Pertanika J. Soc. Sci. & Hum. 32 (4): 1433 - 1450 (2024)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Measuring Spatial Ability: Analysis of Spatial Ability Test for Gulf State Students Using Item Response Theory

Mohammed Al Ajmi[1]*, Siti Salina Mustakim[1], Samsilah Roslan[1] and Rashid Almehrizi[2]

[1]Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]College of Education, Sultan Qaboos University, Alhouz, Muscat, Oman

ABSTRACT

This study evaluates the psychometric properties of the spatial ability test using the three-parameter logistic model within item response theory. The final version of the scale comprised 29 dichotomous items, administered to a sample of 2,694 male and female students from grades 5 and 6 across schools in the Arab Gulf region. The test adhered to the three-parameter model, satisfying the assumptions of unidimensionality and local independence. The item difficulty parameters ranged from -1.541 to 1.735, discrimination parameters spanned from 0.419 to 5.252, and guessing parameters varied between 0.00 and 0.346. With a marginal reliability coefficient of 0.86, the scale demonstrated strong stability. These findings indicate that the test items align with established measurement principles, supporting the spatial ability test as a valid and reliable assessment tool for measuring spatial abilities in the Gulf region. The results have important implications for educational assessment in the Arab Gulf and could guide the development of similar assessments in other educational contexts. Further research is recommended to improve the test's precision and explore its application in diverse educational settings.

Keywords: Item response theory, psychometric characteristics, reliability, spatial ability, three-parameter model

ARTICLE INFO

Article history:
Received: 21 October 2023
Accepted: 30 July 2024
Published: 02 December 2024

DOI: https://doi.org/10.47836/pjssh.32.4.10

E-mail addresses:
mohd7010@gmail.com (Mohammed Al Ajmi)
mssalina@upm.edu.my (Siti Salina Mustakim)
samsilah@upm.edu.my (Samsilah Roslan)
mehrzi@squ.edu.om (Rashid Almehrizi)
*Corresponding author

INTRODUCTION

Cognitive abilities encompass the mental processes and skills that allow individuals to gather, interpret, comprehend, and apply information. These abilities cover a range of thinking functions, including attention, memory, reasoning, problem-solving, and specific verbal, numerical, and spatial competencies. Cognitive ability is known in the literature by several names, such

ISSN: 0128-7702
e-ISSN: 2231-8534

as general mental ability and general intelligence (Carroll, 1993; Deary, 2020; Hunt, 2011; Jensen, 1998; Salgado, 2002). These abilities are foundational to learning and performing various tasks, influencing how effectively individuals can process and apply information in different contexts. Cognitive abilities significantly impact how individuals acquire and process information efficiently (Wong et al., 2023). Enhanced cognitive abilities enable quicker comprehension and more effective problem-solving, which are essential in academic and everyday settings.

Numerous studies have consistently demonstrated a positive correlation between cognitive abilities and academic achievement across a variety of subjects, including mathematics, reading, and science (Li et al., 2022). Students with stronger cognitive abilities tend to excel academically as their enhanced information processing, comprehension, and application skills contribute to better learning outcomes. Recognizing students' cognitive strengths and weaknesses allows educators to provide targeted interventions and individualized instruction, optimizing their learning potential and fostering academic achievement (Keenan & Meenan, 2014). This personalized approach helps maximize students' academic performance by addressing their specific needs and leveraging their cognitive strengths.

Spatial ability is regarded as one of the key cognitive skills in mathematics, attracting significant interest from educators and specialists in curriculum development, particularly in relation to mathematics and its teaching methodologies. Its active role is increasing through what mathematics relies on for the primary stage in solving the issue and learning the relationships and geometric shapes. Thurston defines it as the ability to visualize shapes and perceive their relationships. This ability appears in mental activity that depends on visualizing objects without changing their spatial position (Suleiman, 2010). In educational curricula, spatial ability is crucial for understanding and solving geometry problems, visualizing mathematical concepts, and interpreting data from graphs and charts. Incorporating spatial reasoning tasks in the curriculum helps students develop these skills, which are crucial for success in mathematics and other related disciplines (Uttal et al., 2013).

Studies have shown that individuals with strong spatial abilities often excel in geometry, physics, and engineering, where understanding spatial concepts and visualizing three-dimensional objects are essential (Uttal et al., 2013). These disciplines require the ability to mentally manipulate shapes and visualize spatial relationships, which are well-developed skills in individuals with strong spatial abilities. For example, in physics, students need to visualize forces and motions in three-dimensional space, while in engineering, they must design and interpret complex structures. By recognizing and nurturing students' spatial abilities, educators can employ teaching strategies emphasizing visualizations, hands-on activities, and spatial reasoning tasks to enhance their

understanding and achievement in spatially related disciplines. Also, research confirms that individuals with strong spatial abilities often perform well in various fields, such as academic achievement in mathematics and arithmetic, such as studying (Hallowell & Okamoto, 2015; Verdine, 2011; Weckbacher & Okamoto, 2014).

Due to the significance of spatial ability, the National Council of Teachers of Mathematics in the United States of America has recommended that educational programs from kindergarten to twelfth grade enable students to develop their spatial abilities through geometric content. This includes identifying locations, describing spatial relationships, spatial visualization, spatial reasoning, and geometric models to solve problems (National Council of Teachers of Mathematics, 2023). The council emphasizes that developing spatial skills is essential for students' mathematical understanding and problem-solving abilities. Integrating spatial reasoning into the curriculum helps students better grasp complex mathematical concepts and apply them in various contexts, thus enhancing their overall cognitive development and preparing them for advanced studies and careers in STEM fields.

