



The latent structure of spatial skill: A test of the 2×2 typology

Kelly S. Mix^{a,*}, David Z. Hambrick^b, V. Rani Satyam^b, Alexander P. Burgoyne^b, Susan C. Levine^c

^a University of Maryland, United States

^b Michigan State University, United States

^c University of Chicago, United States

ARTICLE INFO

Keywords:

Spatial skill
Children
Latent structure

ABSTRACT

Multiple frameworks for categorizing spatial abilities exist but it has been difficult to verify them using exploratory factor analysis. The present study tested one of these frameworks—a 2×2 classification scheme that crossed the dimensions of static/dynamic and intrinsic/extrinsic (Uttal et al., 2013)—using confirmatory factor analysis with data on spatial performance from kindergarten ($N = 251$), third grade ($N = 246$) and sixth grade students ($N = 241$). For kindergarten and third grade students, four models were tested at each grade level: A 1-factor model, two 2-factor models (one static vs. dynamic, the other intrinsic vs. extrinsic), and a 4-factor model. In sixth grade, only the 2- and 1-factor models could be tested given the available data. Evidence that the 4-factor model was the best fit would have validated the 2×2 model. However, the 4-factor models failed to converge in kindergarten and third grade. Both the 1- and 2-factor models converged in these age groups, and chi-square tests demonstrated that the 2-factor intrinsic-extrinsic model was the best fit at both grade levels. In sixth grade, only one of the 2-factor models converged and it did not fit significantly better than the 1-factor model. Thus, there was limited validation of the model in these grades, as well as a trend toward less dimensionality in spatial skill over development.

1. Introduction

A longstanding question in cognitive psychology is whether spatial ability is a unitary construct or decomposable into distinct subfactors (Carroll, 1993; Hegarty & Waller, 2005; Lohman, 1979; Newcombe, 2018). There are obvious commonalities among spatial tasks that allow us to recognize them as spatial. Yet, there also are differences among these tasks. For example, the demands involved in locating a position on a map seem quite different than those involved in copying a form or recognizing a structure from different angles. Are these differences reflected in the cognitive structure of spatial ability?

This question has practical implications. Some have argued that training in domain general processes could have cascading effects in academic achievement by improving children's information processing capacity, efficiency, and speed (e.g., Blair & Razza, 2007), with several papers focusing on the potential benefits of spatial training in particular (Newcombe, 2010; Newcombe & Frick, 2010; Stieff & Uttal, 2015). Spatial ability is a particularly strong candidate for training because of its well established correlations with science and mathematical achievement (e.g., Hodgkiss, Gilligan, Thomas, Tolmie, & Farran, 2018; Wai, Lubinski, & Benbow, 2009), evidence of its malleability (Uttal et al., 2013), and positive effects of spatial training on performance in

science, technology, engineering, and mathematics (STEM) subjects in adults (e.g., Sorby, 2011, see Newcombe, 2018, for a review).

Still, transfer from domain general training to academic performance has been elusive for the most part (e.g., Melby-Verlag, Redick & Hulme, 2016; NRC, 2006). In the domain of spatial cognition, several studies have demonstrated improvement in specific skills and within-domain transfer (Casey et al., 2008; DeLisi & Cammarano, 1996; Terlecki, Newcombe, & Little, 2008; Wallace & Hofelich, 1992). However, attempts to demonstrate transfer to academic skills, such as mathematics, have been mixed. For example, Cheng and Mix (2014) found improvement in first graders' mathematics performance following mental rotation training, whereas Hawes, Moss, Caswell, and Poliszczuk (2015) did not. One reason for these mixed findings may be that broad domains, such as working memory, executive function, and spatial ability, represent collections of related but distinct sub-abilities rather than a single ability that is tapped by every task. If so, it is possible that only training in particular subskills would transfer to learning in school, or that the training in one subskill would transfer to some academic tasks and not others. For this reason, it is important to understand the substructure of cognitive domains, such as spatial skill.

The question of whether spatial skill is unitary or multidimensional is also important from a theoretical standpoint. Previous attempts to

* Corresponding author.

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

determine the cognitive structure of spatial ability have yielded inconsistent results. Exploratory factor analyses have indicated that spatial skill is a unitary construct (Mix, Levine, Cheng, Young et al., 2016; Mix, Levine, Cheng, Young, et al., 2017; Slater, 1940) and possibly indistinguishable from general intelligence (Smith, 1964). Other studies have found evidence of separable subfactors, but there is disagreement about the number of subfactors and what tasks should be included in each (e.g., Carroll, 1993; Höffler, 2010; Lohman, 1979; Michael, Guilford, Fruchter, & Zimmerman, 1957; Thurstone, 1944). Reviews of factor analytic studies have suggested that spatial skill may be most accurately characterized as a loose constellation of overlapping subskills (e.g., Hegarty & Waller, 2005; Miyake, Friedman, Rittenger, Shah & Hegarty, 2001; Mix & Cheng, 2012).

It is difficult to know what to conclude from these studies. One reason these factor analyses may have been inconclusive is that spatial skill does not have a stable, multi-dimensional structure and is actually best viewed as unidimensional. In a sense, arguing against multidimensionality is arguing from a null result as one can always argue that a factor analysis could have been more comprehensive or included a different balance of measures. Still, after multiple attempts to demonstrate dimensionality with mixed results, one could argue that it is parsimonious to conclude dimensionality is not present.

In contrast, some have argued that the factor analytic approach may be inconclusive because the structures it probes are unconstrained and not theory-driven (Uttal et al., 2013; Young, Levine, & Mix, 2018). Many theoretical distinctions have been delineated in the spatial literature, including categorical versus coordinate representations of space (e.g., Jager & Postma, 2003; Kosslyn, Koenig, Barrett, Cave, Tang, & Gabrieli, 1989), near versus far perceptions of space (e.g., Cowey, Small & Ellis, 1994), global versus local processing (e.g., Navon, 1977), and allocentric versus egocentric perspectives (e.g., Kesner, Farnsworth, & DiMattia, 1989), all of which derive support from behavioral and neurological evidence. Perhaps factor analyses that target these distinctions would yield more convincing evidence.

Based on this reasoning, Uttal et al. (2013) adopted Newcombe and Shipley's (2015) theory-driven framework for classifying spatial tasks to guide their meta-analysis of spatial training studies, rather than relying on the results of previous factor analytic studies. The framework is a 2×2 typology that results from crossing two dimensions that have been discussed in the psycholinguistic literature (e.g., Chatterjee, 2008; Palmer, 1978; Talmy, 2000). The static-dynamic dimension reflects the observation that spatial tasks sometimes involve objects arranged in stable positions (static) and sometimes involve objects in motion (dynamic). The intrinsic-extrinsic dimension reflects the observation that spatial relations can be inherent to an object and its parts (intrinsic), or can exist among multiple objects or between an object and its context (extrinsic). Crossing these two dimensions results in a 2×2 matrix with four task types.

The spatial cognition literature provides considerable support for the 2×2 model (see Newcombe, 2018, for a review). For example, Kozhevnikov and colleagues (Kozhevnikov, Hegarty & Mayer, 2002; Kozhevnikov, Kosslyn, & Shephard, 2005) examined individual differences in spatial visualization, and found that artists tend to excel at object visualization (or intrinsic-static processing) whereas scientists tend to excel at shifting visualization tasks (or intrinsic-dynamic processing). Evidence for the intrinsic-extrinsic distinction comes from studies such as one showing that wayfinding in adults is partially dissociable from completing paper and pencil tasks (e.g., Hegarty, Montello, Richardson, Ishikawa & Lovelace, 2006), and another showing that third grade children's performance differed on a Piagetian perspective-taking task (e.g., Huttenlocher & Presson, 1979) depending upon whether the children moved to a different view of the model (i.e., what could be considered extrinsic as it is similar to navigation), or the model itself was moved (what could be considered intrinsic as it is

similar to object manipulation) (Newcombe, 2018).

