Different Analyses, Different Conclusions? Validity Evidence From the EGMA Spatial Reasoning Subtask

Lindsey Perry
Southern Methodist University

Abstract

As the global development community shifts its focus from improving access to education to improving learning and instruction, the need for instruments that accurately measure student achievement in mathematics and meet technical standards is increasing. This paper explores the importance of collecting high-quality validity evidence that aligns with an instrument's intended uses and interpretations by discussing a new subtask developed for the Early Grade Mathematics Assessment (EGMA). The EGMA Spatial Reasoning subtask was developed by RTI International with funding from the United States Agency for International Development (USAID). To collect validity evidence to support the assumption that the EGMA Spatial Reasoning subtask could be used to determine overall student proficiency in spatial reasoning, the items developed for the subtask were pilot tested with 1,426 students in Jordan. Pilot test data was initially analyzed using Item Response Theory. However, Item Response Theory assumptions were not met, thus, supplemental analyses were conducted using Classical Test Theory. There were differences in the findings using the two different methods, which impacts the interpretations made using this instrument. This paper illustrates the importance of choosing analytic techniques that align with an instrument's intended use in order to make valid interpretations from the data to inform policy and practice.

Keywords

Early Grade Mathematics Assessment (EGMA), validity evidence, test instrument intended uses, test results interpretation

Introduction

One mathematical topic consistently identified as being foundational for future success is spatial reasoning (Learning Metrics Task Force [LMTF], 2013; NCTM, 2000; National Research Council [NRC], 2009). The LMTF (2013) identified spatial reasoning as essential content for all children, including those at the early childhood level, and the NRC (2009) identified spatial reasoning as part of two core mathematical topics for young children. Spatial reasoning helps students investigate and navigate their own environment by enabling them to visualize objects and locations from different perspectives and orientations. One

Corresponding Author:
Lindsey Perry, Research in Mathematics Education,
Simmons School of Education and Human Development
Southern Methodist University, PO Box 750114
Dallas, TX 75275-0114
Email: leperry@mail.smu.edu

Citation: Perry, Lindsey (2018). Different analyses, different conclusions? : Validity evidence from the EGMA spatial reasoning subtask. *Global Education Review*, *5* (3), 125-142.

reason spatial reasoning is often identified as being critical for students to learn is that students' ability to reason spatially is highly predictive of overall mathematics achievement in the short- and long-term (Gilligan, Flouri, & Farran, 2017; Markey, 2009; Robinson, Abbott, Berninger, & Busse, 1996). Spatial reasoning also supports number sense (van Nes & de Lange, 2007) and the development of problem solving (Battista, 1990; Hegarty & Kozhevnikov, 1999; van Garderen, 2006). Additionally, spatial reasoning is crucial for many careers, such as engineering (Olkun, 2003) and medicine (Allahyar & Hunt, 2003).

However, while spatial reasoning is widely cited as being important for young children, few assessments exist that primarily focus on spatial reasoning, and most do not assess all aspects of spatial reasoning. Furthermore, the assessments that do exist have not been used in low-income contexts. As the global development community shifts its focus from improving access to education to improving learning and instruction, the need for instruments that accurately measure student achievement in mathematics and meet technical standards is increasing.

To fill this assessment gap, RTI International developed a Spatial Reasoning subtask for the Early Grade Mathematics Assessment (EGMA) and piloted the Spatial Reasoning items in Jordan. The purpose of this paper is two-fold: 1) to present information on this experimental measure and the results from the pilot testing, and 2) to discuss how differences in item analysis techniques may lead to different validity conclusions. For the first purpose, one primary research question with two sub-questions was proposed: Is the Spatial Reasoning subtask a technically adequate measure that reliably estimates students' spatial reasoning abilities? A) Does the two-parameter item response theory model fit the Spatial Reasoning pilot test data with acceptable item parameters and fit statistics? B) Is the reliability of the data generated by the Spatial Reasoning subtask sufficient for the intended interpretation of the subtask? During the investigation into these research questions, analyses from both Item Response Theory (IRT) and Classical Test Theory (CRT) frameworks were conducted. Therefore, in addition to providing evidence for these specific research questions, the second purpose of this paper was to draw attention to the differences between item analysis techniques, the potential differences in interpretations made using these techniques, and the implications of using each method. Information about spatial reasoning, the EGMA Spatial Reasoning subtask, and validity are included to provide pertinent background and theoretical foundations for these research questions.

Spatial Reasoning

Spatial abilities are an important component of aptitude and have been researched and investigated for over 100 years (Galton, 1883; Thurstone & Thurstone, 1941). Spatial reasoning moves beyond geometric ideas of shape and properties, as theorized by van Hiele (1983), and instead focuses on understanding the complexity of one's environment. While spatial tasks require many different skills, the general consensus of researchers is that spatial reasoning consists of two factors: spatial visualization and spatial orientation (Bishop, 1980; McGee, 1979; NRC, 2009; Sarama & Clements, 2009). The distinction between these two factors lies in what is being moved. With spatial visualization, a person mentally visualizes and transforms an image; the visualized image is being moved mentally and the person's perspective remains stationary. With spatial orientation, a person mentally views environments from different perspectives; in this case, the environment remains stationary and the person mentally moves their own position around the environment. In general, spatial reasoning focuses on mentally transforming objects or seeing these objects from different perspectives. The EGMA Spatial Reasoning subtask primarily focuses on spatial visualization skills. This section provides additional information on the construct of spatial visualization and details when spatial visualization develops in young children.

