFAs are notoriously difficult to replicate (Costello & Osborne, 2005; Osborne & Fitzpatrick, 2012). For example, using samples of 260 participants drawn randomly from a very large data set (n = 24,599), Costello and Osborne (2005) found that the factor structure of repeated EFAs replicated only 70% of the time (given a 20:1 ratio of participants to measures). This failure to replicate can occur if a sample, though random, is skewed or otherwise unrepresentative of the overall population. We know already from the MGCFA that the correlation matrices for the two waves in the present study differed, suggesting different sampling distributions, but we do not know what patterns might have emerged if we had simply explored the structure of the Wave 2 data in a separate analysis, without testing a confirmatory model.

To find out, we tested an EFA model using ESEM for the Wave 2 data, following precisely the data analysis procedures outlined by Mix et al. (2016). Specifically, we submitted the children's raw scores on each measure to an oblique Geomin rotation in the Mplus 7.0 program (Muthén & Muthén, 1998-2012) after first partialing out children's WJ-3 Vocabulary scores. For each analysis, we determined the optimal number of factors using 95% confidence intervals around each factor's eigenvalue (see Larsen & Warne, 2010) and rejected models with factors for which the lower bound of the confidence interval was 1.00 or less. After identifying the number of informative factors, we determined the optimal rotation for each model and evaluated model fit using the indices described earlier. Once the model with the best fit was

identified, we determined which tasks loaded onto each factor significantly using *z* values derived by dividing the factor loading for each measure by its standard error (Cudeck & O'Dell, 1994; Schmitt & Sass, 2011).

INSERT TABLE 5 HERE

The factor loadings are presented in Table 5. At all three grade levels, the first two factors had adequate eigenvalues (kindergarten, Factor 1 = 3.82, 95% CI [3.15, 4.48], Factor 2 = 1.39, 95% CI [1.14, 1.63]; third grade, Factor 1 = 4.65, 95% CI [3.83, 5.48], Factor 2 = 1.30, 95% CI [1.07, 1.53]; sixth grade, Factor 1 = 5.67, 95% CI [4.66, 6.69], Factor 2 = 1.84, 95% CI [1.51, 2.17], whereas the third factor did not (kindergarten, 0.94 [0.77, 1.10]; third grade, 1.08 [0.89, 1.27]; sixth grade, 0.99 [0.82, 1.17]). The fit of the two-factor models was good in kindergarten (RMSEA = .05, 90% CI [.03, .08], CFI = .97, SRMR = .03) and sixth grade (RMSEA = .05, 90% CI [.03, .06], CFI = .97, SRMR = .03), but it was marginal in third grade (RMSEA = .09, 90% CI [.08, .11], CFI = .93, SRMR = .04). As we reported previously for Wave 1 (Mix et al., 2016), one of these factors was primarily spatial and the other was primarily mathematical. Also as before, the factors were highly correlated at each grade level (kindergarten = .53; third grade = .54; sixth grade = .51), and these correlations did not increase or decrease with age (z = 0.16 -0.45, p = 0.9 - 0.7). Thus, the factor structure we reported for Wave 1 was replicated.

We also found significant cross-domain loadings in the Wave 2 ESEM, but they were not exactly the same as in Wave 1 (see Table 3). The key similarities were that a) in kindergarten, performance on mental rotation loaded significantly onto the mathematics factor in both waves, and b) in sixth grade, both VSWM and VMI loaded significantly onto the mathematics factor in both waves. Recall that though these cross-domain loadings were not significant for Wave 2 in

the MGCFA configural model, they were significant in the final model. Another similarity was that fraction understanding was significantly correlated with the spatial factor in third grade in the Wave 1 regression analyses and also appeared as a significant cross-domain loading with the spatial factor in the Wave 2 ESEM. Note that this finding emerged even without including the more simplified fraction-matching items that were added in Wave 2 (and will be analyzed in the "Study 2" section). These are all crucial similarities suggesting at least some consistency in the pattern of cross-domain loadings.

