Thayaseelan *et al.*
Disciplinary and Interdisciplinary Science Education Research (2024) 6:3
https://doi.org/10.1186/s43031-024-00095-8

Disciplinary and Interdisciplinary Science Education Research

RESEARCH Open Access

Revalidating a measurement instrument of spatial thinking ability for junior and high school students

Kannaki Thayaseelan[1], Yanfang Zhai[1,2*], Siqi Li[3] and Xiufeng Liu[1]

Abstract
Spatial thinking is a set of cognitive abilities that enable people to organize, reason about, and mentally manipulate both real and imagined spaces. One of the available measurement instruments is the Spatial Thinking Ability Test (STAT). Given the critical need for spatial thinking ability measurement for junior and high school students, and the popularity of STAT to measure spatial thinking ability, revalidation of STAT is necessary as STAT was developed primarily for university students and validation of the original STAT was based on the classical test theory from which the findings are notoriously sample dependent. We used Rasch modeling to revalidate STAT as it allows parameters to be mutually independent and measures to be interval. The sample included 1340 junior and high school students. Item fit statistics, Item Characteristics Curves, unidimensionality test, and the Wright map provided evidence for the construct validity of STAT measures. The reliability of the instrument was moderate. Wald test for item measure invariance of individual items showed that among sixteen items seven items were variant in measures. The Anderson LR test indicates that the Rasch difficulty measures of STAT were not adequate for invariance. There was no DIF between two subsamples based on gender, suggesting fairness of the instrument in terms of gender. The above results suggest that STAT possesses certain degrees of validity, reliability, and fairness, although there is still room for further improvement.

Keywords Spatial thinking, Rasch measurement, Validity, Reliability, Fairness

Introduction

Spatial thinking is a distinct, universal, and productive form of thinking used in a variety of academic disciplines, ranging from psychology to natural sciences, although each discipline may emphasize different aspects of spatial thinking. Spatial thinking is often used synonymously with spatial ability (Lee & Jo, 2022); however spatial ability and spatial thinking are distinct in that spatial ability is a psychological trait while spatial thinking is a collection of cognitive skills that involves both knowledge and cognitive operations applied to knowledge (NRC, 2006). Spatial thinking has also been called spatial thinking ability. For example, Lohman (1996) defined spatial ability as the "ability to generate, retain, retrieve, and transform

*Correspondence:
Yanfang Zhai
yzhai7@buffalo.edu
[1]Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York, Buffalo, NY 14260-1000, USA
[2]School of Education, Capital Normal University, Beijing, China
[3]College of Education for the Future, Beijing Normal University, Zhuhai, China

well-structured visual images"(p.112). Alkan and Erdem (2011) stated that "spatial abilities are described as the combination of the skills such as creating mental pictures of objects in the universe, recognizing in different ways and budging these objects as a whole or in pieces individually" (p.3446). Spatial thinking is a set of cognitive abilities that enable us to organize, reason about, and mentally manipulate both real and imagined spaces. These include reasoning about the shape, size, orientation, direction, and trajectory of objects, the relationships between objects, mentally visualizing objects and/or their relationships, and reasoning about objects and their spatial and time relationships (Gagnier et al., 2022; NRC, 2006).

Developing spatial thinking ability of students ranging from primary school through college bears great promise of improving Science, Technology, Engineering, and Mathematics (STEM) education (Gagnier et al., 2022). Research has demonstrated that spatial thinking ability benefits students to improve their STEM learning outcomes. Specifically, studies spanning more than 60 years have revealed that spatial thinking abilities are crucial for success in STEM fields (Shea et al., 2001; Wai et al., 2009). Assessment of preschool students' spatial skills followed through high school (Wolfgang et al., 2003) showed that spatial thinking abilities are significant and independent predictors of later school success in STEM education. Furthermore, there is plentiful empirical evidence that spatial thinking ability underpins students' comprehension and reasoning of scientific phenomena (e.g., Gagnier et al., 2017; Jee et al., 2013; Kozhevnikov et al., 2007; Mix, 2019; Verdine et al., 2017).

There is a consensus that measurement of spatial thinking should be contextual and integrated with context (NRC, 2006; Eliot & Czarnolewski, 2007, Hegarty et al., 2002; Lee & Jo, 2022). For example, Nazareth et al. (2019) and van der Ham et al. (2020) assessed spatial thinking in the environmental context. This kind of assessment of spatial thinking is typically at a large geographical scale and has been carried out in disciplines focusing on spatial concepts, such as mathematics, geography, geosciences, and environment science. There is another type of assessment of spatial thinking by using Likert-scale questionnaires (e.g., Erskine et al., 2015; Turgut, 2015). Respondents are asked to express their levels of agreement to a series of statements regarding their attitudes, self-assessed abilities, and ideas about spatial thinking. This type of assessment of spatial thinking does not directly assess a cognitive capacity, but rather a sense of that capacity (Lee & Jo, 2022).

One measurement instrument for spatial thinking that assesses a comprehensive list of spatial thinking components with reported validity and reliability is the Spatial Thinking Ability Test (STAT) (Lee & Bednarz, 2012). STAT measures spatial thinking abilities in the context of geography. The reliability and validity of STAT were examined using the classical test theory (i.e., principal component analysis, Cronbach's internal consistent reliability) based on samples of university students. Specifically, Cronbach's alpha was in the moderate range from 0.70 to 0.72 for the components. Although the authors hypothesized that there were eight components within the 16 items, principal components analysis with varimax rotation revealed that there were only six components with eigenvalues greater than 1 and many items loaded on multiple components, suggesting that the hypothesized components of STAT were not independent. Thus, evidence did not support the hypothesis that spatial thinking abilities consist of eight independent skills. It remains unknown how many independent skills form spatial thinking abilities.