Yet such findings, while suggestive of an underlying dimensional structure, do not provide direct evidence of it. The standard approach to investigating the multidimensionality of hypothesized constructs in research on individual differences is factor analysis. This approach assumes that if there is some common cause of individual differences in multiple measures (e.g., a common process or structure), those measures should correlate significantly with each other (assuming acceptable reliability). Multidimensional scaling can also be used to investigate the nature of an individual difference construct, although a limitation of this approach is that there is no clear way to compare the fit of alternative models. One might also use an experimental approach to test for dissociations between different measures (i.e., to test whether manipulations have different effects on different measures). However, this approach should be considered a complementary rather than alternative approach to factor analysis. That is, an experimental approach focuses on mean-level differences, whereas a factor analytic approach investigates individual differences. In short, a factor analytic approach was the clear choice to address the research questions we set out to address.

In particular, we chose confirmatory factor analyses as our approach, because exploratory factor analyses lack *a priori* theoretical constraint. As Uttal et al. (2013) argued, this may be why previous factor analytic studies failed to demonstrate dimensionality in spatial processing. Confirmatory factor analysis methods may be more appropriate because they can be used to test theoretically specified models against each other and thus are well suited to detect subtle structural differences. Greater sensitivity may be particularly important given that the existing behavioral and neural evidence is not itself completely clear-cut. That is, of the two dimensions, Newcombe (2018) presented more extensive evidence for the intrinsic-extrinsic distinction than the static-dynamic distinction, and pointed out several ways in which existing research on static versus dynamic processing of objects is complicated and limited. A confirmatory approach would allow a direct test of both dimensions that may reveal structural differences not apparent in either exploratory factor analyses or the existing behavioral evidence.

Specifically, we used confirmatory factor analysis to test for the existence of intrinsic-extrinsic/static-dynamic dimensions at three age points: kindergarten, third, and sixth grade. We used an existing dataset that was part of a larger study of the relations between spatial skill and mathematics (Mix & Levine, 2018) but analyzed children's performance on only the spatial tests (i.e., mental rotation, visual-spatial working memory, figure copying, block design, map reading, perspective-taking, and proportional reasoning). Although the original study included sixth grade students, we were not able to test the full 4-factor model because there were not enough measures to provide two measures per quadrant. Specifically, because all of the map reading items for sixth graders required mental rotation, we could not divide this measure into separate static and dynamic subscores. We were able, however, to test the 1- and 2-factor models in sixth grade, the results of which we report below.

Interestingly, in a previous study using the same tasks, exploratory factor analyses including both spatial and mathematics measures failed to demonstrate substructures for spatial skill (Mix et al., 2017). Instead, all of the spatial tasks loaded onto the same factor. This finding suggested that the latent structure for spatial skill is unitary. However, as noted above, exploratory factor analyses are unconstrained by theory and thus may not detect differences in latent structure that are subtle, yet theoretically meaningful. It is also possible that in the context of an analysis with both spatial and mathematics measures, dimensionality within each domain was obscured. Because of this, we revisited the dimensionality of spatial skill in the present study using confirmatory factor analysis and examining only the spatial measures.

2. Method

2.1. Participants

A total of 738 children participated in the study. They were divided into three age groups: kindergarteners ($n = 251$, 132 boys, mean age = 6.03 years, $SD = .36$), third graders ($n = 246$, 96 boys, mean age = 9.07 years, $SD = .38$) and sixth graders ($n = 241$, 121 boys, mean age = 11.83 years, $SD = .44$). Children were recruited from 33 schools from a range of rural, suburban, and urban settings in nine communities in the Midwestern United States. Parents were contacted via their children's teachers at school, and invited to participate in the study. Only children whose parents signed an IRB-approved consent form were tested.

2.2. Procedure

As noted above, the present study is a secondary analysis of the data collected for a separate confirmatory factor analysis on spatial skill and mathematics (Mix et al., 2017). In the source study, children were tested in three 1-h sessions that took place over the course of two weeks. The tests were presented in one of three randomized, blocked orders. Within each block, the order of presentation for group versus individual tests also was randomized and counterbalanced across children. Children received a decorative folder as a reward for participation.

2.3. Measures

The proposed 2×2 typology consists of four task types (see Table 1). To test the fit of this structure to the data, we needed to identify measures for each task type. In the source study, children completed six tests commonly used to measure spatial skill in children—mental rotation, visual-spatial working memory, figure copying, block design, map reading, and perspective taking. These six tasks were distributed among the four categories as shown in Table 1, with one minor modification. That is, for kindergarten and third grade students we split the map reading items into two groups—one of which included static items and one which included dynamic items (see below). The larger battery also included one test hypothesized to be

spatial—proportional reasoning—which we included in the static-extrinsic category because it involved forming a relation between two separate entities and determining whether that relationship was the same as that shown in a target proportion.

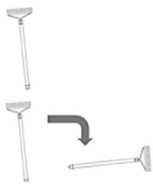



Detailed descriptions of each spatial measure follow. Reliabilities were calculated from the data unless otherwise noted. The measures had generally acceptable reliability with most reaching Cronbach's alpha $> .70$ (but see below for exceptions). Although not every measure exceeded this widely-accepted cut-off, note that one of the major advantages of structural equation modeling is that it corrects for measurement error at the level of individual measures. That is, it permits use of latent variables, which are free of random measurement error because they capture systematic variance common to a set of measures (Kline, 2005). Furthermore, simulation studies have demonstrated that confirmatory factor analyses are robust even for measures with very low internal consistency (Little, Lindenberg, & Nesselroade, 1999).

For six of the eight measures—mental rotation, VSWM, map reading:static, map reading:dynamic, perspective taking, and proportional reasoning—children's scores were the total number of items correct. Block Design and Figure Copying were scored according to their respective test manuals and raw scores were submitted for analysis. Our rationales for classification of each task in the 2×2 model are also described below. These decisions were fairly straightforward for the static-dynamic dimension, which is arguably the more obvious of the two. There is either movement or there is not. In contrast, intrinsic-extrinsic classifications seemed more open to interpretation, perhaps because none of the measures involved clear dynamic evidence provided by bodily movement through space, but rather involved manipulation of objects or paper-and-pencil tasks. This difficulty with accurate categorization is a point to which we will return in the results section.

2.3.1. Intrinsic-static tasks

Visual Spatial Working Memory (adapted from Kaufman & Kaufman, 1983). On each test trial, children saw a 14 cm \times 21.5 cm grid that was divided into squares (e.g., 3×3). Drawings of objects were displayed at random positions within the grid, though gridlines were not provided. Item difficulty was manipulated by adding divisions to the grid (up to 5×5 , again not provided to the participant) and by manipulating the number of to-be-remembered objects (up to 9). On

Table 1
Definitions, examples, and instruments for classifying spatial abilities.