Spatial Visualization

Spatial visualization is the ability to transform figures mentally (McGee, 1979). This is a complex mental process that first requires a person to visualize a static image (Clements, 2004; Kosslyn, 1983), which is similar to a photo (i.e., the objects can be seen but can't be moved). The image must be maintained and held in the person's mind, and then the image can be moved mentally to rotate or transform it in some way. Spatial visualization can be used to determine if two figures are congruent or to determine if there are hidden parts of three-dimensional figures. Spatial visualization has not been as heavily researched as spatial orientation (Sarama & Clements, 2009), primarily because it is difficult to observe spatial visualization.

Spatial visualization develops early in life. At first, the images children imagine mentally are static. However, around the age of 4, children demonstrate the ability to rotate objects mentally (Frick, Hansen, & Newcombe, 2013; Marmor, 1975). Then, at the age of 5, children can translate and reflect images (Sarama & Clements, 2009). Children can rotate images with increasingly complex angles (e.g., 45°) at age 6 and can perform diagonal translations by age 7. While these spatial visualization abilities appear and develop in early childhood, these abilities continue to improve through adolescence (Ben-Chaim, Lappan, & Houang, 1988) and into adulthood.

Spatial Structuring

One way children demonstrate spatial visualization skills is through spatial structuring tasks. Spatial structuring is "the mental act of constructing an organization or form for an object or set of objects" (Battista & Clements, 1996). Spatial structuring entails systematically organizing objects into component parts so that the objects can be enumerated more easily. For example, the volume of a rectangular prism can be calculated by first finding the area of the base and then adding it repeatedly based on the height. This systematic process prevents counting errors that would occur if a child attempted to calculate the volume by counting cubes randomly. Spatial structuring is often used to find the area or volume of two- or three-dimensional figures, respectively, and to utilize spatial structuring, one typically visualizes the object and transforms it mentally to determine the component parts.

EGMA Spatial Reasoning Subtask

In 2013, RTI International began the development of a Spatial Reasoning subtask for the EGMA. The EGMA Spatial Reasoning subtask was designed to complement the existing Core EGMA. The Core EGMA is a set of eight numeracy subtasks: Number Identification, Quantity Discrimination, Missing Number, Addition – Level 1, Subtraction – Level 1, Addition – Level 2, Subtraction – Level 2, and Word Problems. The EGMA is an orally and individually administered assessment. The EGMA was first developed in 2008 by RTI International with funding from USAID and is primarily administered to children in grades 1-3 to make summative decisions, including determination of overall student performance and to program evaluation (Platas, Ketterlin-Geller, Brombacher, & Sitabkhan, 2014). The EGMA has been used in over 25 countries worldwide; however, it is not meant to be used to compare performance across countries.

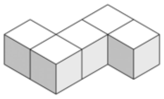
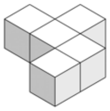
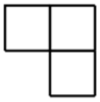
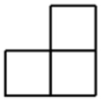
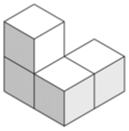
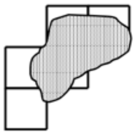
A panel consisting of early mathematics and assessment experts recommended adding a subtask to the EGMA to test spatial reasoning (Platas et al., 2014). To begin development, a literature review was conducted to define the construct and to determine how spatial reasoning had previously been tested. Preliminary items were developed, and multiple rounds of pre-pilot testing and cognitive interviews with students occurred. Information from these interviews informed the final development of the items.

The final development of items for the EGMA Spatial Reasoning focused on spatial visualization and spatial structuring of two- and three-dimensional figures. Table 1 shows the types of items developed. The spatial visualization items required children to look at two figures to determine if the figures were the same or not the same. Children responded by answering, “the same” or “not the same.” To create a range of item difficulties, the number of cubes/squares, the types of transformations, and the angles of rotation varied. The spatial structuring items asked children to determine the number of cubes or squares needed to create a figure. Children responded with a numerical value. To vary the difficulty in the spatial structuring items, cubes/squares were hidden or covered. After the final development, the items were translated from English to Arabic and pilot tested in Jordan.

An expert review was conducted to receive feedback on the items after the pilot test was conducted (Perry, 2017). Four experts in early mathematics education and assessment provided feedback on the items. Overall, the experts rated all of the items as mostly to extremely age-appropriate and representative and relevant to the construct of spatial reasoning. However, the experts did raise a few important issues regarding the words used in the items and the ambiguity of some of the graphics. More specifically, reviewers noted that “the same” and “not the same” could be interpreted differently by students, which may impact how students respond to items. Additionally, reviewers noted that the potential for hidden cubes could confuse some students on some of the three-dimensional spatial structuring items. These comments can be used during future refinement of the items. Even with these comments, however, the experts rated the items highly, noting all items were age-appropriate and representative and relevant to spatial reasoning.

Table 1

Items Developed for the EGMA Spatial Reasoning Subtask

	Item Type	Item Prompt Read by Test Assessor	Sample Items	Number of Items Developed for Pilot Testing
Spatial Visualization	Three-dimensional	Look at these pictures of objects. Please tell me if the two objects are the same or not the same.	 	17
	Two-dimensional	Look at these pictures of objects. Please tell me if the two objects are the same or not the same.	 	17
Spatial Structuring	Three-dimensional	Look at these pictures of objects. Please tell me how many cubes were used to make this object.		17
	Two-dimensional	Look at these pictures of shapes made with squares. Some of the squares are covered. How many squares were used to make this shape?		13 (4 items without squares covered; 9 items with some squares covered)

Validity

The primary purpose of this paper is to present validity evidence collected for the EGMA Spatial Reasoning subtask and to investigate how different analyses may impact interpretations about validity. Therefore, it is critical to discuss the purpose of validity evidence and how evidence impacts interpretations.