Since its publication, STAT has been applied to many published studies. For instance, based on STAT, Kim and Bednarz (2013) developed an interview-based critical spatial thinking oral test. Huynh and Sharpe (2013) created a test of geospatial thinking, built on the work of Lee and Bednardz (2012) and Battersby et al. (2006), to enabled teachers to benchmark student performance levels of understanding. Tomaszewski et al. (2015) modified the STAT, used it in the Rwandan cultural context, and discovered that, in terms of spatial thinking abilities, urban and male students outperformed rural and female students. In Liu et al.'s (2019) study, a 28-item modified STAT test was administered at the end of an undergraduate geography course. The reliability of STAT measure in Cronbach's alpha was calculated to be 0.71, whereas the validity of STAT measure was not examined in this study. Duarte et al. (2022) used a modified STAT, composed of 15 multiple choice questions to test the spatial thinking of 83 students in two different curricular units involving geographical information systems (GIS) concepts and software. The authors conducted content analysis and calculated reliability of the STAT measure. It can be seen that no published studies that applied STAT systematically established validity and reliability, likely assuming the STAT is valid and reliable. Also, there is still a confusion on if spatial thinking abilities should be unidimensional or multidimensional.

Given the increasing prevalence of using STAT in various research studies, it is important to ensure the validity and reliability of STAT measures. STAT was validated using the Classical Test Theory (CTT). When CTT is used to develop and validate measurement instruments, a number of fundamental limitations exist (Liu, 2020). One such limitation is that validity and reliability estimates of STAT measures are sample dependent, i.e., different samples may result in different validity and reliability estimates. Another limitation is that measures of STAT

are in total raw scores. Total raw scores are ordinal, not interval, a requirement for inferential statistical analysis such as t-test. Third, the standard error of measurement for STAT measures (i.e., person abilities) is an overall statistic for all persons collectively; person specific standard error of measurement is not possible. In addition to the above limitations associated with validation of STAT, it is necessary to clarify the nature of spatial thinking abilities, i.e., whether or not they are unidimensional or multidimensional, and if multidimesional how many dimensions exist, because Rasch modeling approaches dimensionality through standardized residuals instead of raw scores. Principal component analysis results of raw scores are sample dependent, while principal component analysis of Rasch standardized residuals is sample independent. Finally, the STAT was validated using undergraduate students; if it is used for junior and high school students, which is the purpose of our study, we need to establish evidence for the validity, reliability as well as fairness of the STAT measures.

The purpose of this study was to revalidate STAT for junior and high school students as a unidimensional measure. It used the Rasch modeling (Liu, 2020) to establish evidence of validity, reliability, and fairness of STAT measures. Due to the unique characteristics of Rasch models (Liu, 2020), revalidating STAT using Rasch modeling can address the limitations associated with the reported validation of STAT using CTT. First, estimates of the person and item parameters are in logits from $-\infty$ to $+\infty$, thus truly interval. Second, student measures and item measures are independent of one another. Therefore, a person's ability is unaffected by the difficulty of items of the instrument that he or she answers, and an item's difficulty is unaffected by the ability of the persons who take it. Because of the above, the validity and reliability estimates based on those item and person measures are sample independent. Third, Rasch modeling produces student and item specific standard errors of measurement. Fourth, we will clarify the nature of spatial thinking abilities, i.e., whether or not they are unidimensional or multidimensional, using Rasch principal component analysis of standardized Rasch residuals; the result will be sample independent. Finally, we will use a data set of junior and high school students to revalidate STAT, making it usable for junior and high school student.

The specific research questions for this study are:

(a) what evidence is available to support the validity claims of STAT for measuring junior and high school students' spatial thinking ability?
(b) what evidence is available to support the reliability claims of STAT for measuring junior and high school student's spatial thinking abilities?
(c) Is STAT fair in measuring male and female students' spatial thinking abilities?

Research questions a and b pertain to validity and reliability of measurement instruments. According to the Standards for Educational and Psychological Testing (Joint committee of AERA, APA and NCME, 2014), a third foundation of measurement is fairness. The goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure (Joint committee of AERA, APA and MCME, 2014, p. 51). Fairness can be demonstrated in many different ways. In this revalidation study, we address fairness in terms of gender equality, i.e., lack of bias against a particular gender, which is what research question c is about.

Theoretical framework

The nature and characteristics of spatial thinking conceptualized in the National Research Council Report (NRC, 2006) guided our study. According to the report, spatial thinking is an approach to problem solving via coordinated use of space, representation, and reasoning. Space makes spatial thinking a district form of thinking from other forms of thinking such as verbal thinking, mathematical thinking, hypothetical thinking. There are three contexts that define space: life space, physical space, and intellectual space. Life space refers to the world people live in; physical space refers to the four-dimensional world of space-time, and the intellectual space refers to the representations people construct such as a concept map. Concepts of space provide a conceptual and analytical framework that enables the integration, correlation, and arrangement of data into a unified entity. Representations, whether they are internal and cognitive, or external and graphical, provide the capacity to store, analyze, comprehend, and communicate organized information to others. Reasoning processes enable the manipulation, interpretation, and explanation of structured information. This cognitive process involves perceiving connections, conceptualizing changes in magnitude, mentally manipulating an item to observe its different aspects, generating a new perspective or viewpoint, and retaining mental representations of images within specific locations and environments. Spatial thinking serves three purposes: (a) a descriptive function to capture, preserve, and convey the appearance of and relations among objects; (b) an analytic function to enable an understanding of the structure of objects, and (c) an inferential function to generate answers to questions about the evolution and function of objects.