Spatial skill	Definition	Example	Tasks used in the present study
Intrinsic-static	Perceiving objects, paths, or spatial configurations amid distracting background information		VSWM VMI
Intrinsic-dynamic	Piecing together objects into more complex configurations, visualizing and mentally transforming objects, often from 2-D to 3-D, or vice versa. Rotating 2-D or 3-D objects.		Block Design (Blocks) Mental Rotation (MR)
Extrinsic-static	Understanding abstract spatial principles, such as horizontal invariance or verticality.		Map Reading: Static (Map: S) Proportional Reasoning (PR)
Extrinsic-dynamic	Visualizing an environment in its entirety from a different position.		Map Reading: Dynamic (Map: D) Perspective Taking (PT)

The definitions and examples of the four spatial categories are adapted from “The malleability of spatial skills: A meta-analysis of training studies,” by D. H. Uttal et al., 2013, *Psychological Bulletin*, 139(2), p. 354–355. Copyright 2012 by American Psychological Association, with permission.

each trial, the stimulus display was left in full view for 5 s and then it was removed. Next, a blank grid with gridlines was displayed and children indicated where the drawings had appeared by marking an X in the positions that they thought had previously contained an object. Stimulus displays were presented on a laptop computer and children made their responses in individual, paper test booklets. The test was introduced with three practice items for which children received feedback on the correctness of their answers and were allowed to compare their responses to the stimulus display. There were 15–20 test trials, depending on the age of the child. Test trials began immediately after the final practice trial. The test was group administered and its reliabilities were $\alpha = .80$, $.69$, and $.81$ for kindergarten, third grade, and sixth grade, respectively.

We classified this task as intrinsic-static because it did not require imagined movement, and children could remember the locations as constellations of pictures, without analyzing the relations between the pictures themselves. Also, similar tasks used by D'Amico (2011) and Mullin (2006) were classified as intrinsic-static by Uttal et al. (2013). However, one could argue that the VSWM task could involve analysis of relations among the pictures even if that is not required. If so, the task could be considered extrinsic-static. Because classification along the intrinsic-extrinsic dimension was less clear-cut, we will report the results of our analyses both ways.

Figure Copying (Test of Visual-Motor Integration, or 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 $.93$. The figure copying task was considered intrinsic-static because it did not involve imagined movement and it focused mostly on perceiving and copying objects within rather than among objects.

2.3.2. Intrinsic-dynamic tasks

Mental Rotation (adapted from Neuburger, Jansen, Heil, & Quaiser-Pohl, 2011 and Peters, Laeng, Latham, Jackson, et al., 1995). Two variations of Vandenberg and Kuse's (1978) mental rotation task were used. In the kindergarten and third grade version (Novack, Brooks, Kennedy, Levine, & Goldin-Meadow, 2013), small groups of children were shown four unfamiliar figures (i.e., forms based on manipulating components of capital letters so they were no longer recognizable as 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 presented on a laptop screen. Children received feedback on the correctness of their performance, and also were shown animations with the correct answers rotating to match the target. Following the practice session, children completed the 16 test items in a paper booklet (kindergarten: $\alpha = .77$; third grade: $\alpha = .86$). 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 ($\alpha = .83$). The task was classified as intrinsic-dynamic because it focuses on an individual object rotating through space. The same task also was classified by Uttal et al. (2013) as intrinsic-dynamic.

Block design (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. Items ranged in difficulty and

children completed different numbers of items depending on their basal and ceiling performance. The reliability coefficient reported in the WISC-IV manual for the Block Design subtest is between $.83$ and $.87$ depending on age group. Like mental rotation, this task involves individual objects rotating through space so it was classified as intrinsic-dynamic. The same task also was classified by Uttal et al. (2013) as intrinsic-dynamic.

2.3.3. Extrinsic-static tasks

Map Reading: Static (adapted from Liben & Downs, 1989). Working one-on-one with the experimenter, children completed 10 items in which they were shown a location on a model and then indicated where it would appear in a corresponding map. The model was a full color 3-dimensional model town with buildings, roads, a river, and trees. It measured 10 in. \times 10 in. in area and the tallest structure was $.50$ in. high. Children marked the corresponding locations on a black and white, 2-dimensional, scale map (6 in. \times 6 in.). Item difficulty was manipulated by varying the scale ratio of the map (1:1, 1:2.5) based on previous research (i.e., Boyer & Levine, 2012; Vasilyeva & Huttenlocher, 2004). The items were ordered from easiest to most difficult based on the results of pilot-testing, and feedback was given on the first three test questions to ensure that children understood the task. Reliability was $\alpha = .57$ in kindergarten and $\alpha = .47$ in third grade. In sixth grade, there were no static map reading items (i.e., every item required rotation). As we will see, this prevented us from testing the 4-factor model in sixth grade. We considered static map reading items extrinsic because answering correctly required children to analyze relations among objects within and across two scenes. These items were considered to be static because the model and map were in the same orientation and thus, did not require imagined movement to respond.

Proportional Reasoning (adapted from Boyer & Levine, 2012). Children were shown a stimulus display on a laptop computer that depicted three columns with different proportions of red space versus blue—a standard and two choices. Next to the standard, there was a picture of pig who was introduced as “Harry the Hog.” Children were told, “Harry 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 were tested in groups, but circled their responses in individual, paper test copies. There were 20–24 test trials depending on grade. On each trial, the target appeared on the left side of the screen and the two response choices on the right side. Two spatial arrangements of the correct answer relative to the foil (i.e., above or below) were counterbalanced across trials. The reliabilities were $\alpha = .70$ for kindergarten, $.90$ for third grade, and $.57$ for sixth grade. We considered proportion matching extrinsic because, like the map reading task, children had to compare within and across two displays to respond. As for VSWM, this classification was not entirely clear-cut, however, as the task also required children to process the ratio of full to empty containers and this could be considered an intrinsic task. Because of this ambiguity, we will present our results with both classifications. In terms of static-dynamic, the task clearly did not require imagined movement and so it was classified as static.

2.3.4. Extrinsic-dynamic tasks

Map Reading: Dynamic (adapted from Liben & Downs, 1989). For kindergarten and third grade participants, this task was a continuation of the Map Reading: Static task, but consisted of 4 items for which the map differed from the model by 180 degrees of rotation. Reliabilities were $\alpha = .48$ in kindergarten and $\alpha = .55$ in third grade. In sixth

Table 2a
Descriptive statistics and correlations for spatial tasks for kindergarten ($N = 251$).

	Mean	SD	1	2	3	4	5	6	7	
<i>Intrinsic-static</i>										
1	Figure Copying (VMI)	16.35	1.79							
2	VSWM	8.82	3.38	.32**						
<i>Extrinsic-static</i>										
3	Map Reading: Static (Map: S)	5.49	1.59	.30**	.48**					
4	Proportional Reasoning (PR)	14.20	4.05	.24**	.10	.13*				
<i>Intrinsic-dynamic</i>										
5	Mental Rotation (MR)	4.37	3.22	.20**	.34**	.26**	.22**			
6	Block Design (Blocks)	14.07	8.35	.40**	.45**	.48**	.17**	.41**		
<i>Extrinsic-dynamic</i>										
7	Map Reading: Dynamic (Map: D)	0.31	0.66	.23**	.29**	.41**	.15*	.16*	.44**	
8	Perspective Taking (PT)	11.35	3.26	.18**	.31**	.31**	.11	.18**	.29**	.37**

* $p < .05$.
** $p < .01$.

grade, there were 8 map reading items and all were dynamic. Sixth grade children completed the task in small groups ($\alpha = .55$). As for map reading:static items, we considered map reading:dynamic items extrinsic because children had to analyze relations among objects within and across two scenes. However, we considered these items to be dynamic because the map and the model were rotated to be in different orientations and, thus, required imagined movement to align.

Perspective Taking (Frick, Mohring, & Newcombe, 2014). The perspective taking task required children to imagine a scene from different points of view. In the kindergarten/third grade version, children were shown a set of Play Mobil figures in a particular arrangement. Then, they were shown four pictures and asked to indicate which picture was taken from each character's perspective. Items varied in difficulty based on the number of objects in the pictures and the angles of view. The 27 test questions were preceded by 4 practice items with feedback (kindergarten $\alpha = .69$; third grade $\alpha = .87$). Sixth grade children saw six to eight objects arranged in a circle. They were asked to imagine standing next to one object while directly facing another object, and then draw an arrow toward a third object to indicate their angle of view from this perspective (Kozhevnikov & Hegarty, 2001). There were two 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 ($\alpha = .82$). In all three grades, children were tested individually. The perspective-taking task was considered extrinsic because participants answered questions about relative placements of objects, and dynamic because answering correctly required children to mentally rotate the scene to match their own viewing perspective.