Numerous research studies emphasize the significance of evaluating spatial ability in gifted individuals, suggesting that talent searches could enhance their selection criteria by incorporating spatial ability measures. This approach would broaden the scope of identifying intellectually capable youth, offering them educational experiences in civil engineering, aviation, and mechanical sciences (Wai et al., 2009). Including spatial ability in gifted searches ensures that students with exceptional spatial skills are recognized and given opportunities to excel in areas where these abilities are crucial. This helps identify a more diverse group of gifted students and provides them with the resources and support needed to develop their unique talents further. (Lohman, 2005; Shea et al., 2001).

Understanding and measuring spatial ability is crucial, as it is associated with performance in STEM, where spatial reasoning and visualization skills are essential (Uttal et al., 2013). Students with strong spatial abilities can better understand complex scientific concepts, visualize engineering designs, and interpret data from graphs and models. Measuring spatial ability is important for assessing individuals' cognitive strengths and weaknesses in this domain. By identifying students with strong spatial abilities, educators can tailor instruction and provide enrichment opportunities that foster their skills and interest in STEM, ultimately contributing to their success and innovation in these areas. One commonly used spatial ability test is the Gulf Scale of Mental Abilities (GMMAS), developed by Alzayat et al. (2011).

The development of the scale is grounded on Thurstone's theory of Primary Mental Abilities, which is a significant and influential concept in the field of psychology. Developed by the American psychologist Louis Leon Thurstone in the

mid-20th century, this theory challenges the notion of a singular, unitary concept of intelligence. Instead, it proposes that intelligence can be broken down into several distinct "primary mental abilities," each representing a specific facet of cognitive function (Thurstone theory of intelligence, 2023). These primary abilities encompass a wide range of skills, including mathematical reasoning, verbal comprehension, memory, spatial visualization, and perceptual speed. Thurstone's work demonstrated that these mental abilities are relatively independent and can be measured separately, providing a more nuanced understanding of human cognitive functioning. This theory has had a lasting impact on the study of intelligence and continues to influence the field of psychology today (Gill et al., 2020). Understanding these primary abilities allows for a more detailed assessment of cognitive strengths and weaknesses, leading to more targeted and effective educational and psychological interventions.

The Gulf Scale of Mental Abilities (GMMAS) assesses various cognitive abilities, including spatial visualization. It provides a standardized and reliable measure to evaluate an individual's spatial aptitude and compare it to a normative sample. Therefore, when designing scales, especially measures of cognitive and mental abilities, one of the two famous measurement methods is used: Item Response theory (IRT) and classical Test Theory (CTT), which was widely used in the twentieth century to try to avoid shortcomings in the test instrument (Jabrayilov et al., 2016).

The IRT and CTT are two foundational educational and psychological measurement approaches. CTT emphasizes the total test score, assuming all test items contribute equally to the final score. It views the observed score as a blend of the true score and random error but lacks the capacity to assess individual item characteristics. In contrast, IRT delves deeper into item-level analysis, proposing that the likelihood of correctly answering an item depends on the examiners' latent ability and specific item attributes like difficulty, discrimination, and guessing. This framework allows a more detailed understanding of how each item influences the measurement process. Different IRT models, such as the one-parameter (1PL), two-parameter (2PL), and three-parameter (3PL) logistic models, provide a more accurate estimation of a person's ability and offer richer insights into item performance (Embretson & Reise, 2000).

The main benefit of IRT compared to CTT is its capability to offer detailed insights into how items perform across different levels of ability, making it more effective in developing adaptive tests and providing accurate measurement across a wider range of abilities. Modern models in educational measurement, including IRT, offer more accurate indicators of item difficulty and discrimination when analyzed statistically, making them superior to CTT, which mainly depends on the overall score (Subali et al., 2021).

Given the reliance on classical measurement theory in the establishment and

standardization of psychological tests and scales used within the field of Humanities, particularly in the Arab and Gulf regions, and as a result of the emergence of some disadvantages associated with this theory, the idea came to use one of the modern models in measurement in order to know the most important psychometric characteristics achieved by one of the scales that was built in the light of classical theory.

Some students may get results that do not express their real abilities in a test. The reason for this may be due to a defect in the tool applied to them, whether it is in the low truthfulness and stability of this tool or it is in the difficulty or ease of paragraphs of this test, which may be higher or lower than their ability level, giving inaccurate results about the level of real ability that students possess. Despite the advantages of IRT, which would improve the accuracy of the scale results, the scale's psychometric properties have yet to be studied according to this theory. Therefore, the current research seeks to study psychometric properties of spatial ability in the Gulf Test of Mental Abilities (GMMES) based on the three-parameter model.

Therefore, the current research attempts to answer:

1. What psychometric characteristics are available in the Spatial Ability Test and its items according to the IRT?

Four questions arise from it:

1. To what extent are the assumptions of the IRT fulfilled in the spatial ability test data?
2. What is the suitability of the 3PL for the spatial ability test data?
3. What is the estimate of the three parameters of the items considering the 3PL?
4. How much information does the test provide at different ability levels?

METHODS

Participants

This study adopts a descriptive approach to examine the statistical features of the spatial ability test within GMMAS. The researcher utilizes secondary data from the GMMAS standardization conducted by the Arab Office for the Gulf States in 2011. The sample consists of fifth—and sixth-grade students aged between 9 and 12. The total sample includes 2,694 students, with 1,273 females and 1,416 males, all within the specified age range.

Developing the Item Bank for Spatial Ability

Measure

The research employs the spatial ability test from GMMAS, as developed by Alzayat et al. (2011). This evaluation comprises three independent exams designed to measure verbal, numerical, and spatial abilities.