Spatial thinking process can be decomposed into various competences or components (NRC, 2006). Example components include representation and transformations

of representations. Combining different representations and transformations gives rise to various complex spatial reasoning. Referencing various competences or components reported in the literature, Lee and Bednarz (2012) identified eight components to develop their STAT test items: (a) comprehending orientation and direction, (b) discerning spatial patterns and graphing a spatial transition, (c) comprehending overlay and dissolve and inferring a spatial aura (influence), (d) recognizing spatial form and transforming perceptions, representations and images from one dimension to another and the reverse, and graphing a spatial transition, (e) comprehending spatial association, making a spatial comparison, and assessing a spatial association, (f) transforming perceptions, representations and images from one dimension to another and the reverse, (g) Overlaying and dissolving maps, (h) comprehending integration of geographical features represented as points, networks, and regions, and comprehending spatial shapes and patterns. Lee and Bednarz further hypothesized that the above eight components form eight distinct measurement components with insignificant correlations among them. However, principal component analysis showed that the eight components are highly correlated. Given the above empirical finding by Lee and Bednarz (2012), and spatial thinking is a constructive amalgam of space, representation and reasoning, and involves a process of four levels of representations and reasoning starting with a set of primitives, adding language of space, deriving spatial concepts, and finally performing cognitive operations (NRC, 2006), it is reasonable to conceptualize various spatial thinking components to constitute a coherent system distributed along the process of above four levels. Thus, the eight components measured by STAT should be unidimensional psychometrically.

Methods

This study adopted Rasch modeling to validate the STAT. Rasch models were originated from the pioneering work of the Danish mathematician Georg Rasch (Rasch,1960/1980). A Rasch model assumes that there exists a linear measure common to both items and examinees. For items, this measure is item difficulty and for examinees it is ability. The unit of measures is logit, i.e., the natural logarithm of odds.

Specifically, we used the binary Rasch model. According to Rasch, for any item *i* with a difficulty D_i that can be scored as right (X = 1) or wrong (X = 0), the probability (*P*) of a person n with an ability B_n to answer the item correctly can be expressed as

$$P(\mathrm{X}=1|\mathrm{Bn},\mathrm{Di})=\frac{\mathrm{e}^{(\mathrm{Bn}-\mathrm{Di})}}{1+\mathrm{e}^{(\mathrm{Bn}-\mathrm{Di})}}$$

We can see from the above equation, the probability of a person to answer an item correctly is determined solely by the difference between that person's ability and item's difficulty. By applying this model to our data set, we analyzed Item Characteristic Curves (ICCs), item fit statistics, unidimensionality based on principal component analysis of standardized residuals and Martin-Löf Likelihood ratio test, the Wright map, and invariance of item measures based on Wald test and Anderson's Likelihood test to establish evidence of construct validity of measures to answer research question a; we also analyzed the person separation index, discernable strata, and Cronbach's alpha to establish evidence of reliability to answer research question b; finally we conducted Differential Item Functional (DIF) to establish evidence for the absence of bias to answer research question c.

Stat test

The STAT consists of sixteen multiple-choice questions (Lee & Bednarz, 2012). All questions are framed within the disciplinary context of geography. Questions are scored correct or incorrect; two sample items are shown in Figure 1.

Data set

Data were collected as part of an NSF funded teacher professional development (PD) project focused on integration of GIS technologies into junior and high school STEM teaching. Each year for three years (2017–2020), 10 teachers from two districts were recruited to participate in the PD project. The teachers were grades 7–12 social science, science, and technology teachers. In the US, physical geography/Earth science is part of Social Studies curriculum in junior high school; it is within the domain of STEM education. The teachers attended a six-week summer institute learning GIS technologies (i.e., Google earth, ESRI story map, sketch up, ArcGIS online and collector, and drones); the teachers also developed lessons plans during the summer institute to integrate GIS technologies into the courses they would teach in the upcoming academic year. During the academic year, the teachers taught the GIS integrated lessons to their students. Over three years, a total of 26 teachers were involved. Students of those teachers completed a pre-test and a post-test using the STAT questions; the pre-test took place in Sept. and the post-test took place in May next year. Altogether, there were 1340 students who completed the STAT: 854 for pre-test, and 486 for post-test. The pre-test and post-test data were pooled into one data set for analysis.

The research protocol was approved by the Institutional Review Board (IRB); parents of student participants as well as students (if they were 18 years or older) received and signed the consent form to participate.