However, there also were differences. First, in contrast to Wave 1, there was a significant cross-domain loading for VSWM in kindergarten that was not obtained previously. Second, we did not find that algebra and place value (as measured on the Rational Numbers subtest of the CMAT) cross-loaded significantly onto the spatial factor in sixth grade. In Wave 1, these variables had cross-loaded significantly, but only in analyses that included the supplemental sixth-grade tasks (e.g., CMAT Charts and Graphs, CMAT Geometry, and fraction number line estimation). When the supplemental tasks were excluded, the cross-domain loadings were no longer significant. Here, even with these tasks included, the cross-domain loadings were not obtained. These shifting patterns of significance again suggest caution in interpreting these particular cross-domain loadings. Third, there were two unexpected, negative cross-domain loadings in Wave 2 (whereas there were none in Wave 1). In third grade, map reading loaded negatively onto the mathematics factor, and in sixth grade, fractions loaded negatively onto the spatial factor. Negative loadings imply that children who performed better on these tasks conversely performed worse on the skills that formed the factor. So, for example, third graders with strong map-reading performance actually performed worse on mathematics than did children who were not as skilled at map reading. Note that these two measures

nonetheless had positive loadings onto their respective domain-specific factors, and these domain specific factors were positively correlated, so the overall picture includes strong positive cross-domain relations for these tasks. Therefore, only some of the processes measured by these tasks load onto the other factor negatively. Further research will be needed to specify what these aberrant processes might be.

In sum, the ESEM for Wave 2 replicated the ESEM for Wave 1 in terms of the factor structure, all the within-domain factor loadings, and some of the cross-domain factor loadings. As noted earlier, the likelihood of replicating an ESEM in two samples is far from certain (Costello & Osborne, 2005), so it is a remarkable finding and one that inspires confidence in the previously reported results. However, not all cross-domain loadings from Wave 1 were replicated in Wave 2, providing further evidence that these effects are highly sensitive to sampling differences and perhaps, in some cases, spurious.

Study 2

Wave 2 exploratory analysis with new measures

Recall that two new measures were added to the test battery in Wave 2: proportion matching and fraction identification. In the analyses reported so far, we have not included these measures because our aim was to determine whether the latent structure was the same in two samples using the same battery. In the present analysis, we examined only the Wave 2 data but included these new tasks to see if they caused the latent structures to shift. As noted earlier, such shifts due to inclusion or exclusion of measures have been documented in related work (e.g., Caviola et al., 2012; Trbovich & LeFevre, 2003).

There were theoretical reasons for examining these two skills in particular. As noted in the introduction, spatial scaling is one possible process hypothesized to have particularly strong

ties to mathematical skill (e.g., Möhring et al., 2014; Newcombe et al., 2015). In Wave 1, we failed to obtain evidence for this connection, but it may have been because we included only an indirect measure of spatial scaling (i.e., map reading). In Wave 2, we added a proportion-matching task that provides a more direct assessment of spatial scaling and has been used in previous research demonstrating a connection between spatial-scaling and mathematical skills, such as fraction understanding (e.g., Möhring et al., 2016; Ye et al., 2016).

With respect to fraction concepts, we sought to follow up on an interesting developmental pattern from Wave 1, in which fraction understanding was particularly related to spatial skill in third but not sixth graders and missing-terms problems/algebra were strongly related to spatial skill in sixth graders but not third graders. In kindergarten children, there was also a strong relation between spatial skill and calculation. This pattern suggested that spatial skill may be recruited for new or challenging content (i.e., calculation in kindergarten, fractions in third grade, and algebra in sixth grade). The finding that some components of algebraic reasoning, as measured in missing-term problems, were not strongly related to spatial skill raised the question of whether fraction concepts in kindergarteners would show the same pattern (i.e., no particular relation to spatial skill, despite its novelty). This pattern might arise if content is so novel that children are either performing at floor or have not encountered sufficiently challenging material to recruit spatial skills. If so, it would suggest a sweet spot in learning for which spatial skills become particularly relevant, while children are in the process of mastery and struggling with novel skills, but only when the content is neither too novel nor too familiar. We could not test this prediction in kindergarteners previously because in Wave 1, we had not given fraction tasks to the kindergarten sample. In Wave 2, these measures were included. We also added these relatively simplistic fraction identification items to the third-grade battery on the hypothesis

these items may reveal a stronger relation between spatial skill and fraction understanding in particular, consistent with recent work (Möhring et al., 2016; Ye et al., 2016).

Measures. Children in Wave 2 received the same measures presented to children in Wave 1 (see Study 1 for a full description), plus two additional measures (described in this section). All measures, including these two, were presented in a random order that varied across children.

