

# Psychometric characteristics of the numerical ability test for Gulf students

Mohammed Al Ajmi<sup>1</sup>, Siti Salina Mustakim<sup>1</sup>, Samsilah Roslan<sup>1</sup>, Rashid Almehrizi<sup>2</sup>

<sup>1</sup>Faculty of Educational Studies, Universiti Putra Malaysia, Serdang, Malaysia

<sup>2</sup>Department of Psychology, College of Education, Sultan Qaboos University, Muscat, Oman

## Article Info

### Article history:

Received Oct 13, 2023

Revised Dec 21, 2023

Accepted Jan 22, 2024

### Keywords:

Item response theory

Marginal reliability

Numerical ability

Psychometric characteristics

Three-parameter logistic model

## ABSTRACT

This study investigates the psychometric properties of the numerical ability test using a three-parameter logistic (3PL) model within the framework of item response theory (IRT). The test comprises 30 dichotomous items and was administered to 2,689 fifth and sixth-grade students in schools across the Arab Gulf countries. The findings indicate a strong alignment of the test items with the three-parameter model, affirming the validity of the IRT approach. The test also meets the criteria for unidimensionality (UD) and local independence, establishing its psychometric soundness. Notably, the numerical ability test excels in discriminating between examinees with varying levels of numerical ability, particularly those with low or average abilities. Moreover, the scale exhibits a high level of reliability, with a marginal reliability coefficient of 0.83. These results suggest the potential for future research aimed at further enhancing the test's precision and effectiveness.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Mohammed Mubarak Al Ajmi

Faculty of Educational Studies, Universiti Putra Malaysia

43400 Serdang, Selangor, Malaysia

Email: mohd7010@gmail.com

## 1. INTRODUCTION

Reevaluating cognitive abilities within the educational system stands as a matter of utmost significance. It not only facilitates the assessment of students' performance but also delves deeper into their grasp of taught concepts and knowledge, thus equipping educators with valuable insights. This empowerment subsequently allows educators to tailor and refine their teaching methods, effectively addressing the unique needs of their students [1]. Moreover, as numerical ability is a skill that extends beyond classroom boundaries, this research also explores its relevance in real-life contexts, career development, and problem-solving abilities in an increasingly quantitative world. By pinpointing areas of weakness through such measurements, educators can readily identify subjects or topics necessitating additional support, ultimately ensuring the delivery of a more enriching and effective education [2]. Additionally, these assessments play a crucial role in appraising the overall quality of the education system. If the results reveal widespread underperformance among students, it can serve as an indicator of systemic issues that demand comprehensive attention to improve the entirety of the education system [3].

Numerical ability is a fundamental component of cognitive development and plays a pivotal role in one's ability to solve mathematical problems, reason logically, and excel in various academic disciplines, such as mathematics, science, and even critical thinking [4]. Understanding the cognitive processes involved in numerical tasks and accurately measuring them is indispensable for educators, policymakers, and researchers alike [5]. It serves as a cornerstone in tailoring educational strategies that cater to individual

student's needs, identifying areas requiring additional support, and fostering a more inclusive and effective learning environment [6]. Moreover, as numerical ability is a skill that extends beyond classroom boundaries, this research also explores its relevance in real-life contexts, career development, and problem-solving abilities in an increasingly quantitative world [7].

In light of the aforementioned importance of measuring cognitive abilities within the educational system, observers of developments in the field of psychological assessment note an increasing focus on the precision and objectivity of the measurement process. This is evidenced by recent changes in the construction and development of assessment tools, leading to a shift in the trajectory of numerous studies [8], [9]. As Subali *et al.* [10] indicates, significant advancements in test construction strategies and item analysis based on classical theory, despite offering solutions to some challenges faced by researchers in test construction and development, have fallen short in addressing other issues.

One such issue is the reliance on an assumption lacking precision, which equates the standard error in measurement for all examinees. Expressing an individual's ability is done through the true score formed by their performance on the entire test, not at the item level. This results in variations in an individual's ability based on changes in the test level. Additionally, test and item characteristics change with changes in individual characteristics [11].