The current investigation centers on the assessment dedicated to spatial ability, which consists of 30 multiple-choice questions. Spatial ability is measured by shape completion test (10 items), paper bending and unfolding test (10 items), and rotation test figure (10 items). In this evaluation, the right response is awarded one score, while an incorrect answer is

assigned a score of zero. Consequently, the cumulative score can range from 0 to 30.

It was confirmed through concurrent validity by knowing the correlation between the spatial ability test and the colored Raven matrix with the students of Kuwait. The value of the correlation between them for the fifth and sixth grades was 0.52 and 0.49, and it was statistically significant, which indicates the concurrent validity of the spatial test. Also, we verified the test's predictive validity by evaluating the correlation values between spatial ability and mathematical achievement across all Kuwaiti grade levels,. At the 0.05 level of significance, the 0.163 value of the correlation values between spatial aptitude and arithmetic performance in fifth grade is significant. At the 0.05 level of significance, the value of the correlation values between spatial ability and mathematical performance in sixth grade is 0.178. Although the sample sizes in each discipline are relatively small, these values are still found to be significant.

The spatial ability test demonstrated strong reliability with a test-retest coefficient of 0.85. Internal consistency remained consistently high for spatial ability across all grade levels, as reflected in the Cronbach alpha coefficients, which ranged from 0.80 to 0.82 in the context of Gulf countries (Alzayat et al., 2011).

Item Response Theory assumptions

Unidimensionality

IRT is built on some assumptions that the researcher should verify before using it. The first of these assumptions is unidimensionality, a fundamental assumption in IRT, which posits that the test items measure a single, dominant latent trait. In simpler terms, the test items are all related to a common underlying construct or ability and do not tap into multiple unrelated dimensions. The unidimensionality assumption is crucial in IRT because the precision of item parameter estimates and the authenticity of interpreting test scores largely depend on it. The results may be confounded and less interpretable if the test is not unidimensional.

Above all, the unidimensional model would be checked by confirmatory and exploratory confirmatory factor analyses were used to verify a unidimensional assumption. Two conditions must be met to establish unidimensionality during exploratory analysis. First, Reckase (1979) states that the dominant component should explain at least 20% of the variance. Second, Reeve et al. (2007) emphasized that the variance of the first factor should be at least four times that of the second factor. For the confirmatory factor, the following two indicators have been used: the Root Mean Square of Residuals (RMSEA) according to the specified criteria by Edelen and Reeve (2007) and Smits et al. (2011), which indicates a good fit when the RMSEA is 0.08 or less, and the Tanaka Index (GFI) which the criterion for a good fit value is 0.90, according to Tanaka and Huba (1985).

Local Independence (LI)

The second premise, LI, requires that item answers be independent. In other words,

once an examiner's latent trait level is known, their answer to one item should not predict their response to another beyond what is expected given the latent trait (Embretson & Reise, 2000). To assess this assumption, the researcher used a statistical measure proposed by Yen (1993), which calculates the correlation coefficient between the residuals of item pairs after adjusting for the individual's ability. The Local Dependence Indices for Dichotomous Items (LDID) software was utilized to test local independence in the spatial ability test. Typically, a critical threshold of 0.2 for the absolute value of Q3 is used as a benchmark (Chen & Thissen, 1997).

IRT Model Comparison

Once the item bank fulfills the unidimensionality and local independence assumptions, an appropriate IRT model must be chosen for parameter estimation. IRT models are selected based on the characteristics of the test items and the nature of the data. According to Jabrayilov et al. (2016) and Baker and Kim (2017), the suitability of different IRT models varies depending on the types of tests and their specific requirements.

One-Parameter Logistics Model (1PL or Rasch Model): This type assumes all items have equal discrimination and only estimates the difficulty parameter for each item. It is suitable for tests where all items are assumed to have similar discriminatory power (Uniform Discrimination Tests). It is common in educational assessments where items are designed to be equally challenging across different difficulty levels. Also, it requires smaller sample sizes compared to more complex models, making it suitable for pilot studies or small-scale assessments.

Two-Parameter Logistics Model (2PL): Estimates each item's difficulty and discrimination parameters. Allows items to vary in how well they discriminate between individuals with different latent trait levels. This type is suitable for tests where items have varying abilities to discriminate between individuals (Variable Discrimination Tests). It is useful in tests covering a broad range of difficulties and is designed to distinguish between different ability levels. It is suitable for psychological tests where different items may have different levels of effectiveness in measuring the latent trait.

Three-Parameter Logistics Model (3PL): Estimates difficulty, discrimination, and guessing parameters and accounts for the possibility that some respondents may guess the correct answer. This type is suitable for multiple-choice tests where guessing can influence responses. It adjusts for the probability that a low-ability test-taker might guess an answer correctly. Also, it is used in assessments where the probability of guessing must be accounted for (Complex Assessments), such as certain types of aptitude or intelligence tests.

To determine the most suitable IRT model and assess its accuracy, four commonly applied model fit indices were used: the Akaike Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978), Root Mean Square

Error of Estimates (RMSE), and the average information value. These metrics were utilized to evaluate how well each model fits the data, aiding in selecting the best-fitting model. Lower values for AIC, BIC, and RMSE suggest a better model fit. Model comparison and selection were conducted using the mirt R package (Chalmers, 2012) and BILOG-MG software.

Item Parameters Estimate

In IRT, key item parameters such as difficulty, discrimination, and guessing are determined through statistical models that map the relationship between an individual's latent ability and the probability of answering a given item correctly. The difficulty parameter (b) represents the point on the latent trait continuum where an individual has a 50% likelihood of providing the correct response. It is estimated using the Item Characteristic Curve (ICC), which shows how the probability of a correct answer changes with different levels of the latent trait (θ). The b-parameter is identified at the point where this probability reaches 0.50.