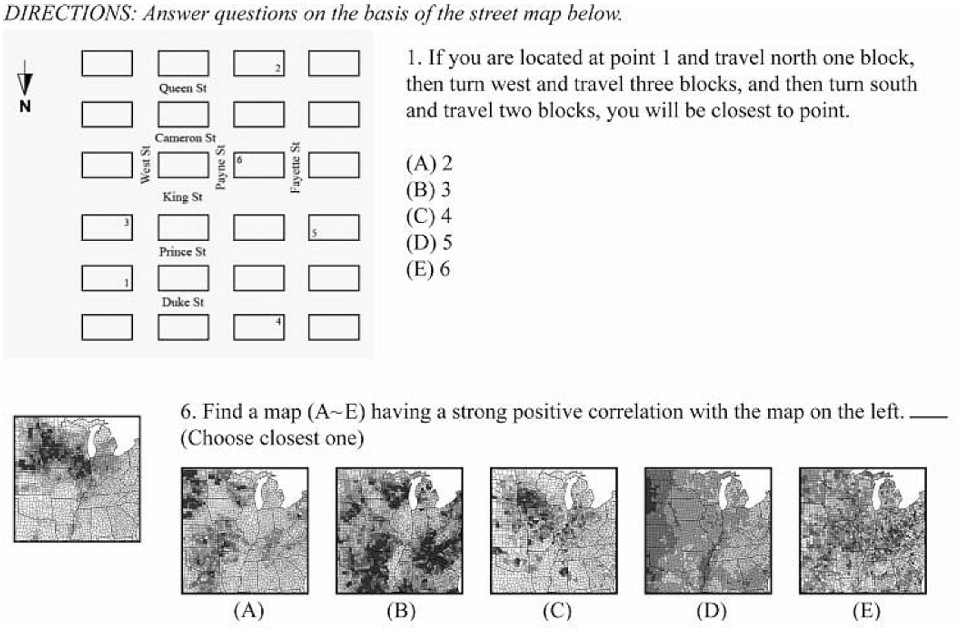

Fig. 1 Two sample STAT questions that are scored correct or incorrect

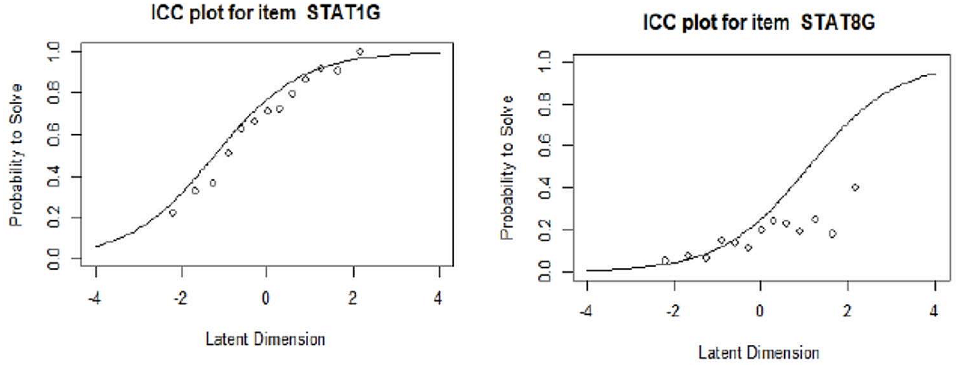

Fig. 2 Item Characteristics Curves for the Item 1 (STAT1G) and Item 8 (STAT8G). Item 1 has a good agreement between the expected probability (the solid curve) and the observed probabilities (the dots). Item 8 has a poor agreement between the expected probability (the solid curve) and the observed probabilities (the dots)

Analysis

The binary Rasch model was used in the analysis. Rasch analysis was conducted using R software. R is an open-source software which provides a statistical computing environment and graphics. Specifically, "eRm", "psych", TAM, and "diffr" packages were used for Rasch analysis. Package 'eRm' allows analyses of Item Characteristic Curves (ICCs), item fit statistics, the Wright map, and Wald test and Anderson's Likelihood test for measure invariance (Mair et al., 2023); package 'psych' allows principal component analysis of standardized residuals and Martin-Löf Likelihood ratio test for dimensionality (Revelle, 2023); package 'TAM' allows analysis of the separation index, discernable strata, and Cronbach's alpha (Robitzsch et al., 2023); finally package 'diffr' allows analysis of Differential Item Functional (DIF) (Muschelli, 2022).

Results

Validity evidence

Item characteristic curves (ICCs)

Item characteristic curves plot the probability of answering each item correctly on a continuum. If there is a good model data fit, the expected and observed ICCS should be within a 95% confidence interval. Figure 2 presents two sample ICCs, with one demonstrating a good agreement between the expected probabilities and the observed probabilities, and another with a poor agreement between the two. From Fig. 2, we see that Item 1 has a good agreement between the expected probability (the solid curve) and the observed probabilities (the

Table 1 Fit statistics of items of STAT

Item	Outfit MNSQ	Infit MNSQ
STAT1G	1.048	1.035
STAT2G	1.023	1.032
STAT3G	0.868	0.902
STAT4G	0.785	0.885
STAT5G	1.010	0.960
STAT6G	0.834	0.881
STAT7G	1.206	1.146
STAT8G	1.382	1.098
STAT9G	0.964	1.011
STAT10G	0.764	0.848
STAT11G	0.926	0.899
STAT12G	1.125	0.949
STAT13G	0.957	0.976
STAT14G	1.135	1.093
STAT15G	0.862	0.874
STAT16G	1.137	1.120

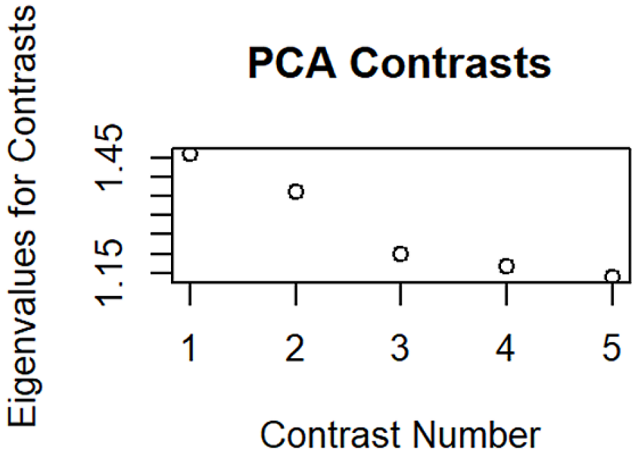

Fig. 3 Principal Component Analysis of standardized residuals. X-axis shows the contrast number, and Y-axis shows the eigenvalue of the contrasts. No contrast had eigenvalue greater than 2, suggesting that the measures were unidimensional

dots), because the dots are evenly distributed along and close to the curve; on the other hands, item 8 has a poor agreement between the expected probability (the solid line) and the observed probabilities (the dots), because the dots above logit 0 are far away from the curve and cluster between 0 logit and 2 logit. Item 1 has a good model-data-fit, while item 8 has a poor model-data-fit. Overall, with the exception of item 8 (STAT8G), there was a good model data fit for all items.