Al-Saikhan and Al-Momani [12] describes this situation as leading to an unstable psychological measurement system, given that item difficulty coefficients fluctuate with changes in the characteristics or abilities of the sample to be tested. Simultaneously, measuring the characteristics or abilities of individuals fluctuates with changes in the difficulty of test items. The stability of scores derived from the test varies with changes in the level and dispersion of the characteristics or abilities of the sample. The scores derived from these tests lack meaning or significance in themselves, as the meaning of the score differs with changes in the difficulty or ease of test items and the narrowness or wideness of the test range.

Therefore, if we intend to compare the performance levels of two individuals or compare an individual's performance in different situations, we must use the same test in either case. Suppose we construct two tests containing the same number of items that measure a specific trait or ability in a group of individuals. In that case, each individual's score will vary with the difficulty of the items in each test. It is said that tests built using classical psychometric methods are limited by the sample of items and the individuals in the sample used for test standardization. Hence, the need for a new theory in measurement has emerged to address theoretical and practical problems that the classical approach has struggled to overcome. This is achieved primarily by liberating an individual's score from being constrained by a specific measurement tool and freeing it from association with the performance of a specific group of individuals [13].

In order to assess the psychometric quality of the numerical ability test and its items, the researcher in the current study turned to the utilization of modern theories in psychological measurement, specifically the latent trait theory (LIT) or what is known as the item response theory (IRT). This choice was motivated by the interest of these theories in establishing a connection between an individual's response to a specific test item and the characteristics inherent in that item [14]. The LIT, with its emphasis on the IRT, provides a contemporary framework for evaluating and understanding the interplay between an individual's responses and the underlying traits being measured. This approach goes beyond traditional psychometric methods, offering a nuanced perspective on the psychometric properties of the numerical ability test and its items.

The IRT is grounded in a set of concepts and principles that fundamentally differ from those upon which its classical predecessor in measurement was based. Bichi and Talib [15] note the existence of two foundational principles that shape the essence of the IRT. The first principle revolves around the ability to predict individuals' performance on a test item through a set of factors referred to as latent abilities or traits. The second assumption encompasses the ability to describe the relationship between individuals' performance on a test item and a set of traits presumed to influence performance on that item. This is achieved through an increasing monotonic function known as the Item Characteristic Function, indicating a higher likelihood for examinees with higher scores to correctly answer the test item compared to their counterparts with lower scores in the trait [16].

This study endeavors to explore the psychometric properties of the numerical ability test through the lens of IRT. Through a meticulous examination of the test's items, their difficulty levels, and patterns in students' responses, our objective is to offer a comprehensive evaluation of the test's measurement properties. Furthermore, this research seeks to contribute significantly to the field of educational measurement and assessment. It is anticipated that the findings of this study will not only enhance our understanding of the numerical ability test's validity and reliability but also shed light on how modern measurement theories, such as IRT, can be effectively employed to optimize the assessment of cognitive abilities in educational contexts. This, in turn, can inform educational practices, improve test construction, and ultimately benefit both students and educators by ensuring more accurate and meaningful assessments of numerical abilities. Therefore, the research attempts to answer the following questions: i) to what extent do the items of the numerical ability

test align with the assumptions of the three-parameter logistic (3PL) model?; ii) what are the parameter values of the items in numerical ability test according to the 3PL model?; and iii) what are the implication?

## 2. METHOD

### 2.1. Participant

In this investigation, a quantitative research methodology is employed, utilizing a descriptive approach to examine the statistical characteristics of the numerical ability assessment within the Gulf Multiple Mental Abilities Scale (GMMAS). The study relies on secondary data derived from the standardization process of the GMMAS, which was conducted by the Arab Office for the Gulf States in 2011. The research sample is composed of students in the fifth and sixth grades, ranging in age from 9 years and 3 months to 12 years and 3 months. Multistage cluster sampling was performed in this study. Samples representing an equal size of the student population in all the Gulf countries were randomly selected within the different groups of students by region, grade, and gender. The overall sample size comprises 2,689 individuals, with 1,273 females and 1,416 males. This sample is deemed suitable for the 3PL model and the 30 items of the numerical ability test [17].