The discrimination parameter (a) reflects the effectiveness of an item in distinguishing between individuals with varying levels of the latent trait. A higher value of the discrimination parameter signifies that the item is more adept at differentiating between examinees whose abilities are closely matched—the estimation process of discrimination parameter (a) by Slope of ICC. The discrimination parameter is the slope of the ICC at the point of inflection (where the probability of a correct response is 50%). The guessing parameter (c) represents the probability that an individual with a very low latent trait level will correctly guess the answer to an item. This parameter is particularly relevant for multiple-choice items where guessing can play a significant role—the estimation process of guessing parameter (c) by the Lower Asymptote of ICC. The guessing parameter is the lower asymptote of the ICC, indicating the probability of a correct response due to guessing (Baker & Kim, 2017; Jabrayilov et al., 2016).

This research utilized the mirt package (Version 1.24) in R to estimate item parameters. The software uses the Expectation A Posteriori (EAP) approach, which uses Bayesian estimation methods. Through the application of the three-parameter logistic model (3PL), the analysis provided estimates for item difficulty, discrimination, and the pseudo-guessing parameter.

Reliability

In the realm of IRT, marginal reliability emerges as a valuable metric for evaluating the reliability of test scores. This measure revolves around estimating reliability depending on the marginal distribution of the test scores, involving the analysis of item parameters and associated standard errors. Marginal reliability in IRT reflects the degree to which test scores are resilient to measurement error. Higher reliability coefficients indicate greater accuracy and consistency in the test scores, signifying that the assessment yields more reliable and precise measurements with minimal

error. A marginal reliability coefficient of 1 signifies perfect reliability, while a value closer to 0 suggests diminished reliability due to increased measurement error. It has significant implications for test interpretation, as it enables educators and psychologists to understand the strengths and limitations of test scores across the ability spectrum, leading to more targeted and effective interventions and support for test-takers (Embretson & Reise, 2000).

FINDINGS AND DISCUSSION

Psychometric Evaluation of the Spatial Ability Item Bank

Unidimensionality

To confirm the unidimensionality assumption of the test, the adequacy of the sample size was evaluated using the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests. The analysis yielded a chi-square value of 10,208.063 with a significance level of 0.001 and 435 degrees of freedom, indicating that the sample size was sufficient for performing exploratory factor analysis. Subsequently, the analysis was conducted on the principal components of the correlation matrix for the 30 spatial ability items.

The results revealed four latent factors with eigenvalues exceeding one, collectively explaining 42.76% of the variance. The ratio of the eigenvalue of the first factor (4.70) to the second factor (1.82) was 2.58, surpassing the value of two, which supports unidimensionality as suggested by Reckase (1997, cited in Matarneh and Oalla, 2018). Moreover, the first factor accounted for 37.26% of the total variance, satisfying Reckase's recommended 20% threshold for a unidimensional test.

Furthermore, Cattell's scree plot (1966) for the 30-item factor analysis confirmed the test's unidimensionality, as the first factor was clearly distinct from the remaining factors (Figure 1).

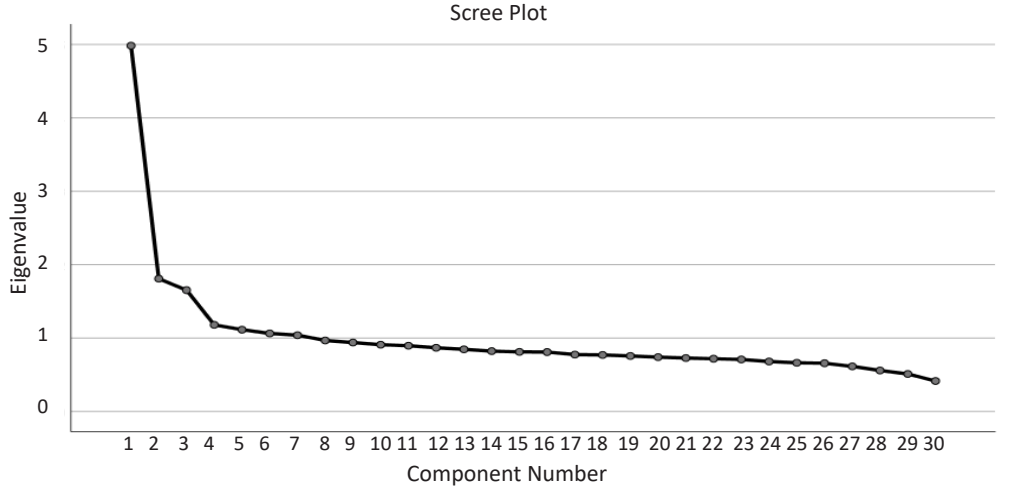

Figure 1. Factor scree plots from principal component analysis of 30 item
Source: Authors' work

Confirmatory Factor Analysis (CFA) was conducted using the AMOS software to compute the Root Mean Square Error of Approximation (RMSEA) and the Goodness of Fit Index (GFI). The factor loadings of the observed variables on a single latent parameter, along with the residual error values, are shown in the CFA results. The analysis produced an RMSEA value of 0.054, which meets the standards set by Edelen and Reeve (2007) and Smits et al. (2011). Additionally, the GFI was calculated at 0.90, indicating a good model fit (Figure 2).