Proportion matching. *(adapted from Boyer & Levine, 2012).* Stimulus displays consisted of three columns with different proportions of red space versus blue—a standard and two choices—presented on a laptop computer. Next to the standard, there was a picture of a pig, and children were told, "Harry the Hog enjoys drinking all kinds of juice, and likes to mix the juice himself. Harry must be careful to have the correct mix of water and juice for each type of mix. Which of these two (pointing to the two alternatives) is the right mix for the juice Harry the Hog is trying to make? Which of these two would taste just like Harry's juice? Circle one!" Children sat in groups for stimulus presentation, but they circled their responses in individual paper-test copies. There were 20 to 24 test trials depending on the child's grade. On each trial, the target appeared on the left side, and the two response choices appeared on the right side in a horizontal row. The side of the correct choice relative to the foil (either near or far) was counterbalanced across items (kindergarten, $\alpha = .70$; third grade, $\alpha = .90$; and sixth grade, $\alpha = .82$).

Fraction identification *(adapted from Miura et al., 1999).* Children were shown a written fraction (e.g., $\frac{1}{2}$) and were asked to mark one of four schematic drawings that showed the correct portion shaded. On each trial, the four drawings varied in shape (circle, square, rectangle) and in how the portions were divided (bars, quadrants, etc.). The foils were

constructed so that one matched the prompt in terms of the numerator, one matched in terms of the denominator, and one matched neither the numerator nor the denominator. For 3 of the items, the nonmatch was replaced with a Numerator + Denominator foil (Paik & Mix, 2003). In this foil, the numerator was represented correctly, but the denominator was represented by unshaded pieces, so that the total number of pieces was the numerator added to the denominator (e.g., the fraction $\frac{1}{2}$ would be represented as one part shaded and two parts unshaded, or three parts total). In previous research, the Numerator + Denominator foil was particularly difficult for first-grade students (Paik & Mix, 2003), so it was included to provide a sufficient range of difficulty for the third-grade students. On each trial, the experimenter read the fraction name aloud and said, "Circle the picture that is [fraction name] shaded." Kindergarten students received only the 11 fraction-matching items. For third-grade students, we combined these items with the 4 fraction items used in Wave 1 (i.e., 2 equivalence and 2 calculation) and analyzed the summed score out of 15 possible points (kindergarten, $\alpha = .50$; third grade, $\alpha = .69$ for the 11 new items and $\alpha = .70$ for the 15 items total).

Results and discussion

We tested three ESEMs—one at each grade level—in which both spatial and mathematics measures were considered together. As in our previous work and in Study 1, the analyses were carried out with an oblique Geomin rotation in the Mplus 7.0 program (Muthén & Muthén, 1998-2012) using the raw scores for all measures and MLR to guard against specification errors due to non-normal distribution (Freedman, 2006; Muthén & Muthén, 2012; Wang & Wang, 2012). An examination of the proportion-matching and fraction scores (i.e., fraction identification only in kindergarten and the composite scores used in third grade) used in the present study confirmed

that all fell within an acceptable range of skewedness from -2°to 2° (Chou & Bentler, 1995; Chou, Bentler, & Satorra, 1991).

As before, we first extracted factors until they no longer added significant explanatory power as indicated by their eigenvalue confidence intervals (Larsen & Warne, 2010) and then determined the optimal rotation for each model and evaluated model fit using RMSEA (Steiger, 1990), CFI (Hu & Bentler, 1999; Raykov & Marcoulides, 2006) and SRMR (Kline, 2005). Once the model with the best fit was identified, we determined which tasks loaded onto each factor significantly following our previously established procedures of deriving *z* scores (Cudeck & O'Dell, 1994; Schmitt & Sass, 2011).