Item fit statistics

There are four fit statistics: infit MNSQs, infit ZSTDs, outfit MNSQs and outfit ZSTDs. Because of the large sample size, only Infit MNSQs and Outfit MNSQs were considered (Boone et al., 2014). MNSQ stands for mean square residuals; ZSTD stands for standardized mean square residuals. Infit is a weighted mean square residual by giving more weights to those subjects whose abilities are close to the item difficulty, while outfit is simple average of mean square residuals by weighing all subjects equally. An item with good model-data-fit should have infit and outfit MNSQs within the range of 0.7–1.3 for multiple choice questions (Bond & Fox, 2015). Table 1 presents the fit statistics. From Table 1, we see that only one item, STAT8G, has outfit MNSQ outside the acceptable range; all other items have both Outfit MNSQ and Infit MNSQ within the acceptable range.

Unidimensionality

In this revalidation, we hypothesized that spatial thinking abilities measured by STAT were unidimensional. Unidimensionality refers to the fact that measures describe only one attribute. The Rasch principal component analysis of residuals can be used to identify if additional dimensions may exist in the residuals. The eigenvalue of the contrasts/components should be smaller than 2 if unidimensionality is strictly held (Linacre, 2023). Figure 3 presents the principal component analysis results. The x-axis shows the contrast number, and the y-axis shows the eigenvalue of the contrasts.

From Figure 4, we see that the eigenvalues of first, second, third, fourth and fifth contrasts were less than 2, suggesting unidimensionality of measures.

Martin-Löf (MLoef) likelihood ratio (LR) test

Unidimensionality was also checked by the Martin-Löf (MLoef) Likelihood Ratio (LR) test. MLoef test splits the items into two subgroups based on the median item raw scores and test if the two subgroups are homogenous. The results as follows: LR-value: 77.784, Chi-square=77.784, df=63, $p=0.099$. Because $p>0.05$, item difficulty measures based on two subgroups of students were statistically the same, suggesting that the items measures were unidimensional.

The wright map

The wright map shows how items target persons. A good measurement instrument should be able to target the intended population by matching its difficulty distribution with the sample's ability distribution. A gap indicates that subjects within that gap cannot be accurately differentiated because of the lack of items at that level (Liu, 2020). Figure 3 presents the Wright map.

The bottom of the graph is the Rasch scale in logit, the vertical line on the left-hand presents the items, the top panel of histograms shows the frequency distribution of subjects' ability estimates along the Rasch scale at the bottom, and finally the dots within the square show the locations of item difficulties on the Rasch scale at the bottom. Figure 4 shows several gaps circled in red where

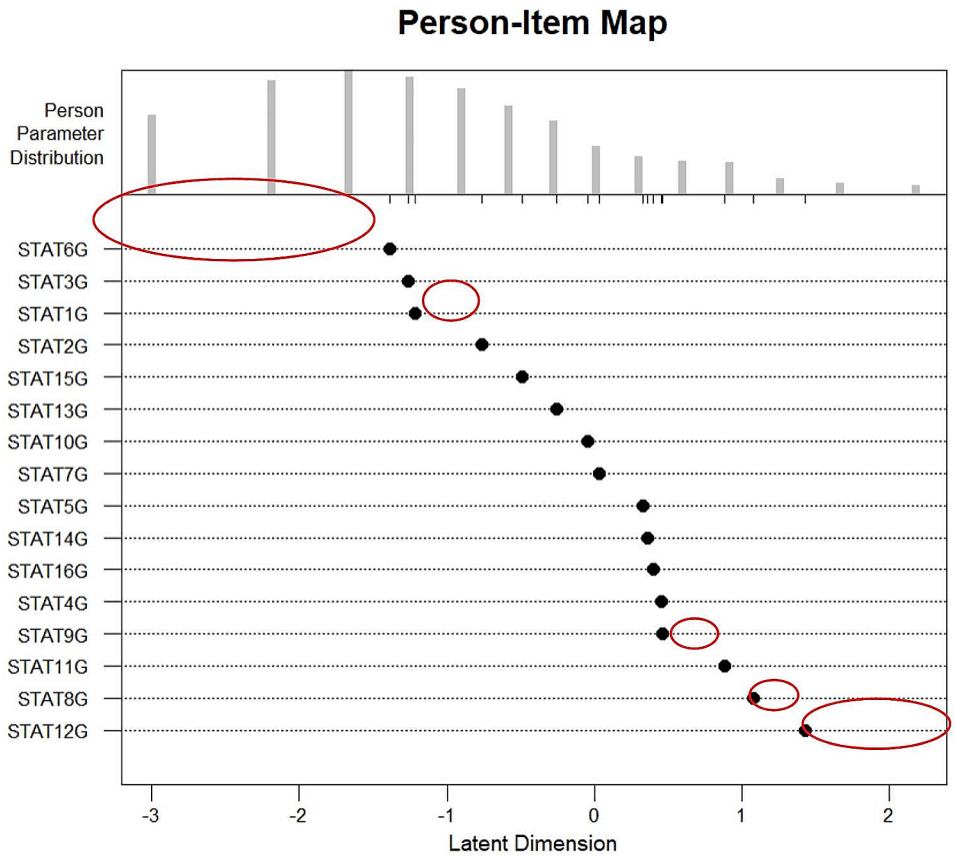

Fig. 4 Wright Map showing how STAT items target student abilities. There is a need for additional items to be added to fill the gaps in student abilities as circled in red

Table 2 Wald test on item level (z-values)

Item	Z statistic	P value
beta STAT1G	0.244	0.807
beta STAT2G	1.443	0.149
beta STAT3G	-2.674	0.007
beta STAT4G	-3.650	0.000
beta STAT5G	0.876	0.381
beta STAT6G	-3.877	0.000
beta STAT7G	4.399	0.000
beta STAT8G	4.806	0.000
beta STAT9G	0.041	0.967
beta STAT10G	-3.862	0.000
beta STAT11G	-1.678	0.093
beta STAT12G	1.972	0.049
beta STAT13G	0.133	0.894
beta STAT14G	2.142	0.032
beta STAT15G	-3.497	0.000
beta STAT16G	2.921	0.003

there were no items to measure the students of certain ability ranges (e.g., below −1.5 and above +1.4, between −1.2 and −0.8, between 0 and 0.2,, and between 0.5 and 0.8 logits). New items are needed to over those ranges of student abilities.