Before the EGMA Spatial Reasoning subtask can be used, the validity of the interpretations made from the assessment must be evaluated. Validity is the most important factor to consider when designing or evaluating tests (AERA, APA, NCME, 2014). The validity evidence collected must be tailored to the intended uses or interpretations of the assessment, and the International Test Commission noted that tests should be used "only for those purposes where relevant and appropriate validity evidence is available" (ITC, 2001, p. 12). Validity is not a property of the test but instead is a property of the interpretation being made using the test scores. Therefore, before collecting validity evidence, the intended uses and interpretations of the EGMA Spatial Reasoning subtask must be identified.

The primary uses of the EGMA are to determine overall student performance and to evaluate programs. To evaluate the validity of these uses, Kane's argument-based approach to validating interpretations (1992, 2013) was used. In this approach, assumptions or propositions are identified that link a score to its interpretation. Then, the types of evidence that can be collected to test those assumptions are identified. For this paper, only one assumption is being tested: the subtask is an accurate measure that reliably estimates students' spatial reasoning abilities. IRT models and reliability estimates were proposed to provide evidence for this assumption. For the full interpretation-use argument for the EGMA Spatial Reasoning subtask, including all identified assumptions, please see Perry (2016).

Methods

To collect validity evidence regarding technical adequacy and reliability, the EGMA Spatial Reasoning items were pilot tested in Jordan. A non-equivalent groups with anchor test (NEAT) design (Holland & Dorans, 2006) was used to allow for item equating. The 64 items were divided among four pilot test forms; four items were used as anchor items and were included on all four forms. Each form had 19 items.

Participants

A total of 1,426 students in Grades 2-3 participated in the pilot test. Table 2 provides additional details about the participants. The participants were enrolled in schools that were selected as part of another RTI International project. Stratified sampling was used to select the schools; additional details about the sampling methods used can be found in Brombacher et al. (2014). One second and one third grade class were selected at random from each school, and 10 students from each class were selected at random to participate in the pilot test. As seen in Table 2, numbers of males and females and Grade 2 and Grade 3 students were approximately equal. Most students were between the ages of 7-9.

Administration

Trained test assessors administered the EGMA Spatial Reasoning subtask to each selected student. Item stimulus sheets were used to show students the figures for each item, and iPads were used by the test assessor to reference the script for each item and to record student responses. The test assessor administered two

Table 2
Pilot Test Participants for the EGMA Spatial Reasoning Subtask

	Form A	Form B	Form C	Form D	Total
Number of students	340	348	369	369	1426
Male	152	139	176	195	662
Female	188	209	193	174	764
Urban	207	275	223	197	902
Rural	133	73	146	172	524
Grade 2	177	170	184	186	717
Grade 3	163	178	185	183	709
Age 6	0	1	0	0	1
7	53	28	49	58	188
8	157	176	156	147	636
9	121	128	149	146	544
10	8	14	13	17	52
11	1	1	2	0	4
12	0	0	0	1	1

sample items before each section to familiarize students with the directions and with the format of the items. Tangible objects were used in the sample items to help students understand the tasks. However, the actual items used two-dimensional representations of the figure as seen in Table 1.

Analyses

To collect validity evidence for technical adequacy and reliability, multiple analyses were proposed.

Research Question A

To answer the first sub-question – Does the two-parameter item response theory model fit the Spatial Reasoning pilot test data with acceptable item parameters and fit statistics? – the pilot test data was analyzed using item response theory (IRT). One of the benefits of using IRT instead of Classical Test Theory (CTT) is that IRT models are sample invariant, meaning that the item statistics do not depend on the sample used in the pilot testing. This is particularly pertinent for the EGMA since it is used at a large-scale and in many different contexts. The two-parameter IRT model was proposed in order to examine both item difficulty and item discrimination.

***IRT Assumptions*.** Before applying IRT models, two strong assumptions must be checked: unidimensionality and local independence (Hambleton, Swaminathan, & Rogers, 1991). Unidimensionality refers to the idea that items test a single construct. Local independence refers to the idea that responses on items must only depend on a student's ability with the latent trait (Embretson & Reise, 2000). Local independence is calculated and checked within the IRT framework.

To test for unidimensionality, an exploratory factor analysis was conducted for each form of the Spatial Reasoning pilot test. The analyses were conducted in R using the Psych package (Revelle, 2015). Item loadings above 0.32 were considered acceptable (Tabachnick & Fidell, 2001). The oblique method of rotation was used, and model fit statistics, including the Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA) were calculated. Acceptable fit statistics include TLI values of greater than 0.95 and RMSEA values of 0.06 or less (Hu & Bentler, 1999). A Parallel Analysis was also conducted in R for each pilot test form. A Parallel Analysis tests if actual eigenvalues differ from random data and is considered to be superior to Scree tests or the examination of eigenvalues (Ledesma & Valero-Mora, 2007). The model fit statistics and Parallel Analysis plots were used to assess dimensionality.

Because of the Exploratory Factor Analysis (EFA) results, IRT modeling was not conducted; thus, local independence was not evaluated.

Research Question B

To collect evidence for the second sub-question – Is the reliability of the data generated by the Spatial Reasoning subtask sufficient for the intended interpretation of the subtask? – internal consistency estimates were calculated. Cronbach's alpha was calculated in R for each pilot test form. The level of reliability needed is dependent on the types of decisions being made using the data. Kline suggested that all assessments have Cronbach's alpha values above 0.7, and ideal values are over 0.9 (Kline, 2000). However, if the decisions being made are low-stakes, Cronbach's alpha values can be $0.7 < \alpha < 0.9$. Since the decisions made using the EGMA data are not high-stakes (e.g., do not impact student placement decisions, teacher ratings), an alpha value greater than 0.7 is desired.