In all three grades, the first two factors had adequate eigenvalues (kindergarten, Factor 1 = 3.99, 95% CI [3.29, 4.69], Factor 2 = 1.39, 95% CI [1.15, 1.63]; third grade, Factor 1 = 4.75, 95% CI [3.91, 5.59], Factor 2 = 1.33, 95% CI [1.10, 1.57]; sixth grade, Factor 1 = 5.85, 95% CI [4.81, 6.90], Factor 2 = 1.84, 95% CI [1.51, 2.17]), but the third factor did not (kindergarten, 1.13, 95% CI [0.93, 1.33]; third grade, 1.13, 95% CI [0.93, 1.33]; sixth grade, 1.07, 95% CI [0.88, 1.26].[1] The fit of the two-factor models was good in kindergarten (RMSEA = .05 [range = .03-.07], CFI = .96, SRMR = .04) and sixth grade (RMSEA = .04 [range = .03-.06], CFI = .97, SRMR = .03) but was, again, marginal in third grade (RMSEA = .09 [range = .07-.10], CFI = .92, SRMR = .04).

INSERT TABLE 6 HERE

As can be seen in Table 6, the first two factors at each grade level were the same as before—a spatial factor with all the spatial tasks loaded onto it and a mathematics factor with all the mathematics tasks loaded onto it. Interestingly, proportion matching loaded onto the mathematics factor and not the spatial factor in kindergarten and sixth grade. Proportion

matching loaded onto neither factor in third grade. For kindergarten and sixth-grade students, the cross-domain loadings were the same as in the ESEM without proportional reasoning or the new fraction items. Note that it included the finding of no significant cross-domain loading for fraction understanding onto the spatial factor in kindergarten.

In third grade, however, there was a shift. First, whereas fractions had cross-loaded significantly onto the spatial factor in the Wave 2 ESEM, it was not the case when both the new fraction identification items and the proportion-matching items were included. Instead, fraction skill loaded onto the mathematics factor only. To determine which of the additions caused this shift, we repeated the analysis with only the fraction identification items added (i.e., without adding proportion-matching scores) and found that again, third graders' fraction scores did not cross-load onto the spatial factor (p = .72); however, when proportion-matching items were added and only the previous fraction items (i.e., the fraction calculation items used in Wave 1) were included, fraction skill continued to cross-load onto spatial skill significantly (p = .01). Thus, it appears that only the more advanced fraction skills were related to spatial skill strongly enough to cross-load.

Second, the correlation between the two factors (space and mathematics) was significantly greater in third grade when proportion matching and fraction identification were included (r =.54 vs. .69, z = 2.56, p = .01). We again repeated the analysis adding only one task or the other, and found that when only fraction identification items were added to the model but proportion-matching items were not, the correlation between factors similarly increased (r =.68); however, when only proportion matching items were included, the correlation between factors was the same as in the original model without either proportion matching or fraction identification added (r =.54). Note that the interfactor correlations in the other grades were not

significantly increased or decreased due to the inclusion of either proportion matching or fraction identification items.

Taken together and including the results from Wave 1, there was a reliable relation between fraction skill and spatial skill in third grade. The pattern of cross-domain loadings suggests this relation may be particularly strong for fraction calculation and comparison items—a pattern that may reflect the novel-familiar cycle we noted previously (Mix et al., 2016), wherein spatial skill is recruited for less familiar, more complex mathematics material but may not play a role once skills are mastered and, perhaps, automatized. However, the fact that including simple fraction identification items strengthened the interfactor correlations at this grade level suggests these simpler fraction items also contribute significantly to the shared variance, perhaps at a more general level. In contrast to the findings related to fraction understanding, there was no evidence that including proportion matching in the model affected the factor structure or factor loadings.

General discussion

The present study is a replication and extension of an EFA that examined the latent structure of spatial and mathematical skills combined (Mix et al., 2016). In the previous study, we found that spatial skill and mathematics formed two separate but highly correlated factors, onto which all the variables in the respective domains loaded significantly (i.e., all the spatial tasks loaded onto the spatial factor and all the mathematics tasks loaded onto the mathematics factor). We also found that certain variables significantly loaded onto both factors and, further, that the specific pattern of cross-domain loadings changed from kindergarten to sixth grade. In the present study, we collected a second wave of data using the same measures as before in the same age groups from a mostly overlapping and diverse population. Through a series of

analyses, we sought to determine whether the same patterns were obtained in Wave 2. We also added two measures: a) a simplified fraction measure (i.e., fraction identification) in kindergarten and third grade, and b) a spatial-scaling measure (i.e., proportion matching) in all three age groups. In a second set of analyses, we examined whether these measures changed the pattern of results or revealed evidence of additional shared processing.