3. Results

Descriptive statistics and correlations for the measures are presented in Tables 2a–2c, where it can be seen that scores on most of the measures were significantly correlated. The only exceptions involved proportional reasoning. Specifically, though proportional reasoning was significantly correlated with most of the spatial measures in all three grades, it was not significantly correlated with VSWM or perspective-taking in kindergarten, nor was it significantly correlated with either map reading:dynamic in third grade or VSWM in sixth grade. Still, the overall pattern of significant bivariate correlations suggests there was considerable overlap among the eight measures. To formally investigate the factor structure underlying these correlations, we next conducted a series of confirmatory factor analyses (CFAs).

Table 2b
Descriptive statistics and correlations for spatial tasks in third grade ($N = 246$).

	Mean	SD	1	2	3	4	5	6	7	
<i>Intrinsic-static</i>										
1	Figure Copying (VMI)	20.83	2.94							
2	VSWM	10.22	2.50	.42**						
<i>Extrinsic-static</i>										
3	Map Reading: Static (Map S)	7.27	1.31	.35**	.37**					
4	Proportional Reasoning (PR)	19.66	4.85	.22**	.29**	.16*				
<i>Intrinsic-dynamic</i>										
5	Mental Rotation (MR)	9.13	4.26	.40**	.36**	.36**	.21**			
6	Block Design (Block)	26.71	11.10	.49**	.52**	.47**	.23**	.53**		
<i>Extrinsic-dynamic</i>										
7	Map Reading: Dynamic (Map D)	1.45	1.20	.22**	.31**	.51**	.09	.24**	.40**	
8	Perspective Taking (PT)	17.71	5.37	.39**	.35**	.36**	.20**	.46**	.45**	.35**

* $p < .05$.
** $p < .01$.

Table 2c
Descriptive statistics and correlations for spatial tasks in sixth grade ($N = 241$).

		Mean	SD	1	2	3	4	5	6
<i>Intrinsic-static</i>									
1	Figure Copying (VMI)	23.02	3.05						
2	VSWM	6.84	4.25	.41**					
<i>Extrinsic-static</i>									
3	Proportional Reasoning (PR)	10.73	3.19	.26**	.10				
<i>Intrinsic-dynamic</i>									
4	Mental Rotation (MR)	4.71	3.25	.36**	.35**	.15*			
5	Block Design (Block)	34.37	11.36	.50**	.55**	.21**	.41**		
<i>Extrinsic-dynamic</i>									
6	Map Reading (Map)	−35.77	17.73	.41**	.34**	.18**	.37**	.43**	
7	Perspective Taking (PT)	−75.00	30.78	.48**	.44**	.25**	.53**	.48**	.37**

* $p < .05$.

** $p < .01$.

3.1. Confirmatory factor analyses

We first tested the fit of a single factor model which reflects a unitary structure for spatial tasks. We then tested two 2-factor models to investigate the existence of the hypothesized dimensions (i.e., Intrinsic/Extrinsic and Static/Dynamic). Finally, in kindergarten and third grade only, we tested the full, 4-factor model that results from crossing the two proposed dimensions (i.e., Intrinsic-Static, Intrinsic-Dynamic, Extrinsic-Static, and Extrinsic-Dynamic) at each grade level. Recall that for sixth grade, we could not test the 4-factor model because there were no map reading:static items; thus we did not have at least two measures per cell.

All reported analyses used maximum likelihood estimation (MLR) with robust standard errors in Mplus 7.1 to guard against non-normal distribution for some measures. Specifically, there were non-normal distributions in kindergarten for mental rotation, VSWM, Block Design, map reading:dynamic, and perspective-taking, in third grade for map reading:dynamic and proportional reasoning, and in sixth grade for perspective-taking, VSWM, and mental rotation. 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). Simulation studies have demonstrated that this approach is effective for distributions ranging in skewedness from -2 to 2 degrees (Chou & Bentler, 1995). The distribution of scores used in the present study fell within this range.

3.1.1. Fit statistics

To evaluate model fit, we used several commonly reported statistics, including Chi Square Test of Model Fit (χ^2), comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) (Jackson, Gillaspay, & Purc-Stephenson, 2009; Suhr, 2006). The χ^2 (chi-square) statistic compares the hypothesized and observed covariance matrices. A χ^2 statistic closer to 0 indicates a better fit between the specified model and the observed data. The null hypothesis is that there is no difference between the specified model and the observed data, so contrary to most testing, it is desirable for the null hypothesis to not be rejected. In other words, for this test, a p -value greater than or equal to .05 indicates good fit. However, given a moderate sample size, even slight differences between the hypothesized and observed covariance matrices can result in significant χ^2 values (Bearden, Sharma, & Teel, 1982). Therefore, we evaluated the models in terms of the full pattern of fit statistics. The comparative fit index (CFI) reflects the improvement of model fit over a baseline in which covariances are zero; CFI values range from 0 to 1, with values greater than or equal to 0.95 generally taken to indicate

acceptable fit (Hu & Bentler, 1999). The Root Mean Square Error of Approximation (RMSEA) divides estimated model error by its degrees of freedom and adjusts for sample size (Steiger, 1990). Because RMSEA estimates the “badness of fit,” lower RMSEA values are better. A generally accepted cut-off is .08 (Browne & Cudeck, 1993; Hu & Bentler, 1999; Steiger, 1989). Standardized Root Mean Residual (SRMR) compares an observed correlation to a model-implied correlation matrix and estimates the difference by averaging the absolute values of the correlation residuals. Like RMSEA, lower SRMR values indicate better fit and an SRMR of 0.08 or less is generally considered acceptable (Kline, 2005). Note that although such cut-offs have been debated, the risk of false rejection declines with sample sizes greater than 200 (Chen, Curran, Bollen, Kirby, & Paxton, 2008), as we achieved for each age group in the current study.

3.1.2. Choosing the most parsimonious model

We compared the fit of alternative models using the chi-square (χ^2) difference test on pairs of null models in order to find the most parsimonious for each grade. Pairwise comparisons of model fit are required because there are multiple theoretically plausible models (Werner & Schermelleh-Engel, 2010). Models were compared in order from most to least restrictive. Least restrictive (i.e., reduced) models were those with the fewest factors, most degrees of freedom, and fewest free parameters.

Typically, two models are compared by finding the difference in the χ^2 statistic between the null and alternative model and using the difference in degrees of freedom to determine significance. If the χ^2 difference is significant, then the alternative model is assumed to have a better fit to the data. If the χ^2 difference is not significant, then both models are considered to fit equally well and the “smaller,” more parsimonious model (the null hypothesis) is chosen. However, because we used MLR estimation to correct non-normality in a few variables, this method is not valid because the χ^2 difference is not itself a chi-square distribution. We therefore calculated the Santorra-Bentler scaled chi-square difference test to compare the fits of models, as recommended by Muthén and Muthén (2005), with a critical alpha level of .05 and degrees of freedom derived from the difference between the degrees of freedom of the null and alternative model.

3.1.3. Findings

In kindergarten and third grade, both the 1- and 2-factor models converged. As shown in Table 3, in general, all of these models had acceptable fit. We used χ^2 tests to determine whether the 2-factor models fit better than the 1-factor models in kindergarten and third grade. The Intrinsic/Extrinsic 2-factor model fit better than the 1-factor

Table 3
Goodness-of-fit indicators for confirmatory factor analysis models by grade.