Invariance

For a good measurement instrument, item difficulty measures should be invariant from the sample used to calibrate the item difficulties, and person ability measures should be invariant from the set of items used to produce the ability measures (Liu, 2020). Wald test splits the sample into two sub-samples from the median raw score and test if item difficulty measures obtained from the two sub-samples are statistically the same using a Z test. A Bonferroni adjustment to the alpha value is necessary due to multiple tests being used and inflated type I error. For the 16 STAT items, the cut-off p value for significance should be adjusted, i.e., 0.05/16=0.003. Thus, a cut-off p value of 0.003 should be used to decide if the different is statistically significant. Table 2 presents the Wald test results. From Table 2, we see that items STAT4G, STAT6G, STAT7G, STAT8G, STAT10G, and STAT15G are flagged for lack of invariance, because the p values for those items were smaller than 0.003.

Anderson's likelihood test

To test the overall invariance of items as a whole, Andersen's Likelihood Ratio (LR) test (1973) splits the data into two groups based on the median raw score and compare

Table 3 Mantel-Haenszel Chi-square statistic

Item	Stat	P-value	Adj.P
STAT 1G	0.3993	0.5275	1.0000
STAT2G	3.7427	0.0530	0.8486
STAT3G	0.0197	0.8883	1.0000
STAT4G	2.0613	0.1511	1.0000
STAT5G	0.1168	0.7325	1.0000
STAT6G	1.9028	0.1678	1.0000
STAT7G	1.9278	0.1650	1.0000
STAT8G	0.0043	0.9478	1.0000
STAT9G	0.3089	0.5784	1.0000
STAT10G	0.0012	0.9719	1.0000
STAT11G	1.0029	0.3166	1.0000
STAT12G	0.0046	0.9460	1.0000
STAT13G	0.1875	0.6650	1.0000
STAT14G	0.0009	0.9767	1.0000
STAT15G	2.9398	0.0864	1.0000
STAT16G	2.7862	0.0951	1.0000

Table 4 Effect Size of Mantel-Haenszel Test (ETS Delta scale)

Item	Alpha MH	Delta MH
STAT1G	1.0943	-0.2117 A
STAT2G	1.3045	-0.6247 A
STAT3G	0.9705	0.0703 A
STAT4G	1.2910	-0.6003 A
STAT5G	0.9354	0.1569 A
STAT6G	0.8137	0.4846 A
STAT7G	0.8182	0.4715 A
STAT8G	0.9955	0.0105 A
STAT9G	1.1051	-0.2349 A
STAT10G	1.0073	-0.0171 A
STAT11G	0.8157	0.4786 A
STAT12G	1.0375	-0.0866 A
STAT13G	0.9304	0.1696 A
STAT14G	0.9848	0.0361 A
STAT15G	1.3057	-0.6268 A
STAT16G	0.7664	0.6251 A

Note: 'A': negligible effect, 'B': moderate effect, 'C': large effect

the calibration results of item measures of two groups. Anderson's LR test results are as follows: LR-value: chi-square=120.856, df=15, p=0. The results indicate that overall, the Rasch difficulty measures of STAT were not invariant (p<0.05).

Reliability evidence

Reliability is an important property of measures and essential for any instrument. The person separation index indicates replicability of person ordering on an interval scale if they were given a parallel set of items measuring the same latent trait (Wright & Master, 1982). The person separation index was 1.588, discernable strata (i.e., how many distinct groups can be differentiated) was 2.450, and Cronbach's alpha was 0.716. The desirable reliability for standardized measurement instruments should be 0.8, the desirable separation index should be greater than 2, and the discernable strata should be greater than 3 (Linacre, 2023). Thus, the reliability of STAT measures was moderate.

Fairness evidence: differential item functioning (DIF) Test

In order to examine whether or not there is bias in item difficulty measures (i.e., fairness), it is necessary to have two sub-samples based on subject characteristics such as gender. There should be no significant difference in item difficulty measures obtained from the two sub-samples if STAT is fair. Any difference found may indicate bias in items which is called DIF (Liu, 2020). Table 3 presents the Mantel-Haenszel test results and Table 4 presents the DIF effect sizes. From Tables 3 and 4, we see that there was no DIF in STAT items.

Discussion

According to the educational and psychological testing standards (Joint Committee of AERA, APA and NCME, 2014), validation of a standardized measurement instrument should establish evidence for validity, reliability, and fairness claims. Further, validity, reliability and fairness are not binary concepts; they are matters of degree. For revalidation of STAT, Rasch modeling provides various types of evidence for establishing validity claims. From the above results, the item characteristic curves show that there was a good model data fit for all items with the exceptions of one item (items 8). Item fit statistics of STAT show that, except for one item (item 8), all other items were within the acceptable MNSQ range. Figure 5 presents item 8. Item 8 requires students to mentally visualize a 3-D image based on 2-D information; it assesses students' ability to transform perceptions, representations, and images from one dimension. Item 8 is the second most difficulty question among the 16 STAT items. An item that is too difficult often fits the Rasch model poorly because no student responses are available in high ability region (see Fig. 2 ICC for item 8). In this case, the item itself may not be flawed; if a sample includes more higher ability students, the item may still fit the Rasch model well. Item 8 may not be of too much a concern in terms of threat to validity. The above findings of ICCs and fit statistics suggest that the 16 items of STAT test are of high quality and can result in valid measures of spatial thinking ability.