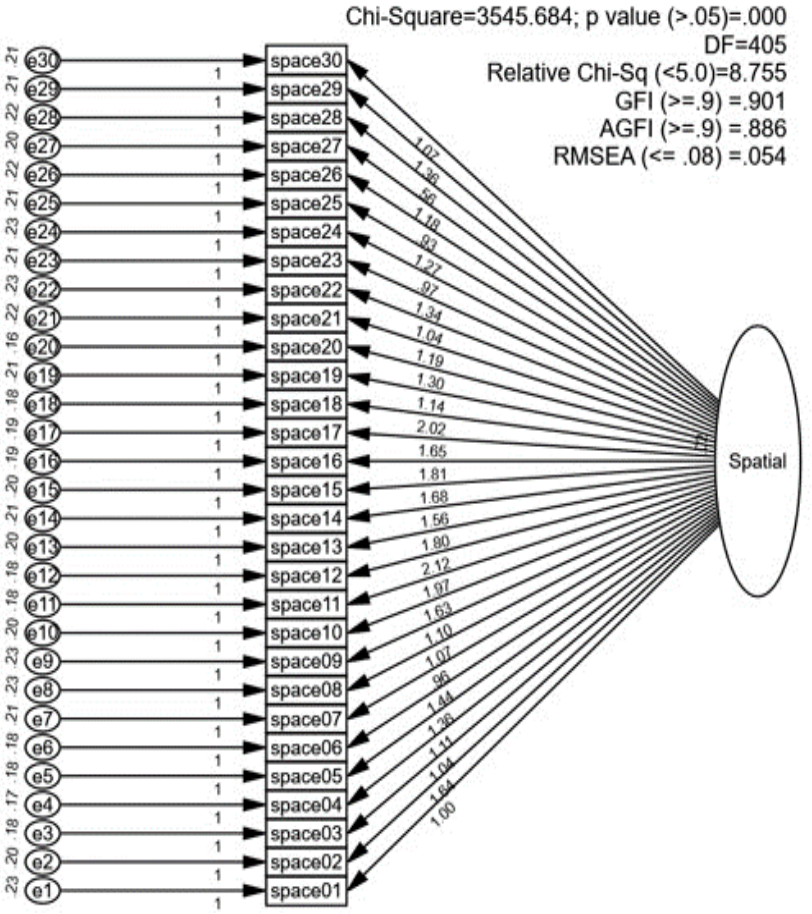

Figure 2. Confirmatory factor analysis for spatial ability
Source: Authors' work

Local Independence

In the 3PL model, local independence was assessed using Q3 statistics. Table 1 summarizes the Q3 values for the test.

The results demonstrate that the mean Q3 value is 0.045, significantly lower than the critical threshold of 0.2. Furthermore, 99.8% of item pairs in the spatial test met the local independence criteria. Only one

Table 1
Indicators of local independence according to Item Response Theory

Ability	No. of test items	items pair	Maximum	Minimum	Mean of Q3
Spatial	30	435	0.44	0.0002	0.045

Source: Authors' work

pair of items (items 2 and 10) exceeded the threshold with a value of 0.44, while all other pairs had values below 0.198. These findings support the notion that the spatial ability test items are largely locally independent.

IRT Model Comparison

We find a compilation of the model fit indices, aiding us in selecting the optimal model for the spatial ability test data (Table 2).

Table 3 clearly indicates that 3PL, which accounts for difficulty, discrimination, and guessing parameters, is the fit model for the spatial test data (Table 3).

Table 2
The indicator values used to select the appropriate model for the spatial ability test data

S	Indicators	Model		
		1PL	2PL	3PL
1	AIC	100088.91	99490.96	99168.61
2	BIC	100271.82	99844.87	99699.47
3	Average Test Information	3.970	4.621	5.678
4	RMSE	0.4524	0.4285	0.4626
5	Reliability Index	0.799	0.822	0.850

Note. IPL, one parameter logarithmic model; 2PL, two-parameter logarithmic model; 3PL, three-parameter logarithmic mode; AIC, Akaike' information criterion; BIC, Bayesian information criterion; RMSE, root mean Square Error
Source: Authors' work

Item Parameters Estimate

The item difficulty parameters range from -1.541 for item 3 to 1.735 for item 19. The average difficulty is 0.450, with a standard deviation of 1.038, indicating that most test items fall within a moderate difficulty level (Table 3). The Item Characteristic Curves for item 3 (the least difficult) and item 19 (the most difficult) are displayed in Figure 3.

Table 3 depicts details, showcasing the item discrimination parameters spanning from 0.419 to 5.252 for items 28 and 23, respectively. Moreover, the mean item discrimination parameter stands at 1.554 with a standard deviation of 1.308, signifying the highest discrimination value. Item Characteristic Curves for item 28, which has the lowest discrimination value, and item 23, which achieved the highest discrimination value, are presented in Figure 4.

Table 3
Item statistics based on the 3PL model

3PLM									
Item	a	b	c	IIC	Item	a	b	c	IIC
1	0.435	0.0202	0.012	0.046	16	0.805	0.8193	0.000	0.162
2	0.859	-0.299	0.000	0.184	17	0.807	1.2152	0.019	0.157
3	0.665	-1.541	0.000	0.11	18	0.973	0.6957	0.001	0.236
4	0.726	-1.471	0.005	0.13	19	0.831	1.7345	0.139	0.132
5	0.856	-1.021	0.000	0.183	20	1.058	1.6001	0.042	0.257
6	0.89	-0.919	0.000	0.198	21	4.241	1.4076	0.325	2.269
7	0.431	-1.302	0.006	0.046	22	3.558	1.3559	0.346	1.586
8	0.534	-0.307	0.000	0.071	23	5.252	1.3257	0.277	4.044
9	0.514	0.0304	0.004	0.066	24	2.022	1.5692	0.31	0.559
10	0.818	-0.172	0.000	0.167	25	4.001	1.3561	0.253	2.445
11	1.397	-0.08	0.08	0.417	26	2.681	1.5003	0.284	1.039
12	1.264	-0.012	0.000	0.399	27	2.477	1.4654	0.202	1.04
13	0.954	0.2015	0.047	0.207	28	0.419	1.4683	0.005	0.043
14	0.737	0.2084	0.000	0.136	29	2.746	1.3301	0.258	1.144
15	0.805	-0.238	0.000	0.162	30	2.865	1.5642	0.247	1.26