A consistent factor structure

With respect to the first question, the latent structure we described in our previous work was clearly replicated. In an MGCFA, we found that a two-factor model provided a strong fit to the data from both waves and, further, that the variable loadings were domain-specific. That is, all the spatial measures loaded significantly onto one factor and all the mathematics measures loaded significantly onto the other. As in our previous work, these two factors, though distinct, were highly correlated. The same pattern was obtained when we fit the confirmatory model to the Wave 2 data alone and when we used an unconstrained exploratory model instead. Regardless of how we constructed the model or to which data we applied it, there were two highly correlated, domain-specific factors.

We previously interpreted this pattern to mean one of two things. One is that the common variance is attributable to general cognitive ability. Although we partialed out children's vocabulary scores to control for general cognitive ability, it is possible that other aspects of intelligence, such as fluid processing, are implicated in both spatial and mathematical processing, which explains the correlation. Alternatively, one could argue that spatial processing itself provides a format for abstract thought that is common to both domains (Lohman, 1996). Because the two domains have unique demands in addition to their shared processing, this commonality may not be enough to yield a single factor but could explain the high correlation.

Distinguishing between these interpretations is not possible based on our data and in some ways hinges on basic, unsettled questions regarding the nature of intelligence, such as whether spatial processing is separable from general intelligence. Still, additional research that includes spatial and mathematics tasks but attempts to control for fluid processing may be helpful.

Partial replication of cross-domain loadings

Across the various analyses, there were significant cross-domain loadings—variables that loaded onto both their domain-specific factor and the other domain-specific factor. Also, as before, these variables varied from one grade level to the next. However, the specific variables that cross-loaded were not entirely consistent across analyses in the present study, nor were they entirely consistent with those we reported previously for Wave 1 (Mix et al., 2016).

The most consistent cross-domain loadings, reaching at least marginal significance in most of our analyses, were mental rotation in kindergarten and VSWM and VMI in sixth grade. In light of the low probability of replication in multiple EFAs (Costello & Osborne, 2005) and the relative fragility of these cross-domain loadings, it is remarkable that these consistent findings were obtained. This outcome may signal important, age-specific points of contact between the two domains.

However, the other cross-domain loadings were less consistent. The spatial measure, VSWM, did not cross-load onto the mathematics factor in Wave 1 for the other grades (see Table 1), but it did cross-load in Wave 2 for kindergarten. This new finding accords with previous work showing that VSWM is related to mathematics in young children (Bull, Espy, & Wiebe, 2008; Holmes, Adams, & Hamilton, 2008), and it tempers our previous claims of a strong age-related trend in shared processing from spatial visualization in younger children to form perception in older children. In terms of mathematics measures, the previously reported cross-

domain loadings for place value and algebra in sixth grade were evident in the MGCFA after relaxing the parameter of word problems, but not in the Wave 2 EFA. In the other grades, there were no significant mathematics-to-spatial cross-domain loadings in Wave 1, but one emerged in the Wave 2 data for third grade (namely, fractions). Interestingly, fractions also cross-loaded onto the spatial factor in the Wave 2 sixth-grade data, but the loading was negative.

The general instability of these cross-domain loadings suggests cautious interpretation. Still, the overall pattern may have meaning. We suggested previously (Mix et al., 2016) that the role of spatial processing in mathematics might vary depending on the familiarity of the content. Indeed, when we divided the mathematics items into novel and familiar categories, we found shifts in the cross-domain loadings such that novel content seemed more strongly related to spatial visualization and familiar content seemed more strongly related to form perception. These patterns were not clear-cut, but they were consistent enough to raise the possibility that shifts in underlying processing might occur at the level of tasks and student knowledge, rather than appearing as broad developmental changes seen in comparisons by age. On this account, VSWM may have cross-loaded in kindergarten for Wave 2 but not Wave 1 because the Wave 2 kindergarteners were more familiar with the mathematics content and thus had transitioned to a stronger role for form perception. Similarly, fractions may have cross-loaded in third grade for Wave 2 but not Wave 1 because Wave 2 third graders were less familiar with the content. Consistent with this idea, the significant cross-domain loading for fractions in third grade disappeared when more simplistic items were added to the battery, perhaps because spatial skill was engaged less for familiar mathematics content. Obviously, even these interpretations are speculative, but they provide an account of these shifts that is at least plausible and may bear further investigation.