Examining the interaction between persons and items can also demonstrate construct validity. A good match in range of measures between persons and items and an even distribution of items along the construct demonstrate construct validity. As presented in Fig. 4, STAT items cover a 3-logit range from −1.5 to 1.5. Typically, a good measurement instrument should cover at least 5 logits. A smaller range of coverage suggests that measures

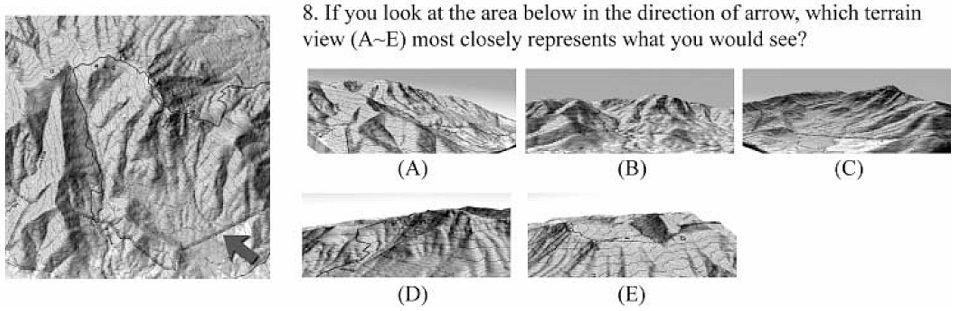

Fig. 5 Item 8 has poor fit statistics. The correct answer is C

of the STAT test possess construct validity for students within a limited range of abilities. Specifically, Wright map shows some ranges of students' spatial thinking ability levels still lacked items of the corresponding difficulties. A direct consequence of the above is that students whose abilities within those ranges may not be accurately measured. Further, as presented above, Wald test and Anderson LR test results show that item difficulties were different for higher ability and lower ability students for six items and overall item difficulties were not invariant, suggesting that STAT item difficulties and student ability measures may still depend on each other. This result is also likely due to the gaps in item difficulties shown in the Wright map.

Besides the above limitation in construct validity of STAT measures, students within those ranges without items may not be reliably measured, i.e., measures of those students will have larger standard errors of measurement. Insufficient items for those ranges of student abilities are a direct cause for the moderate person reliability reported above.

Thus, in order to increase construct validity and at the same time reduce standard errors of measurement for students within certain ranges and the overall person reliability of measures, new items need to be added for these ranges of spatial thinking ability levels in order to increase the construct validity of STAT measures.

Another important evidence for the construct validity of a measurement instrument is unidimensionality. The original STAT was hypothesized as multidimensional, but evidence did not support the multidimensionality of STAT measures (Lee & Bednarz, 2012). In this study, we hypothesized that STAT measures should be unidimensional. The Rasch principal component analysis of standardized residuals and Martin-Löf (MLoef) Likelihood Ratio (LR) test all show unidimensional nature of STAT measures, which provides additional evidence for the construct validity of STAT measures.

Finally, fairness of the test aims to maximize the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure. In this study, we focused on gender difference in spatial thinking ability. Given that no literature has reported statistically significant differences between boys and girls in middle and high school in their spatial thinking abilities, we should not expect difficulties of STAT items to be statistically significantly different. From the DIF analysis results reported above, there is no DIF between two subsamples based on gender on any item, suggesting that STAT measures are fair in terms of gender.

Conclusion

This study aimed at revalidating STAT for junior and high school students as a unidimensional measure. Utilizing Rasch modeling to establish evidence of validity, reliability, and fairness of STAT measures, the result suggests that STAT possesses certain degrees of validity, reliability, and fairness, for students within the middle range of spatial thinking abilities. For students whose spatial abilities are low or high, STAT measures may not be valid and reliable. Thus, in order to further improve the validity and reliability of the STAT test, additional items at low and high ability levels as well as at certain middle ability ranges should be added. Still, the current version of the 16 item STAT test can be used by can be used by teachers and researchers to measure average junior and high school students' spatial thinking abilities.

List of Abbreviations

STA	Spatial Thinking Ability
STEM	Science, Technology, Engineering, and Mathematics
STAT	Spatial Thinking Ability Test
GIS	geographical information systems
CTT	Classical Test Theory
PD	teacher professional development
MNSQ	Mean Square Residuals
DIF	Differential Item Functioning
MLoef	Martin-Löf
LR	Likelihood Ratio

Acknowledgements
Not applicable.

Author contributions
KT contributed to analyzing and interpreting the data as well as writing the article. YF performed the literature review and contributed to writing the article. SQ collected the data. XF provided guidance on conceputaliztion, data analysis and interpretation, and manuscript preparation. All authors read and approved the final manuscript.

Funding
This article is based on work supported by the National Science Foundation under Grant No. DRL-1614976. Any opinions, findings, and conclusions or recommendations expressed in the materials are our own and do not necessarily reflect the views of the National Science Foundation.

Data availability
The datasets used for this study are available from Dr. Xiufeng Liu on reasonable request.

Declarations

Competing interests
KT, YZ and SL have no competing interests in this study; XL is a co-editor-in-chief of DISER.