Supplemental Tests. Because of the results from Research Question A and B, additional supplemental analyses that align most closely with CTT were conducted. The percent of correct responses (i.e., p-values) and the item-total correlations for each item were calculated. Item-total correlations represent the correlation between the scores on an item and the overall score for the subtask. Item-total correlations between 0.3-0.7 are considered acceptable (Ferketich, 1991).

Results

Research Question A

The results from the EFAs and Parallel Analysis plots indicate that the EGMA Spatial Reasoning subtask is not unidimensional. Figure 1 shows the Parallel Analysis plots for the EGMA Spatial Reasoning subtask by pilot test form. These plots indicate that there are multiple factors for each Spatial Reasoning pilot test form. The line representing the actual data should only separate from the simulated and resampled data in one place to represent a one factor solution. These plots indicate that there are at least three factors on each of the pilot test forms.

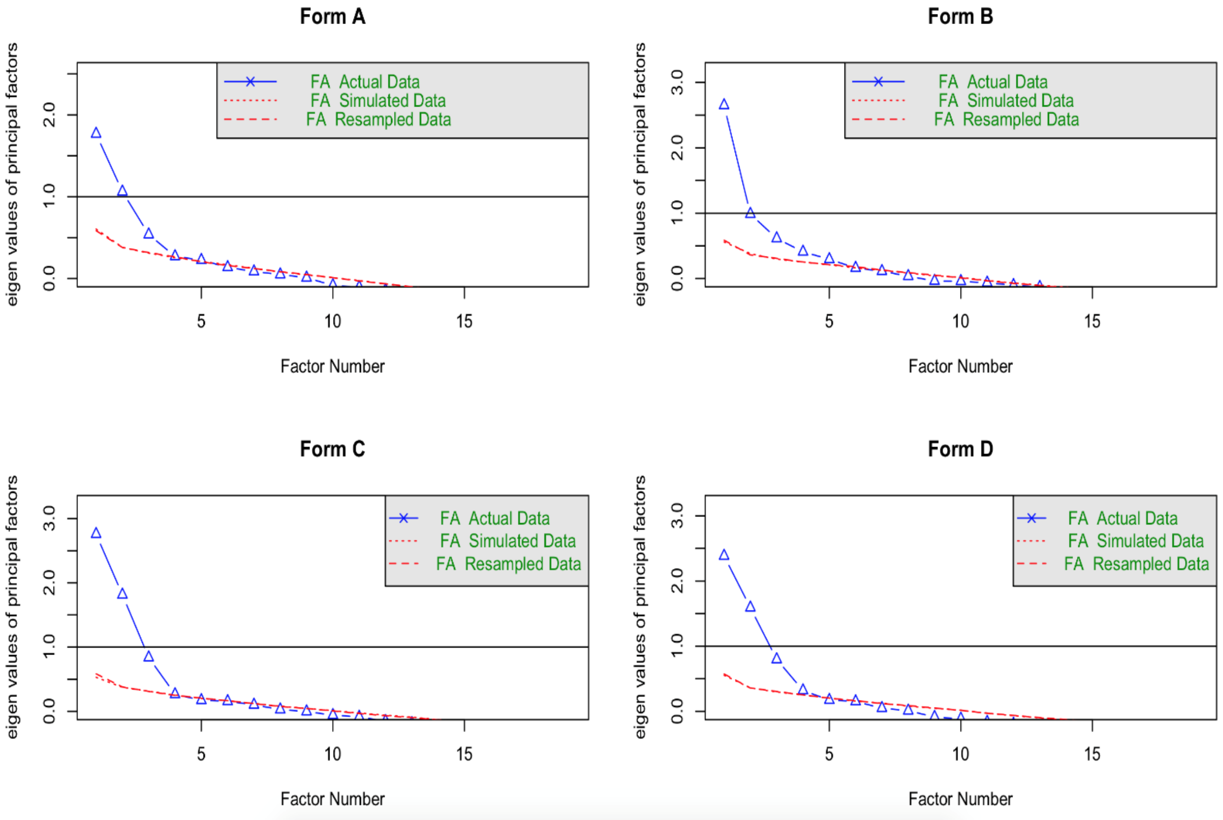


Figure 1. Parallel Analysis plots for the EGMA Spatial Reasoning subtask by pilot test form.

The results from the EFAs are summarized in Table 3. One-factor, two-factor, three-factor, and four-factor solutions were conducted. With the one-factor solution for each pilot test form, only 3, 4, 3, and 6 out of 19 items, respectively, have a factor loading greater than 0.32. As the number of factors increases, the number of items with an acceptable factor loading on one of the factors also increases. The only instance when this does not occur is for Form D; the three-factor solution has a greater number of items with acceptable factor loadings than the four-factor solution. Additionally, model fit statistics also indicate that the one-factor solution is not optimal; the TLI and the RMSEA are outside of the bounds for desired model fit as suggested by Hu and Bentler (1999) (i.e., TLI ≥ 0.95, RMSEA ≥ 0.06). As additional factors are included in the model, the TLI and RMSEA fall within the acceptable ranges for some forms.