Shared processing

We have identified three processes that might support performance of both spatial and mathematical tasks and explain their shared variance: a) spatial visualization, b) form perception, and c) spatial scaling. As noted earlier, the cross-domain loadings obtained here and in Wave 1 are indicative of a role for the first two processes. Specifically, the consistent cross-domain loading of mental rotation onto the mathematics factor in kindergarten is suggestive of a role for spatial visualization, and the consistent cross-loading of VSWM and VMI in sixth grade (as well as the emergence of this cross-loading in kindergarten in Wave 2) suggests a role for form perception and location memory. However, we did not find evidence for a connection between spatial-scaling and mathematics in Wave 1. Our spatial-scaling measure (i.e., map reading) loaded onto the spatial factor only, and number line estimation loaded onto the mathematics factor only. This pattern was replicated in Wave 2 (see Table 5).

It was a bit surprising that number line estimation did not cross-load onto the spatial factor given that spatial skill is a significant predictor of number line estimation (Gunderson, Ramirez, Beilock, & Levine, 2012). Also, models that explain number line estimation in terms of proportional reasoning seem to implicate a spatial process (e.g., Slusser et al., 2013). A possible explanation we evaluated in Wave 2 was whether the map reading task was a poor measure because it involved spatial and nonspatial task demands beyond reasoning about scaling. If spatial scaling is the component shared with number line estimation, it seemed possible that a more direct test of spatial scaling would either lead to a significant cross-domain loading or reveal a third spatial-scaling factor. To find out, we added a proportion-matching task (Boyer & Levine, 2012) to the Wave 2 test battery and included it in an EFA carried out for Wave 2 only (see Table 6). The proportion-matching task requires children to evaluate

quantitative relations (i.e., differences in amount), but it does so through spatial comparisons and not through mathematical symbols or operations. It is worth noting that this same proportion-matching task was significantly correlated with spatial scaling in one study (Möhring, Newcombe, & Frick, 2015) and loaded onto the same factor as number line estimation in another (Ye et al., 2016), so it was difficult to say in advance whether proportion matching would be related to spatial skill, mathematics skill, or both. However, we found that like number line estimation, it loaded onto the mathematics factor. There was no evidence of a cross-domain loading onto the spatial factor for this task in any of the three grades. Note further that the inclusion of proportion matching in Wave 2 did not affect the overall factor structure or the other cross-domain loadings.

Fraction understanding

We added fraction identification items to the kindergarten and third-grade batteries for two reasons. First, we sought to determine whether fraction skill was particularly related to spatial skill in kindergarten—a question we could not pose in Wave 1 because no fraction items were given to kindergarten students. Our prediction, based on the notion that spatial reasoning is recruited when mathematics skills are novel but not completely unfamiliar, was borne out: There was not a significant cross-domain loading for fraction skill on the spatial factor in kindergarten when these new items were added. Though caution is warranted when interpreting a null finding, this result fits into a larger age-related pattern wherein algebra and spatial skills were related in sixth grade but not third grade (where algebra/missing-term problems were likely novel).

Second, we hypothesized that adding these simpler items would increase the relation between fractions and spatial skills for third-grade students, but this prediction was not borne out. We had found previously, in the regression analyses for Wave 1 and the ESEM for Wave 2,

that fractions cross-loaded onto the spatial factor when only the original items (comparing and calculating with fractions) were included. However, when the more simplified fraction identification items were added, there was no cross-loading. Thus, there is sufficient evidence to suggest a strong relation between fraction understanding and spatial skill at this age point, consistent with previous related work (Möhring et al., 2016; Ye et al., 2016), but this relation may be moderated by differences in students' familiarity with fractions and the demands of particular fraction tasks (comparison vs. matching).

Conclusions

In sum, the present study sought to replicate the factor structure obtained previously in a parallel study on the relations between spatial skills and mathematics in kindergarten, third-grade, and sixth-grade students (Mix et al., 2016). The overall pattern obtained in Wave 1—two highly correlated domain-specific factors— was clearly replicated in Wave 2. However, the more subtle effects involving cross-domain loading were only partially replicated and sometimes appeared unstable and context-specific. The inclusion of additional tasks (proportion matching and fraction identification) did not alter the overall factor structure, providing further, converging evidence that spatial skills and mathematics are closely related but unitary constructs. However, evidence of multidimensionality may be forthcoming if additional research can identify and control moderating variables, such as familiarity with mathematics content.