Model	χ^2	df	χ^2 p-value	CFI	RMSEA	SRMR	χ^2 Diff	χ^2 Diff p-value
<i>Kindergarten</i>								
1-factor	41.59	20	.003	.94	.066	.044	–	–
2-factor (Intrinsic-Extrinsic)	36.24	19	.010	.95	.060	.042	7.37*	.007
2-factor (Static-Dynamic)	41.14	19	.002	.94	.068	.043	0.93	.34
<i>3rd Grade</i>								
1-factor	47.13	20	.001	.94	.074	.044	–	–
2-factor (Intrinsic-Extrinsic)	41.01	19	.002	.95	.069	.043	5.07*	.02
2-factor (Static-Dynamic)	47.01	19	.000	.94	.077	.044	0.24	.62
<i>6th Grade</i>								
1-factor	51.24	14	.000	.91	.105	.055	–	–
2-factor (Static-Dynamic)	51.27	13	.000	.91	.111	.054	0.002	.963

Note. The 4-factor models for kindergarten and third grade were non-admissible solutions, so they are not included. The 2-factor (Intrinsic, Extrinsic) model for sixth grade yielded an inadmissible solution, so it is not included. The 4-factor model was not tested in sixth grade because there were not enough measures. Dashes (–) indicate that the row is the least nested model.

* $p < .05$.

model in these grades; the Static/Dynamic 2-factor model did not. The factor loadings for the Intrinsic-Extrinsic model at each grade level are presented in Figs. 1a and 1b.¹ Perhaps not surprising, given that the Static/Dynamic 2-factor model did not fit better than the 1-factor model, the 4-factor model failed to yield an admissible solution in either kindergarten or third grade. Inspection of the inter-factor correlations revealed that half of them (i.e., 3 out of 6 at each grade level) approached or exceeded |1| (Kindergarten: range = .70 to 1.19, median = .94; Third grade: range = .83–1.20, median = .97), suggesting a lack of dimensionality in the 4-factor model (Geiser, 2012).

A different pattern of results emerged in the sixth-grade sample. The Intrinsic/Extrinsic model failed to yield an admissible solution, probably due to a high inter-factor correlation ($r = 1.11$). Although both the 1-factor model and the Static/Dynamic 2-factor model provided convergent and admissible solutions, the difference in fit between the two models was not significant (see Table 3). Thus, we conclude that the 1-factor model fit best in this grade (See Fig. 1c). However, two of the fit statistics did not reach an acceptable level for the 1-factor model, suggesting the model was not particularly robust. This finding was unexpected given that Mix et al. (2016) had reported a unidimensional structure for spatial performance in a similar study that included sixth grade students and many of the same measures.

One change that might explain this discrepancy is the addition of the proportional reasoning task. In Mix et al.’s (2016) exploratory factor analysis, this task was not included, and in the original analysis of the Wave 2 data (Mix et al., 2017), only the combined factor structure of spatial and mathematical tasks was investigated. Children’s performance on spatial tasks alone was not analyzed separately. Because proportional reasoning loaded significantly onto the mathematics factor in sixth grade and not the spatial factor, it might have diminished

¹ One concern with our measures might be the possibility of correlated errors on the map reading task (Brown, 2015). Recall that we derived separate scores from essentially the same task based on whether or not the map was rotated with respect to the model on specific trials, so it is possible that correlated error variances for these two measures could affect the results. We examined whether children’s errors were correlated on the two map reading tasks and determined that the errors were significantly correlated in third grade ($r = .33$), but not kindergarten ($r = .09$). To follow up, we repeated our third grade analyses with the errors on map reading:static and map reading:dynamic correlated. As before, we found that only the 1- and 2-factor models converged and all three models provided acceptable fit, but unlike the results with uncorrelated errors, neither of the 2-factors models fit better than the 1-factor model. An inspection of the loadings showed that when the errors were correlated, the factor loadings for both map reading tasks were lower. This suggests that the correlation may have drawn some of the variance away from their common, extrinsic factor.

model fit when it was included here. To test this possibility, we repeated the same analysis with proportional reasoning removed, but this change did not improve the fit of the 1-factor model (CFI = .904; RMSEA = .114; SRMR = .057). We also examined the factor loadings of the spatial tasks and noticed that mental rotation was the weakest indicator in sixth grade. This was surprising given moderate loadings for mental rotation in a previous factor analysis of spatial performance in sixth grade students (Mix et al., 2016). However, when mental rotation was removed from the 1-factor model, the fit indices reached acceptable levels (CFI = .975, RMSEA = .066, SRMR = .031, $\chi^2 = 18.48$, $df = 9$).

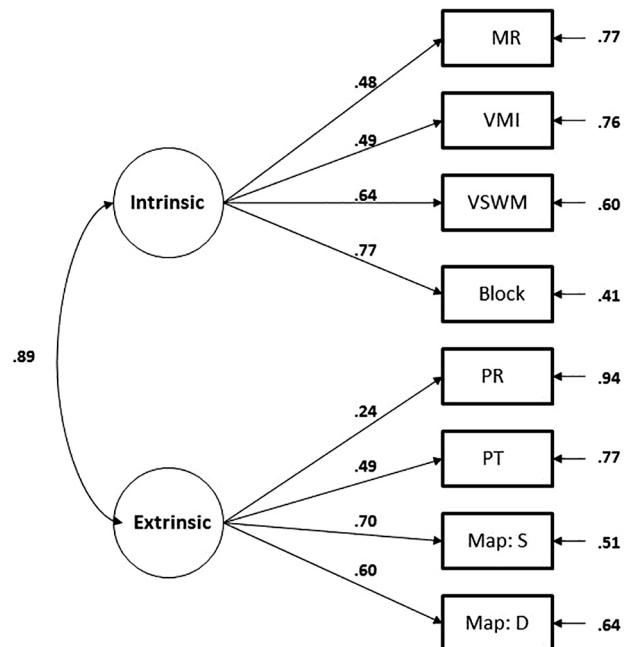


Fig. 1a. The 2-factor (Intrinsic-Extrinsic) model was the most parsimonious model for kindergarten. The one-sided arrows from factors to measures are standardized factor loadings. All of these loadings were significant based on z-values derived by dividing the factor loading for each measure by its standard error. Only tasks with z-values greater than 1.96 were considered significant ($p = .05$). Short arrows that point to the left are the residual variances. Abbreviations: Map: S = Map Reading: Static, Map: D = Map Reading: Dynamic, MR = Mental Rotation, PR = Proportional Reasoning, Block = Block Design, PT = Perspective Taking.

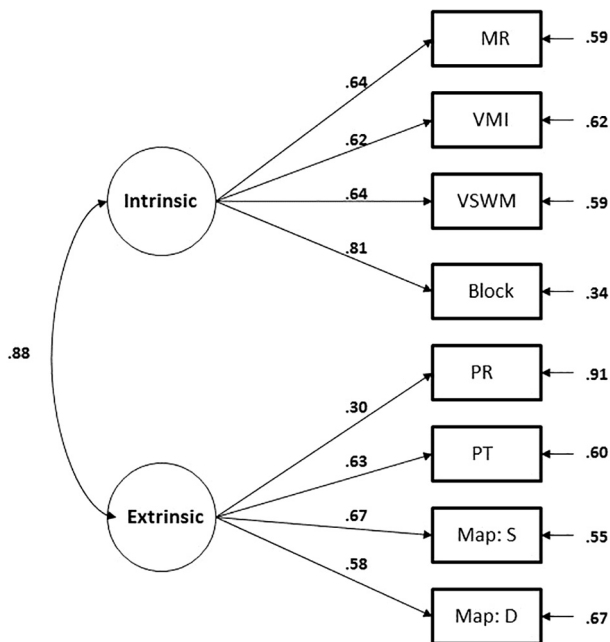


Fig. 1b. The 2-factor (Intrinsic-Extrinsic) model was the most parsimonious model for third grade. The one-sided arrows from factors to measures are standardized factor loadings. All of these loadings were significant based on z-values derived by dividing the factor loading for each measure by its standard error. Only tasks with z-values greater than 1.96 were considered significant ($p = .05$). Short arrows that point to the left are the residual variances. Abbreviations: Map: S = Map Reading: Static, Map: D = Map Reading: Dynamic, MR = Mental Rotation, PR = Proportional Reasoning, Block = Block Design, PT = Perspective Taking.