Taken together, these results indicate that the Spatial Reasoning subtask is not unidimensional. However, a clear factor solution is not present. While the four-factor solution appears to have acceptable model fit statistics, except in Form C, and, in general, has the most items with acceptable factor loadings, the items do not consistently load on factors based on common characteristics or a theoretical rationale. For example, there are four types of Spatial Reasoning items. Therefore, it might be assumed that each distinct type of item would load on a single factor. However, within the four-factor solution, items within one item type

Table 3
Single Factor Exploratory Factor Analysis Results for Spatial Reasoning by Pilot Test Form

Form	Factor Solution	Number of Items	Number of Items with a Single Loading Greater than 0.32	Number of Items with all Loadings Less than 0.32	TLI	RMSEA
A	1	19	3	16	0.491	0.065
	2	19	8	11	0.836	0.038
	3	19	10	9	0.945	0.023
	4	19	11	8	0.994	0.012
B	1	19	4	15	0.612	0.080
	2	19	9	10	0.718	0.069
	3	19	10	9	0.801	0.058
	4	19	12	7	0.853	0.050
C	1	19	3	16	0.391	0.126
	2	19	6	13	0.631	0.098
	3	19	8	11	0.787	0.075
	4	19	8	11	0.808	0.071
D	1	19	6	13	0.346	0.107
	2	19	12	7	0.674	0.076
	3	19	16	3	0.821	0.056
	4	19	14	5	0.888	0.045

load on different factors, and the factors include items from multiple item types. For example, within the four-factor solution for Form D, one of the factors includes three-dimensional spatial visualization, two-dimensional spatial visualization, and two-dimensional spatial structuring items. One of the other factors also includes items from all of these item types. From a different perspective, the two-dimensional spatial visualization items, for example, have acceptable loadings on three different factors. Therefore, while the four-factor solution has acceptable model fit statistics for most forms and the greatest number of items with acceptable factor loadings, a clear theoretical rationale does not support this factor structure.

Therefore, since these results indicate that the Spatial Reasoning subtask is not unidimensional, IRT analyses were not conducted since unidimensionality is a critical assumption for IRT modeling.

Research Question 2B

The internal consistency statistic, Cronbach's alpha, for each Spatial Reasoning pilot test form can be seen in Table 4; these values range from 0.25-0.61. None of these values indicate sufficient internal consistency (α > 0.7).

Supplemental Tests

Because of the results from Research Questions A and B, p-values and item-total correlations were calculated to investigate these items further and can be seen in Table 5.

Table 4
Cronbach's Alpha by Spatial Reasoning Pilot Test Form

Form	Internal Consistency
Form A	0.25
Form B	0.35
Form C	0.17
Form D	0.61

Table 5
Percent correct and item-total correlations for EGMA Spatial Reasoning items

	Percent Correct	Item-total Correlation		
	Mean (*SD*)	Mean (*SD*)	Minimum	Maximum
All items	73% (24%)	0.30 (0.15)	-0.26	0.59
3D Spatial Visualization	79% (22%)	0.29 (0.15)	0.05	0.59
2D Spatial Visualization	55% (24%)	0.23 (0.19)	-0.26	0.45
3D Spatial Structuring	75% (21%)	0.38 (0.10)	0.17	0.51
2D Spatial Structuring	86% (15%)	0.30 (0.12)	0.01	0.54

The percent correct values are very typical of mathematics assessments, with some items being answered correctly more and less often. The mean item-total correlations waver between traditionally acceptable (i.e., > 0.30) and unacceptable values (i.e., < 0.30). The standard deviations of the item-total correlations are also indicative of a wide range of values. Many items were in the unacceptable range for item-total correlations.

Discussion

The results from the planned IRT analyses indicated that the Spatial Reasoning subtask is not unidimensional. However, since this was surprising, additional traditional item analyses, such as calculating p-values and item-total correlations, were conducted during supplemental testing. The intended purpose of these supplemental tests was to better understand why the subtask was not unidimensional and why items were not correlating well with one another; the analyses were expanded to include methods used in CTT to allow for a comparison of the results between the two methods (i.e., IRT and CTT). In fact, some of the results from the CTT analyses seemed contrary to what was expected based on the IRT results. In this section, three questions will be addressed:

(1) Do the results from the two methods of analysis lead to different conclusions about the EGMA Spatial Reasoning subtask?

(2) Why is the method of item analysis critical to consider when making interpretations about results?

(3) What are the implications of the results from this study?

Discussion Question 1

Do the results from the two methods of analysis lead to different conclusions about the EGMA Spatial Reasoning subtask? If examined separately, the EFAs conducted in preparation for IRT modeling and the traditional CTT analyses conducted as supplemental analyses may lead to different conclusions. For example, the EFAs and Parallel Analysis plots indicate that the Spatial Reasoning forms were not unidimensional, and a clear factor structure was not found. This conclusion essentially halts IRT analyses and requires a deeper examination into the items and potential refinement of the construct and/or items. However, only looking at the p-values for the items may lead users to believe that the items are performing relatively well. These interpretations are detailed in the paragraphs that follow.

The results from the EFAs and Parallel Analysis plots suggest that the Spatial Reasoning subtask is not unidimensional. The Parallel Analysis plots and EFAs for the Spatial Reasoning pilot test forms suggest that multiple latent factors exist on the subtask, which is problematic for unidimensional IRT analyses. While the four factor model included the greatest number of items with acceptable factor loadings and the most acceptable fit statistics, there is no theoretical rationale to support how the items load on the factors in this model. For example, items within one item type load on different factors. Factors also include items from multiple item types. Theoretically, since items within an item type are extremely similar and assess the same latent construct, they should, for the most part, load on the same factor. Therefore, there is not a clear factor structure for the Spatial Reasoning subtask based on the EFA analyses. Multiple latent traits are being measured on the Spatial Reasoning subtask. Because unidimensionality is a critical assumption for IRT and the Spatial Reasoning subtask is not unidimensional, it was not possible to continue with the IRT analyses. Taken together, these results do not provide validity evidence for the technical adequacy and reliability assumption.