Note. a: Discrimination parameter; b: Difficulty parameter, IIC: Maximum item information curve
Source: Authors' work

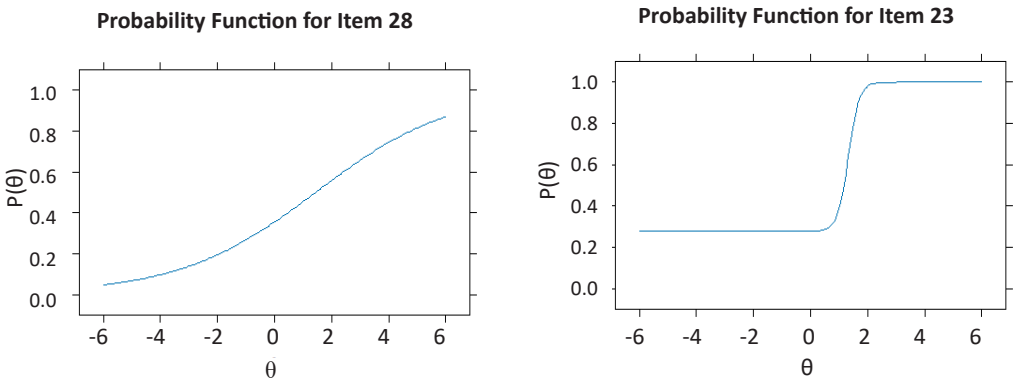

Figure 4. Item characteristic curves for items 28 and 23
Source: Authors' work

The guessing parameters for the items range from 0.000 for item 10 to a notable 0.346 for item 21. The average guessing parameter is 0.096, with a standard deviation of 0.127, indicating minimal dependence on guessing when responding to the test items. These results suggest that examinees seldom employed guessing strategies (Table 3). Furthermore, the characteristic curves for item 10, with the lowest guessing value, and item 21, with the highest guessing value, are displayed in Figure 5.

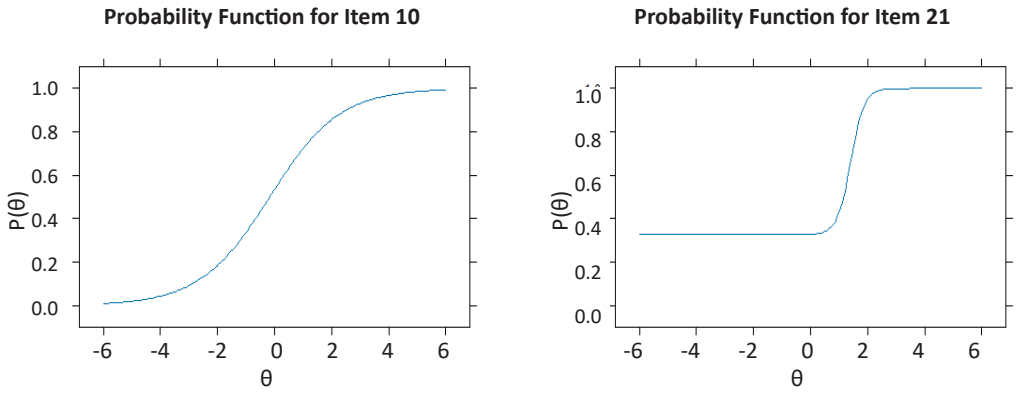

Figure 5. Item characteristic curves for items 10 and 21
Source: Authors' work

We discern that the test items offer varying degrees of information, with values ranging from 0.043 to 4.044. Item 28 yields the least amount of information, starkly contrasting with item 23, which presents the highest information content (Table 3). To grasp these insights visually, the item information curves for items 28 and 23 (Figure 6).

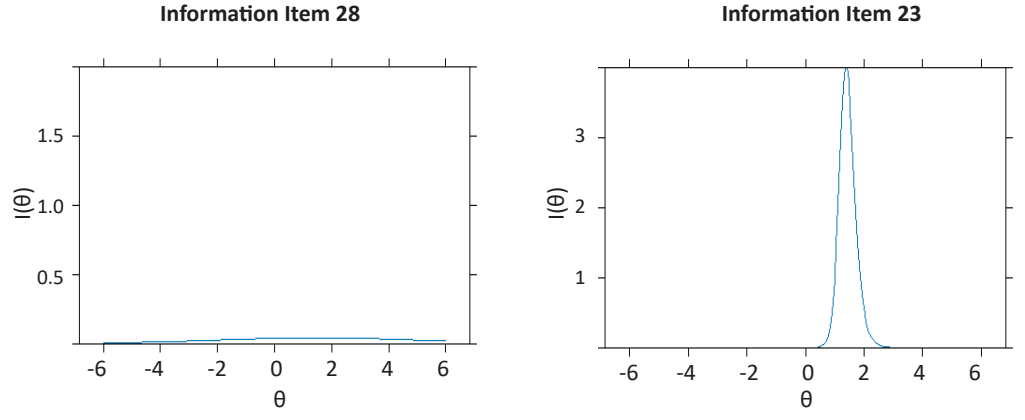

Figure 6. Item information curve of the items 28 and 23
Source: Authors' work

Reliability

The spatial ability test is highly reliable in precisely estimating individuals' abilities, as seen by its obtained marginal reliability of 0.86.