We next assessed whether the results in all three grades changed if we reclassified two of the tasks. Recall that the classification of VSWM and proportional reasoning along the Intrinsic/Extrinsic dimension was debatable. When we reassigned these tasks such that VSWM was classified as extrinsic and proportional reasoning was classified as intrinsic, we obtained slightly different patterns of results. In kindergarten, the 4-factor model now converged but did not fit significantly better than the 2-factor models. As before, the Intrinsic/Extrinsic 2-factor model fit best. In third grade, the 4-factor model failed to converge as it had before, but neither of the 2-factor models fit significantly better than the 1-factor model (see Table 4). The sixth-grade results remained the same as before. Thus, reassigning these two tasks impacted the results inasmuch as the previously significant improvement in model fit did not reach significance in both grades; however, this reclassification did not reveal alternative patterns of significance (e.g., new evidence for the static-dynamic dimension was not revealed).

4. Discussion

The present study tested a theoretical model for classifying spatial tasks (Newcombe & Shipley, 2015; Uttal et al., 2013) that was based on well-known dimensions of spatial skill studied since the 1970s. Our analysis used data from a larger study of spatial skill and mathematics that has been published separately (Mix et al., 2017). We evaluated up to four nested CFA models in kindergarten, third and sixth grade students. The models included (1) a 1-factor “general spatial” model; (2) a 2-factor model that tested only the intrinsic/extrinsic dimension; (3) a 2-factor model that tested only the static/dynamic dimension; and (4) a 4-factor model that reflected all four quadrants of the 2×2 typology (Intrinsic-Static, Intrinsic-Dynamic, Extrinsic-Static, Extrinsic-Dynamic). Based on the proposed theory-driven framework, we hypothesized that the 4-factor model would provide the best fit.

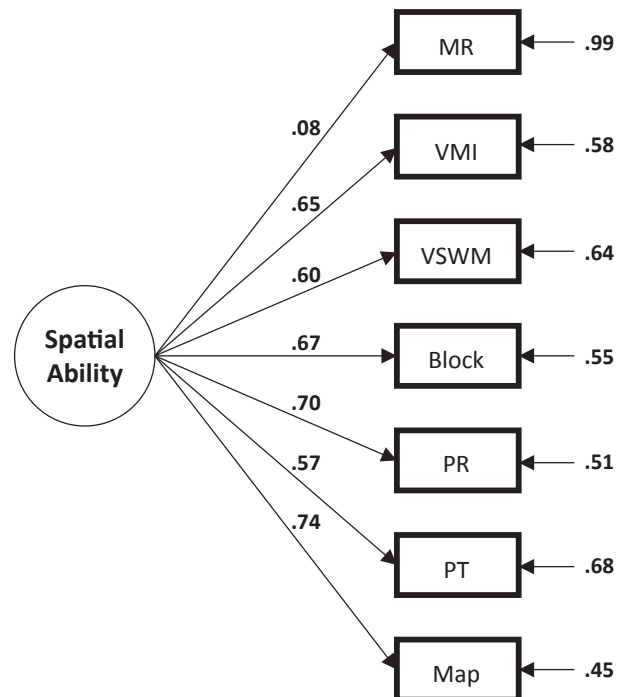


Fig. 1c. The 1-factor model was the most parsimonious model for sixth grade. The one-sided arrows from factors to measures are standardized factor loadings. All of these loadings except mental rotation were significant based on z-values derived by dividing the factor loading for each measure by its standard error. Only tasks with z-values greater than 1.96 were considered significant ($p = .05$). Short arrows that point to the left are the residual variances. Abbreviations: MR = Mental Rotation, Block = Block Design, Map = Map Reading, PR = Proportional Reasoning, PT = Perspective Taking.

Our results provided mixed evidence regarding dimensionality within the spatial measures. In all but one case, the 4-factor model failed to converge in either kindergarten or third grade, and even in that exceptional case, the 4-factor model did not provide a better fit than the 2-factor model based on intrinsic-extrinsic processing. This suggests that the 2×2 typology may not be an accurate characterization of the latent structure underlying spatial performance, at least in elementary school aged children.

In kindergarten and third grade, the two 2-factor models consistently converged, and one of these models—intrinsic-extrinsic—generally provided a better fit to the data than the 1-factor model. This finding stands in contrast to the unidimensionality reported in a previous exploratory factor analysis that used a closely related dataset (Mix et al., 2016) and suggests spatial skill may have a multi-dimensional structure, but perhaps not one that is based on static-dynamic processing. This structure might be limited to intrinsic-extrinsic processing, or might include other dimensions that were not tested in this model, such as categorical versus coordinate representations of space (e.g., Kosslyn et al., 1989), global versus local processing (e.g., Navon, 1977), and allocentric versus egocentric (e.g., Kesner et al., 1989).

One motivation for testing the 2×2 model was to better understand how spatial skill might relate to performance in STEM fields. If there are separable dimensions of spatial skill, these dimensions might relate differently to mathematics in general, or to specific mathematics tasks (e.g., algebra). In light of the present results, one might ask how intrinsic-extrinsic processing relates to STEM performance in the early elementary age range. Intrinsic processing involves noticing spatial relations within an object, and individual differences in this skill could translate into individual differences in symbol differentiation and symbol reading in early mathematics. For example, strong intrinsic

Table 4
Goodness-of-fit indicators for confirmatory factor analysis models with VSWM and proportional reasoning reassigned, by grade.

Model	χ^2	df	χ^2 p-value	CFI	RMSEA	SRMR	χ^2 Diff	χ^2 Diff p-value
<i>Kindergarten</i>								
1-factor	41.59	20	.003	.94	.066	.044	–	–
2-factor (Intrinsic/Extrinsic)	32.69	19	.026	.96	.054	.039	11.19*	< .001
2-factor (Static/Dynamic)	41.14	19	.002	.94	.068	.043	0.93	.34
4-factor	22.59	14	.067	.98	.049	.031	9.77	.08
<i>3rd Grade</i>								
1-factor	47.13	20	.001	.94	.074	.044	–	–
2-factor (Intrinsic/Extrinsic)	45.78	19	.001	.94	.076	.043	1.68	.19
2-factor (Static/Dynamic)	47.01	19	.000	.94	.077	.044	0.24	.62
<i>6th Grade</i>								
1-factor	51.24	14	.000	.91	.105	.055	–	–
2-factor (Static/Dynamic)	51.27	13	.000	.91	.111	.054	0.002	.963

Note. VSWM was reassigned to extrinsic-static and proportional reasoning was reassigned to intrinsic-static. The 4-factor model in third grade was not an admissible solution, so it is not represented in this table. The 2-factor (Intrinsic, Extrinsic) model for sixth grade yielded an inadmissible solution, so it is not included. Dashes (–) indicate that the row is the reduced model.

* $p < .05$.

spatial processing may help children discriminate between written numerals, such as 8 versus 3. Intrinsic processing may also be important for differentiating geometric shapes. In contrast, extrinsic processing involves noticing relations among objects, or between objects and their immediate environments. This skill may be integral to reading equations and reading multi-digit numerals, for which relative position carries most of the meaning (e.g., 27 versus 72). Extrinsic processing may also help children ground the meaning of small numbers, by representing quantities as a list ordered in space, as on a spatial representation like the number line. One intriguing possibility is that mathematical performance is supported by the ability to move flexibly between intrinsic and extrinsic spatial processing. For example, children solving multi-digit addition problems must both recognize individual digits (intrinsic) and interpret the place value meaning of the digits and track their place in the equation as they carry from one place to the other (extrinsic). Being able to rapidly switch between processing types may be particularly helpful in such cases.