However, many of the traditional CTT statistics, such as p-values and item-total correlations (see Table 5), were in the typical range of acceptable values. Except for the two-dimensional spatial visualization items, all other item types had mean p-values greater than 70%. If a user just examined p-values of items, many of which were greater than 90%, it could appear that students understand and perform relatively well on the Spatial Reasoning items. This conclusion may be supported by the item-total correlations. The mean item-total correlations were all slightly above or below the acceptable value of 0.30. Many were above the acceptable cut-off and many were below. However, as noted in Table 4, the internal consistency estimates ranged from 0.17 to 0.61, indicating poor internal consistency. The internal consistency estimates, which are typically computed within a CTT framework, were not in the acceptable range and support the conclusion from the EFA analyses that the Spatial Reasoning subtask assesses more than one latent construct. While both IRT and CTT analyses on the whole are not indicative of technical adequacy and reliability, if certain analyses were examined in isolation, different, albeit unfounded, interpretations may exist.

Discussion Question 2

Why is the method of item analysis critical to consider when making interpretations about results? The method of item analysis impacts the interpretations that can be made using the results. While many differences exist between Classical Test Theory (CTT) and Item Response Theory (IRT), one primary difference is that CTT is sample-dependent and item-dependent. The item difficulty estimates (i.e., p-value), item discrimination estimates (i.e., item-total correlation), and ability estimates (i.e., raw total scores or percent correct) are all dependent on the sample of students who took the items and the items on the assessment. In contrast, Item Response Theory (IRT) is sample- and item-independent. The ability estimates and item parameters (e.g., b-parameters) are not dependent on the students taking the test or on the items chosen for the test. With IRT, students' ability estimates can be compared even if students took different items on an assessment. IRT analyses do require a significantly larger sample size than CTT analyses, which can be a hindrance to using IRT.

To illustrate the importance of considering sample and item independence when making interpretations, examine Table 6, which shows the average percent correct for students for each form, a traditional CTT scoring statistic. It appears that students performed better on Form D than the other Forms. However, it is impossible to understand the differences between these scores. For example, if all forms had an equivalent number of lower- and higher-ability students, the items on Form A may have been more difficult than the items on Form D, as it might appear. However, if the sample for Form D was inadvertently skewed toward lower-ability students, the items on Form D could have been even much easier than the items on other forms (since the students taking Form D received higher scores). Similarly, if the sample for Form D was inadvertently skewed toward higher-ability students, the items on Form D may actually be much harder than those on the other forms, even though it may not appear that way. Using the CTT method of scoring, comparisons cannot be made between Forms without considering the sample and the items on the test.

Table 6
Average Student Score by Spatial Reasoning Pilot Test Form

Form	Average Student Score – Percent Correct (*SD*)
Form A	70% (10%)
Form B	73% (10%)
Form C	75% (9%)
Form D	76% (13%)

Because CTT is sample and item dependent, IRT is a more appropriate method to analyze the pilot test data and to build a final Spatial Reasoning form, especially since the EGMA is used widely and with many different populations. If CTT statistics were used to create the final form for Spatial Reasoning by choosing items from the four pilot test forms, the final form may perform very differently in the field than expected due to the fact that the pilot test statistics were based on a very specific pilot test sample. For example, a form could be constructed to have an average item difficulty of 0.70 and an average item-total correlation of 0.35. However, once the sample of students changes, those estimates could vary greatly, impacting the interpretations and usability of the results.

Because the EGMA is used widely around the world, IRT is the preferred method of analysis since IRT estimates are not dependent on the sample or the items on the assessment. However, since IRT has strong assumptions, including unidimensionality, that may be difficult to meet, IRT is not always used in the field. This may lead to using CTT statistics and running into interpretation challenges as illustrated above.

Discussion Question 3
What are the implications of the results from this study? One major implication from this study is that since the EGMA Spatial Reasoning subtask is not unidimensional, a single score should not be used to describe students' performance on this subtask. Because the internal consistency estimates and the EFA analyses suggest that the subtask is assessing multiple constructs, using a single score may be inappropriate. It may be more appropriate to separate the current Spatial Reasoning subtask into unique subtasks in order to report subscores for the different dimensions assessed. However, since a clear, theoretically driven factor model was not found for the Spatial Reasoning subtask, additional research should be conducted to investigate how to separate the items into subtasks.

One possibility would be to include a subtask and subscore for each Spatial Reasoning item type: three-dimensional spatial visualization, two-dimensional spatial visualization, three-dimensional spatial structuring, and two-dimensional spatial structuring. Since the items within an item type are theoretically similar and were written to assess the same latent trait, subtasks created using these item types theoretically should be unidimensional. Preliminary EFA results indicate that splitting the Spatial Reasoning subtask into subtasks based on item type may be promising for further development. Table 7 shows the EFA results for a one factor solution when each item type is treated as a separate subtask. The three-dimensional and two-dimensional spatial structuring tasks appear to be mostly unidimensional and have no items with negative factor loadings. Many of the fit statistics for the spatial structuring tasks are in the acceptable range. The three-dimensional and two-dimensional spatial visualization tasks, however, have items with negative loadings, and most fit statistics are outside of the acceptable range. While these results indicate that separating the Spatial Reasoning subtask into subtasks by item type may be promising, not all items are performing as intended and development should continue in order to improve the items.