DISCUSSION

This study aims to assess the spatial ability test's psychometric qualities. The Gulf version was given to 2694 students from six nations in a random sample. The test's validity and measurement accuracy

depend heavily on the assessment of IRT presumptions. Within the framework of a spatial ability test, this study investigates two fundamental assumptions of IRT: unidimensionality and local independence. The results indicated that the spatial ability test adhered to the unidimensionality assumption. The data supported the presence of a single underlying factor influencing the test responses, which validates the use of unidimensional IRT models for this test, ensuring that the spatial ability construct is consistently measured across all items. As for the second assumption, local independence, the results showed that this assumption was generally met across the test items. However, there was one exception: a particular item exhibited dependency on other items, violating the local independence assumption. The problematic item was removed from the analysis to maintain the integrity of the IRT model. This dependency could stem from factors such as similar content or overlapping skills required to answer the items. Identifying and addressing such violations is crucial because they can lead to biased estimates of item parameters and latent traits.

The results presented in this investigation unveil the superior performance of the 3PL over the one-parameter logarithmic model (1PL) and the two-parameter logarithmic model (2PL) in evaluating the spatial ability test. This advantage of the 3PL model can be attributed to the multiple-choice format of the test questions, which, as asserted by Haladyna and Downing (2004), is widely employed in educational institutions. It is possible to estimate examinees' ability parameters accurately by adopting a model that takes into account the three parameters of difficulty, discrimination, and guessing. Incorporating the guessing parameter in the 3PL model effectively accounts for this behavior, leading to a better fit for the data. This observation aligns with earlier research conducted by Fu (2010) and Gao (2011).

The researcher highlights key findings from calibrating items in the spatial ability test, notably the variation in item difficulty levels. Despite this variation, the average difficulty parameter (0.450) suggests that the items, overall, fall within the medium difficulty range. This result carries important implications for both test design and test-taker performance. The test comprehensively evaluates test-taker abilities by incorporating a balanced mix of easy, medium, and difficult items. It effectively challenges fifth and sixth-grade students by presenting a well-balanced array of item difficulties. This outcome is consistent with measurement theory, which underscores the necessity of including items with varying difficulty levels to accurately assess test-takers abilities.

The analysis of the item discrimination parameter reveals that most test items show high discrimination. High discrimination values indicate that these items effectively distinguish between test-takers with varying ability levels, which is crucial for the precision and accuracy of the assessment. This desirable characteristic enhances the validity of the test by ensuring that it accurately reflects differences in abilities.

The findings suggest that the test items successfully differentiate between higher- and lower-performing students, thereby improving the overall reliability and effectiveness of the assessment.

Additionally, the low values of the guessing parameter indicate that test-takers are not significantly reliant on random guessing while responding to the items. It indicates that the items are carefully constructed to minimize the likelihood of guessing, thereby allowing the test to more accurately capture test-takers true abilities. It contributes to the overall validity and reliability of the test.

Examining the item information curves shows a considerable variation in the information provided by the spatial ability test items, ranging from 0.043 to 4.044. This variation underscores the items' capacity to effectively distinguish between individuals with varying latent trait levels. The diverse range of information values delivers key insights into the precision and discriminatory power of the test across different ability levels. Such data enables test developers and researchers to pinpoint items that offer the most meaningful information, assess the test's overall measurement accuracy, and make informed decisions regarding item selection, refinement, or elimination to optimize the test's reliability and effectiveness. Additionally, items with high information values at specific ability levels can be strategically targeted to improve the test's sensitivity within those ranges. This process leads to a more balanced and diagnostically effective tool capable of accurately assessing spatial abilities across diverse individuals.

CONCLUSION

This research explored the psychometric characteristics of the Spatial Ability Test using item response theory. The findings affirm that the test items are effectively designed, displaying moderate difficulty levels, strong discriminatory power, and limited dependence on guessing. These results emphasize careful construction and precision embedded in the test's development.

The study holds several theoretical implications as it makes a significant contribution to our comprehension of IRT and its application in assessments and measurements within the Gulf countries. Moreover, it advances our understanding of methodologies for evaluating cognitive abilities, ultimately guiding us toward enhancing assessment and measurement procedures. From a practical perspective, this study can serve as a blueprint for enhancing the quality of the spatial ability test for fifth and sixth-grade students by furnishing actionable recommendations for refining and enhancing the test items. Furthermore, it significantly contributes to improving the accuracy of estimating students' spatial abilities by utilizing IRT and focusing on the psychological attributes of the test.

The study's findings offer several practical applications for enhancing spatial ability assessments. By identifying high-quality test items with strong discriminatory

power, developers can refine the item pool, ensuring a more accurate and reliable measure of spatial ability. Educators can use these refined tools to create targeted intervention programs, addressing specific strengths and weaknesses in students. It leads to personalized learning approaches that optimize student outcomes in spatially demanding subjects like mathematics and engineering. Additionally, the insights from IRT facilitate the transition to computerized adaptive testing (CAT), which tailors the test to each individual's ability level, making the assessment more efficient and reducing test-taking time. CAT also conserves resources, allowing for more frequent and less burdensome testing. Furthermore, these refined assessments can guide curriculum development and educational policy, ensuring that programs support the development of spatial skills. The study's findings can significantly improve spatial ability assessments' accuracy, efficiency, and effectiveness by leveraging these practical applications.