The present results contribute to scientific understanding of the dimensionality of spatial performance, but for two major reasons, they do not provide a definitive test of the 2×2 typology. First, it is important to note that our results are based on data from children, whereas most research indicating dimensionality in spatial skill has focused on adolescents and adults. Perhaps spatial skills become more differentiated with development. If so, the static-dynamic dimension may become evident later in life, in which case the 2×2 typology might be supported. Our findings with sixth grade students seem to argue against this possibility, as we did not observe developmental change on a trajectory toward more differentiation but rather, the reverse. Indeed, assuming the lack of dimensionality was not due to our limited measures, one could characterize our findings as a pattern of gradually weakening multi-dimensionality from kindergarten to sixth grade. However, it is possible that further research with adolescents and adults would reveal a non-linear developmental trend in which more differentiated structures become evident in adulthood.

Another reason why the present study should not be considered a definitive test of the 2×2 typology is that it was a secondary analysis of an existing dataset that was not specifically designed to test this typology. As such, it offered only the minimum number of measures needed to test the 4-factor model (i.e., eight tasks, or two per quadrant) in kindergarten and third grade, and not enough measures to test the 4-factor model in sixth grade. Perhaps with a more extensive set of measures, a different pattern would have emerged. That stated, even when we had 4 tasks per dimension in the 2-factor models, we found no evidence for the static-dynamic dimension. Without confirmatory evidence for the static-dynamic dimension, it would be impossible to

obtain confirmatory evidence for the 4-factor model, so it seems unlikely that simply adding measures would change the overall outcome.

A third, more minor concern is that a few of the measures had relatively low reliability. Although latent variables are free of measurement error, it is possible that factor loadings for these measures would increase with higher reliability. This would not have changed the overall pattern of results, however.

Our failure to find support for the static-dynamic dimension is, in itself, rather surprising because this categorization seems straightforward—that is, the presence of movement or transformation is easily judged. Yet grouping tasks based on this criterion did not seem to mirror the latent structure underlying performance. This finding points to a potential problem inherent to models like the 2×2 typology—namely, the complexity involved in categorizing spatial tasks (Newcombe, 2018). We have already discussed the ambiguity involved in categorizing tasks such as VSWM and proportional reasoning in terms of the intrinsic-extrinsic dimension. This difficulty could also extend to other tasks as well. For example, although the block design task involves constructing a design within a single frame, the design is composed of intrinsic elements that need to be put together properly to make the whole. Thus, one could argue that constructing the design rests on extrinsic relations among elements.

Additionally, there are other, less observable complications that could introduce noise. With respect to the static-dynamic dimension, research indicates that dynamic processing depends on the quality of the representation of the object that is transformed. This pattern is apparent on mental rotation tasks, where the representation of the object being rotated can affect mental rotation performance (e.g., Folk & Luce, 1987; Heil & Jansen-Osmann, 2008; Xu & Franconeri, 2015). More generally, a task might fit multiple categories because different strategies for attacking a task engage different processing, leading to individual differences in the categorization itself. For people who encode displays in the VSWM task as gestalt patterns, the VSWM task would be intrinsic, but for people who encode the same displays as collections of individual objects, the task would be extrinsic. Both strategies are likely sufficient to support accurate responses so it is not obvious how the task could be categorized accurately.

Further, it is possible for different spatial skills to become engaged at different stages of processing, even in the same task and the same people. In the Shepherd-Metzler cube task, for example, people must first encode the figure in order to imagine it moving. The encoding stage could be considered intrinsic-static whereas the imagined movement stage would be intrinsic-dynamic. Others have questioned these distinctions on similar grounds, pointing to the variance introduced by different contexts, measurement approaches, and order of processing

(e.g., parallel vs. sequential) (Burgess, 2006; Lourenco & Longo, 2009). In short, such typologies, even theory-driven and empirically-based typologies, may not be capable of accurately capturing all the complexities of spatial thought in a given task.

In summary, we used confirmatory factor analysis to test the fit of a theory-driven model for distinguishing among spatial tasks in three age groups. We did not find evidence to support the static-dynamic dimension at any age, and consequently did not find evidence for the overall 2×2 model. We did find evidence for the intrinsic-extrinsic dimension in younger children, thus lending support to the general idea of dimensionality in spatial processing and to the specific proposal that spatial skills may be distinguished along the lines of intrinsic and extrinsic relations. Further research with adults and perhaps a more extensive array of measures may provide additional insight into these structures, but a fundamental problem to overcome is the complexity inherent in categorizing spatial tasks.

5. Author note

We are grateful to all of the children and school personnel who participated in this study, including those from the Alsip-Hazlgreen-Oaklawn Public Schools, Bath Public Schools, Chicago Public Schools, Dewitt Public Schools, Eaton Rapids Public Schools, Hazelgreen School, Holt Public Schools, Independence Schools, Lane Technical High School, Potterville Public Schools, Resurrection School, St. Malachy Catholic School, St. Martha's School, St. Michael Parish School, University of Chicago Charter School Donoghue Campus, Valley View Community Unit Public Schools, and community programs in Chicago. This research was supported by a generous grant to the first and last authors from the Institute of Education Sciences (#R305A120416). The opinions and positions expressed in this report are the authors' and do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