References

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

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

Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: Is 1/3 = 2/6 = 3/9 = 4/12? *Journal of Experimental Child Psychology, 111*, 516-533. doi:10.1016/j.jecp.2011.11.001

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, 33, 205−228. doi:10.1080/87565640801982312

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466. doi:10.1037/0033-2909.105.3.456

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697-705. doi:10.1037/0012-1649.31.4.697

Caviola, S., Mammarella, I. C., Cornoldi, C., & Lucangeli, D. (2012). The involvement of working memory in children's exact and approximate mental addition. *Journal of Experimental Child Psychology, 112,* 141-160. doi:10.1016/j.jecp.2012.02.005

Chou, C. P., & Bentler, P. M. (1995). Estimation and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks, CA: Sage.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*(2), 347-357.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cudeck, R., & O'Dell, L .L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115*, 475-487. doi:10.1037/0033-2909.115.3.475

Freedman, D. A. (2006). On the so-called 'Huber sandwich estimator' and 'robust standard errors.' *The American Statistician, 60,* 299. doi:10.1198/000313006X152207

Frick, A., Mohring, W., & Newcombe, N. (2014). Picturing perspectives: Development of perspective-taking abilities in 4- to 8-year-olds. *Frontiers in Psychology, 5,* 386. doi:10.3389/fpsyg.2014.00386. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012199/

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*, 337-353. doi:10.1006/jecp.2000.2594

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability manual* (3rd ed.).

Austin, TX: Pro-Ed.

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number line. *Developmental Psychology*, *48*, 1229-1241. doi:10.1037/a0027433

Holmes, J., Adams, J. W., & Hamilton, C. J. (2008). The relationship between visuospatial sketchpad capacity and children's mathematical skills. *European Journal of Cognitive Psychology, 20*, 272–289. doi:10.1080/09541440701612702

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

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

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

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, *29*, 745-756. doi:10.3758/BF03200477

Lakoff, G., & Núñez, R. (2000). *Where mathematics come from: How the embodied mind brings mathematics into being*. New York, NY: Basic Books. https://www.amazon.com/Where-

Mathematics-Come-Embodied-Brings/dp/0465037712

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.

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(*4), 720-733.

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*, *42*, 871-876. doi:10.3758/BRM.42.3.871

Laski, E. V., Casey, B. M., Yu, Q., Dulaney, A., Heyman, M., & Dearing, E. (2013). Spatial skills as a predictor of first grade girls' use of higher level arithmetic strategies. *Learning and Individual Differences*, *23*, 123-130. doi:10.1016/j.lindif.2012.08.001

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. doi: 10.1016/S0065-2407(08)60414-0

Lohman, D. F. (1996). Spatial ability and g. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 97-116). Hillsdale, NJ: Lawrence Erlbaum.

Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. *Current Directions in Psychological Science*, *1*, 61-66. doi:10.1111/1467-8721.ep11509746

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

Miura, I. T., Okamoto, Y., Vlahovic-Stetic, V., Kim, C. C., & Han, J. H. (1999). Language supports for children's understanding of numerical fractions: Cross-national comparisons. *Journal of Experimental Child Psychology, 74*(4), 356-365.

Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and math: Developmental and educational implications. In J. B. Benson (Ed.), *Advances in child development and behavior (Vol. 42,* pp. 201-247). New York, NY: Elsevier.

Mix, K. S., Levine, S. C., Cheng, Y. -L., 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*, 1206-1227. doi:10.1037/xge0000182

Möhring, W., Newcombe, N. S., & Frick, A. (2014). Zooming in on spatial scaling: Preschool children and adults use mental transformations to scale spaces. *Developmental Psychology*, *50*, 1614-1619. doi:10.1037/a0035905

Möhring, W., Newcombe, N. S., & Frick, A. (2015). The relation between spatial thinking and proportional reasoning in preschoolers. *Journal of Experimental Child Psychology*, *132*, 213-220. doi:10.1016/j.jecp.2015.01.005

Möhring, W., Newcombe, N. S., Levine, S. C., & Frick, A. (2016). Spatial proportional reasoning is associated with formal knowledge about fractions. *Journal of Cognition and Development*, *17*, 67-84. Doi:10.1080/15248372.2014.996289

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17. Available online at: https://www.semanticscholar.org/paper/BSEM-Measurement- https://www.semanticscholar.org/paper/BSEM-Measurement-Invariance-Analysis-Muth%E9n-Asparouhov/59ac7664b27fdf8579d23984b353ea0eb3348002

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* Washington, DC: Authors.