While the current findings generally provide positive indications for the psychometric characteristics of the spatial ability test, it is essential to acknowledge several limitations. First, the findings presented in this research relied on a smaller set of 29 items compared to the larger item banks commonly employed in IRT. Expanding the item bank could further enhance the benefits of administering the spatial ability test, offering greater precision and flexibility in assessing a wider range of abilities. Second, the conclusions drawn in the current study relied exclusively on a dichotomous item response model, which involves binary outcomes (correct or incorrect). To gain deeper insights, future investigations on the spatial ability test could benefit from a comparative analysis involving both dichotomous and polytomous IRT models. In light of these limitations, the study suggests that future research should promptly expand the spatial ability test's item repository, encompassing a multitude of items, potentially reaching into the hundreds, to cover the entire spectrum of abilities from -4 to +4. Additionally, incorporating a diversity of test questions, including multiple-choice and open-ended questions, would allow for measuring spatial ability according to multiple IRT models.

ACKNOWLEDGMENTS

We thank the Arab Bureau of the Gulf States for approving the GMMAS scale and benefiting from the data collected. We also thank those concerned with the Universiti Putra Malaysia for their support.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control IEEE Trans, 19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alzayat, F., Almahrazi, R., Arshad, A., Fathi, K., Albaili, M., dogan, A., Asiri, A., Hadi, F., & Jassim, A. (2011). *Technical report of the Gulf Scale for Multiple Mental Abilities (GMMAS)*. Arab Gulf University.

Baker, F. B., & Kim, S. (2017). *The basics of Item Response Theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge University Press. https://doi.org/10.1017/cbo9780511571312

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the *R* environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Deary, I. J. (2020*). Intelligence: A very short introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780198796206.001.0001

Edelen, M. O., & Reeve, B. B. (2007). Applying Item Response Theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(S1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Fu, Q. (2010). *Comparing Accuracy of parameter estimation using IRT Models in the presence of guessing* [Unpublished doctoral dissertation]. University of Illinois at Chicago.

Gao, S. (2011). *The exploration of the relationship between guessing and latent ability in IRT models* [Unpublished doctoral dissertation]. University of Southern Illinois.

Gill, P., Marrin, S., & Phythian, M. (2020). *Developing intelligence theory: New challenges and competing perspectives*. Routledge. https://doi.org/10.4324/9780429028830

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hallowell, D., Okamoto, Y., Romo, L. F., & La Joy, J. R. (2015). First-graders' spatial-mathematical reasoning about plane and solid shapes and their representations. *Zdm – Mathematics Education*, *47*(3), 363–375. https://doi.org/10.1007/s11858-015-0664-9

Hunt, E. (2011). *Human intelligence*. Cambridge University Press.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Jensen, A. R. (1998). *The G factor: The science of mental ability*. Praeger. https://www.amazon.com/Factor-Science-Evolution-Behavior Intelligence/dp/0275961036

Keenan, T. D., & Meenan, C. E. (2014). Using cognitive predictors to identify response to intervention groups. *Journal of Learning Disabilities, 47*(4), 348–360.

Li, Y., Zhang, Y., Dai, D., & Hu, W. (2022). The role of cognitive abilities and self-control in high school students' academic performance: Evidence from Chinese compulsory education. *Journal of Educational Psychology, 114*(7), 1553-1566. https://doi.org/10.1037/edu0000701.

Lohman, D. F. (2005). The role of nonverbal ability Tests in identifying academically gifted Students: An Aptitude perspective. *Gifted Child Quarterly*, *49*(2), 111–138. https://doi.org/10.1177/001698620504900203

Matarneh, A. J., & Oalla, B. M. (2018). Differential item functioning of the test subjects in the English language course administered to students of Mu'tah University. *Journal of Educational and Psychological Sciences, 19*(2), 449-475. https://doi.org/10.12785/jeps/190215

National Council of Teachers of Mathematics. (2023). *Principles and standards for school mathematics.* https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.2307/1164671

Reeve, B. B., Hays, R. D., & Bjørner, J. B. (2007). Psychometric evaluation and calibration of health-related quality of life item banks. *Medical Care, 45*(5), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Salgado, J. F. (2002). The big five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment*, *10*(1 & 2), 117–125. https://doi.org/10.1111/1468-2389.00198

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

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. https://doi.org/10.1037/0022-0663.93.3.604

Smits, N., Cuijpers, P., & Van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188*(1), 147–155. https://doi.org/10.1016/j.psychres.2010.12.001

St Clair-Thompson, H., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Subali, B., Kumaidi, B., & Nonoh, S. A. (2021). The comparison of item test characteristics viewed from classic and modern test theory. *International Journal of Instruction, 14*(1), 647-660. https://doi.org/10.29333/iji.2021.14139a

Suleiman, S. (2010). *The relationship between spatial ability and academic achievement in mathematics among sixth grade students in UNRWA schools* [Unpublished doctoral dissertation]. Islamic University of Gaza.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 197–201. https://doi.org/10.1111/j.2044-8317.1985.tb00834.x

Thurstone theory of intelligence: Exploring multiple factors. (2023). Testbook. https://testbook.com/ias-preparation/thurstone-theory-of-intelligence

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Verdine, B. N. (2011). *Navigation experience in video game environments: Effects on spatial ability and map use skills* [Doctoral dissertation, Vanderbilt University]. ProQuest. https://www.proquest.com/docview/898587381?accountid14624

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

Weckbacher, L. M., & Okamoto, Y. (2014). Mental rotation ability in relation to self-perceptions of high school geometry. *Learning and Individual Differences*, *30*, 58–63. https://doi.org/10.1016/j.lindif.2013.10.007

Wong, E. H., Rosales, K. P., & Looney, L. (2023). Improving cognitive abilities in school-age children via computerized cognitive training: Examining the effect of extended training duration. *Brain Sciences, 13*(12), Article 1618. https://doi.org/10.3390/brainsci13121618

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x