References

- Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research*, 425–430.
- 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.
- Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78(2), 647–663.
- Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: $1s/3 = 2/6 = 3/9 = 4/12$? *Journal of Experimental Child Psychology*, 111(3), 516–533. <https://doi.org/10.1016/j.jecp.2011.11.001>.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10, 551–557. <https://doi.org/10.1016/j.tics.2006.10.005>.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
- Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., & Copley, J. (2008). The development of spatial skills through interventions involving block building activities. *Cognition and Instruction*, 26(3), 269–309. <https://doi.org/10.1080/0737000802177177>.
- Chatterjee, A. (2008). The neural organization of spatial thought and language. *Seminars in Speech and Language*, 29(3), 226–238. <https://doi.org/10.1055/s-0028-1082886>.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36(4), 462–494. <https://doi.org/10.1177/0049124108314720>.
- 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>.
- 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.
- Cowey, A., Small, M., & Ellis, S. (1994). Left visuo-spatial neglect can be worse in far than in near space. *Neuropsychologia*, 32, 1059–1066. [https://doi.org/10.1016/0028-3932\(94\)90152-X](https://doi.org/10.1016/0028-3932(94)90152-X).
- D'Amico, A. (2011). The assessment and training of working memory for prevention and early intervention in case of reading, writing, and arithmetical difficulties in children. In Eden S. Levin (Ed.), *Working memory: Capacity, developments, and improvement techniques* (pp. 181–248). New York, NY: Nova Science.
- DeLisi, R., & Cammarano, D. M. (1996). Computer experience and gender differences in undergraduate mental rotation performance. *Computers in Human Behavior*, 12, 351–361.
- Folk, M. D., & Luce, D. (1987). Effects of stimulus complexity on mental rotation rate of polygons. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3), 395–404.
- Freedman, D. A. (2006). On the so-called 'Huber sandwich estimator' and 'robust standard errors'. *The American Statistician*, 60, 299–302. <https://doi.org/10.1198/000313006X152207>.
- Frick, A., Möhring, W., & Newcombe, N. S. (2014). Development of mental transformation abilities. *Trends in Cognitive Sciences*, 18(10), 536–542. <https://doi.org/10.1016/j.tics.2014.05.011>.
- Geiser, C. (2012). *Data analysis with MPlus*. New York, NY: Guilford Press.
- Hawes, Z., Moss, J., Caswell, B., & Poliszczuk, D. (2015). Effects of mental rotation training on children's spatial and mathematical performance: A randomized controlled study. *Trends in Neuroscience and Education*, 4(3), 60–68. <https://doi.org/10.1016/j.tine.2015.05.001>.
- Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2), 151–176. <https://doi.org/10.1016/j.intell.2005.09.005>.
- Hegarty, M., & Waller, D. (2005). Individual differences in spatial abilities. In P. Shah, & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). Cambridge University Press.
- Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation of polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *The Quarterly Journal of Experimental Psychology*, 61(5), 683–689. <https://doi.org/10.1080/17470210701822967>.
- Hodgkiss, A., Gilligan, K. A., Thomas, M. S. C., Tolmie, A. K., & Farran, E. K. (2018). Spatial cognition and science: The role of intrinsic and extrinsic spatial skills from seven to eleven years. *British Journal of Educational Psychology*, 88(1), 538–543. <https://doi.org/10.1111/bjep.12211>.
- Höfler, T. N. (2010). Spatial ability: Its influence on learning with visualizations—A meta-analytic review. *Educational Psychology Review*, 22(3), 245–269. <https://doi.org/10.1007/s10648-010-9126-7>.
- 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), 1–55.
- Huttenlocher, J., & Presson, C. C. (1979). The coding and transformation of spatial information. *Cognitive Psychology*, 11(3), 375–394.
- Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R., Jr. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14(1), 6–23. <https://doi.org/10.1037/a0014694>.
- Jager, G., & Postma, A. (2003). On the hemispheric specialization for categorical and coordinate spatial relations: A review of the current evidence. *Neuropsychologia*, 41(4), 504–515.
- Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman assessment battery for children (K-ABC)*. Circle Pines, MN: Wiley.
- Kesner, R. P., Farnsworth, G., & DiMattia, B. V. (1989). Double dissociation of egocentric and allocentric space following medial prefrontal and parietal cortex lesions in the rat. *Behavioral Neuroscience*, 103, 956–961. <https://doi.org/10.1037/0735-7044.103.5.956>.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.
- Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., & Gabrieli, J. D. E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 723–735. <https://doi.org/10.1037/0096-1523.15.4.723>.
- Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29, 745–756. <https://doi.org/10.3758/BF03200477>.
- Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (2002). Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. *Cognition and Instruction*, 20(1), 47–77.
- Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition*, 33(4), 710–726.
- 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.
- Little, T. D., Lindenberger, U., & Nesselrode, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, 4(2), 192–211.
- Lohman, D. F. (1979). Spatial ability: A review and reanalysis of the correlational literature. (Technical Report No. 8). Stanford, CA: Aptitude Research Project, School of Education, Stanford University.
- Lourenco, S. F., & Longo, M. R. (2009). The plasticity of near space: Evidence for contraction. *Cognition*, 112, 451–456. <https://doi.org/10.1016/j.cognition.2009.05.011>.
- Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science*, 11(4),

- 512–534. <https://doi.org/10.1177/1745691616635612>.
- Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. (1957). The description of spatial-visualization abilities. *Educational and Psychological Measurement*, 17(2), 185–199.
- Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and math: Developmental and educational implications. *Advances in Child Development and Behavior*, 42, 197–243. <https://doi.org/10.1016/B978-0-12-394388-0.00006-X>.
- [dataset] Mix, K. S. & Levine, S. C., Space and Math Wave One & Wave Two (v.1), Digital Repository at the University of Maryland (DRUM), 2018, <http://hdl.handle.net/1903/21090>.
- 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(9), 1206–1227.
- 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: A replication of the two-factor model. *Journal of Cognition and Development*, 18(4), 465–492.
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, 130(4), 621–640. <https://doi.org/10.1037/0096-3445.130.4.621>.
- Mullin, L. N. (2006). Interactive navigational learning in a virtual environment: Cognitive, physical, attentional, and visual components (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3214688).
- Muthén, L. K., & Muthén, B. O. (2005). Chi-square difference testing using the Santorra-Bentler scaled chi-square. Retrieved from. <https://www.statmodel.com/chidiff.shtml>.
- Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9, 353–383. [https://doi.org/10.1016/0010-0285\(77\)90012-3](https://doi.org/10.1016/0010-0285(77)90012-3).
- National Research Council (2006). *Learning to think spatially: GIS as a support system in the K-12 curriculum*. Washington, DC: National Academies Press.
- 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. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, Summer, 29–43.
- Newcombe, N.S. (2018). Three kinds of spatial cognition. In J. Wixted (Ed.), Stevens' handbook of experimental psychology and cognitive neuroscience (4th ed.) (pp. 521–552).
- Newcombe, N. S., & Frick, A. (2010). Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, 4(3), 102–111.
- Newcombe, N. S., & Shipley, T. F. (2015). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (Ed.) Studying visual and spatial reasoning for design creativity (pp. 179–192). Springer Netherlands. doi: 10.1007/978-94-017-9297-4_10.
- Novack, M., Brooks, N., Kennedy, D., Levine, S., & Goldin-Meadow, S. (October, 2013). Using action and gesture to improve mental rotation. Poster presented at the Cognitive Development Society Conference, Memphis, TN.
- Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch, & B. B. Lloyd (Eds.). *Cognition and categorization* (pp. 259–303). Hillsdale, NJ: Erlbaum.
- 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.
- Slater, P. (1940). Some group tests of spatial judgment or practical ability. *Occupational Psychology*, 14, 40–55.
- Smith, I. M. (1964). *Spatial ability: Its educational and social significance*. San Diego, CA: Knapp.
- Sorby, S. A. (2011). *Developing spatial thinking workbook*. Independence, KY: Cengage Learning.
- Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180. https://doi.org/10.1207/s15327906mbr2502_4.
- Stieff, M., & Uttal, D. (2015). How much can spatial training improve STEM achievement? *Educational Psychology Review*, 27(4), 607–615. <https://doi.org/10.1007/s10648-015-9304-8>.
- Suhr, D. (2006). Exploratory or confirmatory factor analysis? Statistics and Data Analysis, 1–17. Cary, NC: SAS Institute. doi: 10.1002/da.20406.
- Talmy, L. (2000). *Toward a cognitive semantics: Vol. 1. Concept structuring systems*. Cambridge, Massachusetts: MIT Press.
- Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, 22(7), 996–1013. <https://doi.org/10.1002/acp.1420>.
- Thurstone, L. L. (1944). *A factorial study of perception*. University of Chicago Press.
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, 139(2), 352–402. <https://doi.org/10.1037/a0028446>.
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47(2), 599–604.
- Vasilyeva, M., & Huttenlocher, J. (2004). Early development of scaling ability. *Developmental Psychology*, 40, 682–690.
- 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>.
- Wallace, B., & Hofelich, B. G. (1992). Process generalization and the prediction of performance on mental imagery tasks. *Memory & Cognition*, 20(6), 695–704.
- 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.). San Antonio, TX: NCS Pearson.
- Werner, C., & Schermelleh-Engel, K. (2010). Deciding between competing models: Chi-square difference tests. Introduction to Structural Equation Modelling with LISREL.
- Xu, Y., & Franconeri, S. (2015). Capacity for visual features in mental rotation. *Psychological Science*, 26(8), 1241–1251. <https://doi.org/10.1177/0956797615585002>.
- Young, C. J., Levine, S. C., & Mix, K. S. (2018). The connection between spatial and mathematical ability across development. In H.-C. Nuerk, K. Cipora, F. Domahs, & M. Haman (Eds.). *Special issue: On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations*. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2018.00755.