Received: 29 September 2023 / Accepted: 30 December 2023
Published online: 04 January 2024

References

Alkan, F., & Erdem, E. (2011). A study on developing candidate teachers' spatial visualization and graphing abilities. *Procedia - Social and Behavioral Sciences*, *15*, 3446–3450. https://doi.org/10.1016/j.sbspro.2011.04.316.
Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/bf02291180.
Battersby, S. E., Golledge, R. G., & Marsh, M. J. (2006). Incidental learning of geospatial concepts across grade levels: Map overlay. *Journal of Geography*, *105*(4), 139–146.
Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd edition.). Routledge.
Boone, W. J., Staver, J., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
Duarte, L., Teodoro, A. C., & Gonçalves, H. (2022). Evaluation of spatial thinking ability based on exposure to geographical information systems (GIS) concepts in the context of higher education. *ISPRS International Journal of Geo-Information*, *11*(8), 417. https://doi.org/10.3390/ijgi11080417.
Eliot, J., & Czarnolewski, M. Y. (2007). Development of an everyday spatial behavioral questionnaire. *The Journal of General Psychology*, *134*(3), 361.
Erskine, M. A., Gregg, D. G., Karimi, J., & Scott, J. E. (2015). Geospatial reasoning ability: Definition, measurement, and validation. *International Journal of Human-Computer Interaction*, *31*(6), 402–412.
Gagnier, K. M., Atit, K., Ormand, C. J., & Shipley, T. F. (2017). Comprehending 3D diagrams: Sketching to support spatial reasoning. *Topics in Cognitive Science*, *9*(4), 883–901.
Gagnier, K. M., Holochwost, S. J., & Fisher, K. R. (2022). Spatial thinking in science, technology, engineering, and mathematics: Elementary teachers' beliefs, perceptions, and self-efficacy. *Journal of Research in Science Teaching*, *59*(1), 95–126. https://doi.org/10.1002/tea.21722.
Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, *30*(5), 425–447.
Huynh, N. T., & Sharpe, B. (2013). An assessment instrument to measure geospatial thinking expertise. *Journal of Geography*, *112*(1), 3–17.
Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., & Sageman, B. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, *14*(2), 175–187.
Joint Committee of the AERA, APA and NCME. (2014). *Standards for educational and psychological testing*. the authors.
Kim, M., & Bednarz, R. (2013). Development of critical spatial thinking through GIS learning. *Journal of Geography in Higher Education*, *37*(3), 350–366. https://doi.org/10.1080/03098265.2013.769091.
Kozhevnikov, M., Motes, M. A., & Hegarty, M. (2007). Spatial visualization in physics problem solving. *Cognitive Science*, *31*(4), 549–579.
Lee, J., & Bednarz, R. (2012). Components of spatial thinking: Evidence from a spatial thinking ability test. *Journal of Geography*, *111*(1), 15–26. https://doi.org/10.1080/00221341.2011.583262.
Lee, J., & Jo, I. (2022). Assessing spatial skills/thinking in Geography. In T. Bourke, R. Mills, & R. Lane (Eds.), *Assessment in geographical education: An international perspective. Key challenges in Geography*. Springer. https://doi.org.gate.lib.buffalo.edu/10.1007/978-3-030-95139-9_4.
Linacre, J. M. (2023). A User's Guide to WinSteps. http://winsteps.com.
Liu, X. (2020). *Using and developing measurement instruments in science education: A rasch modeling approach*. IAP Press.
Liu, R., Greene, R., Li, X., Wang, T., Lu, M., & Xu, Y. (2019). Comparing geoinformation and geography students' spatial thinking skills with a human-geography pedagogical approach in a Chinese context. *Sustainability*, *11*(20). https://doi.org/10.3390/su11205573.
Lohman, D. F. (1996). Spatial ability and G. In I. Dennis, & P. Tapsfield (Eds.), *), human abilities: Their nature and assessment* (pp. 97–116). Erlbau.
Mair, P., Much, T., Hatzinger, R., Maier, M., & Debelak, R. (2023). Package 'eRm'. https://cran.r-project.org/web/packages/eRm/eRm.pdf.
Mix, K. S. (2019). Why are spatial skill and mathematics related? *Child Development Perspectives*, *13*(2), 121–126.
Muschelli, J. (2022). Package 'diffr'. https://cran.r-project.org/web/packages/diffr/diffr.pdf.
National Research Council. (2006). *Learning to think spatially*. The National Academies Press. https://doi.org/10.17226/11019.
Nazareth, A., Newcombe, N. S., Shipley, T. F., Velazquez, M., & Weisberg, S. M. (2019). Beyond small-scale spatial skills: Navigation skills and geoscience education. *Cognitive Research: Principles and Implications*, *4*, 1–17.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedogogiske Institute. Chicago: University of Chicago.
Revelle, W. (2023). Package 'psych'. https://cran.r-project.org/web/packages/psych/psych.pdf.
Robitzsch, A., Kiefer, T., & Wu, M. (2023). Package 'TAM'. https://cran.r-project.org/web/packages/TAM/TAM.pdf.
Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, *93*(3), 604–614.
Tomaszewski, B., Vodacek, A., Parody, R., & Holt, N. (2015). Spatial thinking ability assessment in Rwandan secondary schools: Baseline results. *Journal of Geography*, *114*(2), 39–48. https://doi.org/10.1080/00221341.2014.918165.
Turgut, M. (2015). Development of the spatial ability self-report scale (SASRS): Reliability and validity studies. *Quality & Quantity*, *49*, 1997–2014.
van der Ham, I. J., Claessen, M. H., Evers, A. W., & van der Kuil, M. N. (2020). Large-scale assessment of human navigation ability across the lifespan. *Scientific Reports*, *10*(1), 3299.
Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2017). I. spatial skills, their development, and their links to mathematics. *Monographs of the Society for Research in Child Development*, *82*(1), 7–30.
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.
Wolfgang, C., Stannard, L., & Jones, I. (2003). Advanced constructional play with LEGOs among preschoolers as a predictor of later school achievement in mathematics. *Early Child Development and Care*, *173*(5), 467–475.
Wright, B. D., & Master, G. N. (1982). *Rating scale analysis*. Mesa Press.

Publisher's Note

Kannaki Thayaseelan is a Ph.D. student in the Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York.

Yanfang Zhai is a visiting scholar in Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York, and also a Ph.D.student in the School of Education, Capital Normal University, Beijing.

Siqi Li is a lecture in College of Education for the Future, Beijing Normal University, Zhuhai.

Xiufeng Liu is a Distinguished Professor in Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York.