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

Newcombe, N. S. (2010, Summer). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, 29-43.

Newcombe, N. S., Levine, S. C., & Mix, K. S. (2015). Thinking about quantity: The intertwined development of spatial and numerical cognition. *WIREs Cognitive Science, 6*, 491-505. doi:10.1002/wcs.1369

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1-9.

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation,* 17(15), 1-8.

Paik, J. H., & Mix, K. S. (2003). US and Korean Children's Comprehension of Fraction Names: A Reexamination of Cross-National Differences. *Child Development,* 74(1), 144-154.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological*

Science, *7*, 528-530. doi:10.1177/1745691612465253.

http://journals.sagepub.com/doi/full/10.1177/1745691612465253

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*, 39–58. doi:10.1006/brcg.1995.1032

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Robert, N. D., & LeFevre, J.-A. (2013). Ending up with less: The role of working memory in solving simple subtraction problems with positive and negative answers. *Research in Mathematics Education, 15*, 165-176. doi:10.1080/14794802.2013.797748

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29,* 347-363. doi:10.1177/0734282911406661

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514. doi:/10.1007/BF02296192

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350.

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71*(1), 95-113.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-250.

doi:10.1111/1467-9280.02438

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

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General*, *142*, 193-208. doi:10.1037/a0028560

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Thompson, J. M., Nuerk, H.-C., Moeller, K., & Cohen Kadosh, R. (2013). The link between mental rotation ability and basic numerical representations. *Acta Psychologica, 144*, 324-331. doi:10.1016/j.actpsy.2013.05.009

Trbovich, P. L., & LeFevre, J.-A. (2003). Phonological and visual working memory in mental addition. *Memory & Cognition, 31*, 738-745. doi:10.3758/BF03196112

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70. doi:10.1177/109442810031002

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492. doi:10.1080/17405629.2012.686740

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2017). I. Spatial skills, their development, and their links to mathematics. *Monographs of the Society for Research in Child Development*, *82*, 7-30. doi:101111/mono.v82.1

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity.

Trends in Cognitive Sciences, *7*, 483-488. doi:10.1016/j.tics.2003.09.002

Wang, J., & Wang, X. (2012). Structural equation modeling: Applications using Mplus. Hoboken, NJ: John Wiley & Sons.

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maerlender, A. (2004). *WISC-IV: Wechsler Intelligence Scale for Children integrated technical and interpretive manual* (4th ed.). Minneapolis, MN: NCS Pearson.

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

Ye, A., Resnick, I., Hansen, N., Rodrigues, J., Rinne, L., & Jordan, N. C. (2016). Pathways to fraction learning: Numerical abilities mediate the relation between early cognitive competencies and later fraction knowledge. *Journal of Experimental Child Psychology, 152*, 242-263. doi:10.1016/j.jecp.2016.08.001

Footnotes

1. Though the eigenvalue confidence intervals fell below our accepted cut-off, the eigenvalues themselves were acceptable by conventional standards (e.g., the Guttman rule) and the fit was good (Kindergarten: RMSEA = 0.03, range = 0.00; 0.05, CFI = 0.99, SRMR = 0.03; Third Grade: RMSEA = 0.03, range = 0.00; 0.06, CFI = 0.99, SRMR = 0.02; Sixth Grade: RMSEA = 0.03, range = 0.00; 0.05, CFI = 0.99, SRMR = 0.02). Therefore, we examined the structure of the three-factor models and provide them to readers (see supplemental data, Tables S1-S3). We ultimately rejected them, however, because we deemed them uninterpretable. In kindergarten, the third factor was comprised of only one variable—place value. In third grade, it was comprised of several mathematics tasks, and two spatial measures—one of which loaded positively (VMI) and one which loaded negatively (map reading). In sixth grade, the three factor model also consisted of several mathematics tasks and one spatial task that was negatively loaded (VSWM).