Another major implication from these results is that further research is needed in order to understand why the items did not correlate well with one another. A first step to this research may be to re-examine comments from the expert review to understand potential factors that may have impacted item performance. For example, while the expert review revealed that the experts rated the Spatial Reasoning items as age-appropriate and representative and relevant to the construct of spatial reasoning, points were raised about graphics and language that could be contributing to the suboptimal performance of the items; these points should be addressed in

Table 7

Single Factor Exploratory Factor Analysis Results for Spatial Reasoning by Item Type

Item Type	Form	Number of Items	Number of Items with a Factor Loading Greater than 0.32	Number of Items with Negative Factor Loadings	TLI	RMSEA
3D Spatial Visualization	A	5	2	2	0.855	0.059
	B	5	2	0	0.409	0.120
	C	5	2	1	0.282	0.183
	D	5	2	0	0.243	0.190
2D Spatial Visualization	A	5	3	1	0.906	0.054
	B	5	3	2	0.933	0.103
	C	5	2	2	0.998	0.018
	D	5	2	2	0.915	0.081
3D Spatial Structuring	A	5	5	0	0.990	0.022
	B	5	2	0	0.170	0.138
	C	5	3	0	0.625	0.161
	D	5	3	0	0.952	0.036
2D Spatial Structuring	A	4	2	0	1.117	0.000
	B	4	3	0	1.000	0.005
	C	4	4	0	1.009	0.000
	D	4	3	0	0.823	0.076

future item development efforts. First, additional research should be conducted to determine the best graphics to use to assess students' spatial visualization and structuring abilities. The experts suggested that students may have difficulty determining if hidden cubes or squares were present in the three-dimensional and two-dimensional spatial structuring items. Currently, students may not believe that hidden cubes are necessary to the structural integrity of a figure or that rows of squares must be complete. Referencing previous research on spatial structuring and conducting cognitive interviews may provide insight into how to improve the graphics to prevent confusion about hidden cubes or squares. For example, the three-dimensional spatial structuring tasks developed for a research study by Battista and Clements (1996) involved complete rectangular prisms, which must have the same number of cubes on each layer. If students are told that the figure must be a rectangular prism, they should recognize that hidden cubes are necessary for the structural integrity of the prism. If the student does not account for hidden cubes, his/her response is based on a misconception, not confusion caused by the graphic. Similarly, the two-dimensional spatial structuring graphics (see Figure 4) used in a research study by Battista et al. (1998) are complete rectangles, which must have the same number of squares per row. Again, if students are told that the figure must be a rectangle, there is less ambiguity about the possibility of hidden squares. Currently, the spatial structuring items on the Spatial Reasoning subtask utilize irregular figures.

The shading of the graphics should also be examined in order to improve the items. An expert reviewer noted that the shading of the figures, particularly for the three-dimensional spatial visualization items, may cause confusion. The shading of a figure does not remain the same after a rotation. Students may consider the shading of faces when determining whether or not the figures are the same.

Additionally, future research is needed to determine the most appropriate and clear instructions for the spatial visualization items. The language of these items should be clarified to prevent confusion about what constitutes figures as being the "same." An expert reviewer suggested modifying the item prompt from "Please tell me if these objects are the same or not the same" to "If you could hold one of these structures in one hand and one in the other, are there positions in which they would look exactly alike?" An additional suggestion was, "If you could move these structures any way you want, is there a way that you could make them look exactly alike in all ways?" Student interviews could assist in determining how to best modify the item prompt to maximize clarity and prevent confusion.

After determining how to proceed with possible subscores and refining and developing additional items based on expert feedback and further research, the items should be pilot tested again. Then, the technical adequacy of the measures and the reliability of the ability estimates can be re-examined.

Conclusion

The purpose of this paper was two-fold: 1) to present information on the EGMA Spatial Reasoning subtask and the results from pilot testing the subtask in Jordan, and 2) to discuss how differences in item analysis techniques may lead to different validity conclusions. This study collected evidence to investigate one primary and two secondary research questions: Is the Spatial Reasoning subtask a technically adequate measure that reliably estimates students' spatial reasoning abilities? A) Does the two-parameter item response theory model fit the Spatial Reasoning pilot test data with acceptable item parameters and fit statistics? B) Is the reliability of the data generated by the

Spatial Reasoning subtask sufficient for the intended interpretation of the subtask?

Overall, the evidence collected to support the assumption that the Spatial Reasoning subtask is an accurate measure that reliably estimates students' spatial reasoning abilities suggests that future research and refinement of the items is needed before making interpretations about students' spatial reasoning abilities based on results of the Spatial Reasoning subtask.. EFA results indicated that the pilot test forms were not unidimensional, which halted IRT analyses. Because of this, the two parameter model could not be used to analyze the data (Research Question A). The reliability of the pilot test forms was calculated and the estimates were not in the desired range (Research Question B).

While some supplemental CTT statistics, such as item p-values, were in typical ranges, other CTT statistics, such as internal consistency, were not in acceptable ranges. Additionally, because of the desired uses and interpretations of the subtask, IRT is the preferred method of analysis due to it being sample- and item-independent. Future research is needed to investigate the dimensionality of the subtask and ways to refine the items.

Beyond the EGMA Spatial Reasoning subtask, this paper also illuminates the importance of the analyses that are used to evaluate assessments and evaluate the validity of the interpretations made using assessment scores. P-values are one of the most straightforward statistics to run on assessment data; however, p-values may also be one of the most misleading statistics. As noted with the EGMA Spatial Reasoning subtask data, the p-values were in a normal range, and many items had very high p-values. Just focusing on these easy-to-calculate scores may inadvertently lead to invalid interpretations about the functionality of these items. In contrast, IRT analyses indicated that the subtask was not unidimensional, which suggests that the score from the subtask cannot be used to describe a students' ability on a single construct.

Analyses should be chosen based on the purpose of the assessment and must be linked to the interpretations being made. High-stakes assessments, such as those that impact grade progression or student placement, should utilize rigorous analysis techniques, such as IRT, in order to collect technical adequacy and reliability evidence. Assessments, such as class tests, that have lower stakes and that likely do not meet IRT sample size requirements are better suited for CTT analyses. Aligning item analyses with assessment purpose will ensure that the evidence collected best supports the interpretations being made.

Author Note

Acknowledgements: I would like to thank RTI International for allowing me to analyze the data from the EGMA Spatial Reasoning pilot test for my dissertation, which was a basis for this manuscript.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Allahyar, M., & Hunt, E. (2003). The assessment of spatial orientation using virtual reality techniques. *International Journal of Testing, 3*(3), 263-275.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics, 21*(1), 47-60.

Battista, M. T., & Clements, D. H. (1996). Students' understanding of three-dimensional rectangular arrays of cubes. *Journal for Research in Mathematics Education, 27*(3), 258-292.

Battista, M. T., Clements, D. H., Arnoff, J., Battista, K., & Van Auken Borrow, C. (1998). Students' spatial structuring of 2D arrays of squares. *Journal for Research in Mathematics Education, 29*(5), 503-532.

Ben-Chaim, D., Lappan, G., & Houang, R. T. (1988). Visualizing rectangular solids made of small cubes: Analyzing and effecting students' performance. *Educational Studies in Mathematics, 16*(4), 389-409.

Bishop, A. J. (1980). Spatial abilities and mathematics education: A review. *Educational Studies in Mathematics, 11*(3), 257-269.

Clements, D. H. (2004). Geometric and spatial thinking in early childhood education. In D. H. Clements & J. Sarama (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 267–297). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Embretson, S. E., & Reise, S. (2000*). Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Ferketich, S. (1991). Focus on psychometrics: Aspects of item analysis. *Research in Nursing & Health, 14,* 165-168.

Frick, A., Hansen, M. A., & Newcombe, N. S. (2013). Development of mental rotation in 3-to-5-year-old children. *Cognitive Development, 28*, 386-399.

Galton, F. (1883). *Inquiries into the human faculty and its development*. London: Macmillan.

Gilligan, K. A., Flouri, E., & Farran, E. K. (2017). The contribution of spatial ability to mathematics achievement in middle childhood. *Journal of Experimental Child Psychology, 163*, 107-125.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hegarty, M., & Kozhevnikov, M. (1999). Types of visual–spatial representations and mathematical problem solving. *Journal of Educational Psychology, 91*(4), 684–689.

Holland, P. W., & Dorans, N. J. (2006). *Linking and equating*. In R. L. Brennan (Ed.), Educational Measurement (4th ed.) (pp. 187-220). Westport, CT: American Council on Education and Praeger Publishers.

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

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1(*2), 93-114.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Kosslyn, S. M. (1983). *Ghosts in the mind's machine*. New York: Norton.

Learning Metrics Task Force. (2013). *Toward universal learning: What every child should learn* (Report No. 3 of the Learning Metrics Task Force). Montreal & Washington, DC: UNESCO Institute for Statistics & Center for Universal Education at the Brookings Institution.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research, and Evaluation, 12*(2), 1-11.

Markey, S. M. (2009*). The relationship between visual-spatial reasoning ability and math and geometry problem-solving*. (Unpublished doctoral dissertation). American International College, Massachusetts.

Marmor, G. S. (1975). Development of kinetic images: When does the child first represent movement in mental images? *Cognitive Psychology, 7,* 548–559.

McGee, M. G. (1979). Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin, 86*(5), 889-918.

NCTM. (2000*). Principles and standards for school mathematics*. Reston, VA: NCTM.

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. C. T. Cross, T. A. Woods, & H. Schweingruber (Eds.). Committee on Early Childhood Mathematics, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olkun, S. (2003). Making connections: Improving spatial abilities with engineering drawing activities. *International Journal of Mathematics Teaching and Learning*, January 2003.

Perry, L. (2016). *Validating interpretations about student performance from the Early Grade Mathematics Assessment relational reasoning and spatial reasoning subtasks* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (Order No. 10164141).

Perry, L. (2017). EGMA *Spatial Reasoning and Relational Reasoning subtasks: Content evidence*. In T. A. Olson & L. Venenciano (Eds.), Proceedings of the 44th Annual Meeting of the Research Council on Mathematics Learning. Fort Worth, TX.

Platas, L. M., Ketterlin-Geller, L., Brombacher, A., & Sitabkhan, Y. (2014*). Early grade mathematics assessment (EGMA) toolki*t. Research Triangle Park, NC: RTI International.

Revelle, W. (2015). *psych: Procedures for personality and psychological research. R package*. Evanston, Illinois: Northwestern University.

Robinson, N. M., Abbott, R. D., Berninger, V. W., & Busse, J. (1996). The structure of abilities in mathematically precocious young children: Gender similarities and differences. *Journal of Educational Psychology, 88*, 341-352.

Sarama, J., & Clements, D. H. (2009*). Early childhood mathematics education research: learning trajectories for young children*. New York: Routledge.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn and Bacon.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs, 2*.

van Garderen, D. (2006). Spatial visualization, visual imagery, and mathematical problem solving of students with varying abilities. *Journal of Learning Disabilities, 39*(6), 496-506.

van Hiele, P. M. (1986). *Structure and insight. A theory of Mathematics Education*. Orlando, FL: Academic Press, Inc.

van Nes, F., & de Lange, J. (2007). Mathematics education and neurosciences: Relating spatial structures to the development of spatial sense and number sense. *The Montana Mathematics Enthusiast, 4*(2), 210-229.

About the Author

Lindsey Perry, PhD, is a Research Assistant Professor at Southern Methodist University. Her current research interests focus on investigating children's spatial and relational reasoning abilities, developing mathematics assessments for young children, and training educators on how to use data from assessments to make instructional decisions.