### 2.2. Construction of the item bank for numerical ability

#### 2.2.1. Measure

This study utilizes the numerical ability assessment within the GMMAS [18]. This assessment comprises three distinct tests designed to measure verbal, numerical, and spatial abilities. For the purposes of this research, our focus centers on the numerical ability test, which encompasses a total of 30 multiple-choice items. Numerical ability is gauged through various facets, including counting, addition, subtraction, multiplication, division, numerical relations, numerical reasoning, and arithmetic problem-solving. Each correct response garners a score of 1, while incorrect answers are assigned a score of 0, resulting in a total score range of 0 to 30.

To substantiate the predictive validity of the numerical ability test, correlation coefficients were computed to assess the relationship between numerical ability and academic achievement in mathematics, exclusively within the State of Kuwait. In the fifth grade, the correlation coefficient between numerical ability and mathematics performance yielded a statistically significant value of 0.63 at a significance level of 0.05. Similarly, in the sixth grade, the correlation coefficient between numerical ability and mathematics achievement amounted to 0.38, also deemed statistically significant at a significance level of 0.05. It is noteworthy that these statistically significant correlations were achieved despite the relatively modest sample sizes for each academic level. Additionally, the Raven's progressive matrices test validated the construct of the numerical ability assessment, demonstrating positive and statistically significant correlation coefficients that affirm the test's construct validity [18]. The test-retest reliability coefficient for the numerical ability assessment was established at 0.89. Moreover, internal consistency for numerical ability remained consistently high across all grade levels, with Cronbach's alpha coefficients ranging from 0.82 to 0.87 for Gulf countries [18].

#### 2.2.2. Unidimensionality

The concept of unidimensionality (UD) posits that variations in item responses among individuals primarily result from disparities in a single variable. While it is acknowledged that tests and questionnaires inherently involve multiple variables or traits to elucidate response patterns, it is worth noting that certain variables may not significantly contribute to divergent response patterns within a particular population of respondents [19]. The assumption of UD holds significant importance in the context of IRT, as the precision of item parameter estimations and the validity of test score interpretations are profoundly contingent upon this assumption. When a test is not unidimensional, it can lead to confounding and hinder the interpretability of the results [20].

Primarily, the assessment of the unidimensional model involved the utilization of exploratory and confirmatory factor analyses to validate the unidimensional hypothesis. To establish UD during the exploratory analysis, two fundamental criteria needed to be satisfied. Firstly, the initial factor should explain a minimum of 20% of the variance within the test, according to the Reckase criterion [21]. Secondly, as emphasized by Reeve *et al.* [22], the variance associated with the first factor should be at least four times greater than that of the second factor. Regarding the confirmatory factor analysis (CFA), two key indicators were employed. The root mean square error of approximation (RMSEA) was assessed according to the criteria specified by previous studies [23], [24], where a value of 0.08 or less is indicative of a good fit. Additionally, the Tanaka index, goodness of fit index (GFI) was considered, with a good fit denoted by a value of 0.90, following the criteria established by Tanaka and Huba [25].

### 2.2.3. Local independence

Closely linked to the concept of UD is the notion of local independence. Local independence suggests that, given the latent variable(s), item responses are not interrelated. The prevalent form of local independence, often referred to as strong local independence, posits that the likelihood of observing any specific pair of item responses is the product of the probabilities of observing each individually [26].

The researcher utilized a statistical metric introduced by Yen [27], which entails computing the correlation coefficient between the residuals associated with a pair of items, while accounting for the individual's ability ( $\theta$ ). In order to assess the validity of the local independence assumption for the numerical ability test, the software tool "Local Dependence Indices for Dichotomous Items" (LDID) was employed. It is customary to employ a standardized threshold value of 0.2 for the absolute magnitude of Q3 [28].

### 2.2.4. Item response theory model comparison

Eleje *et al.* [29] succinctly summarized the various IRT models by highlighting that they vary based on the characterization of the relationship between item performance and knowledge, categorized into one-, two-, or 3PL functions. Each of these IRT parameterization models addresses distinct item characteristics, resulting in varying approaches to ability estimation. The 1-parameter (1-PL) IRT model adjusts for item difficulty, the 2-parameter (2-PL) IRT model takes into account both item difficulty and discrimination, while the 3-parameter (3-PL) IRT model considers item guessing, difficulty, and discrimination effects. Additionally, a widely employed one-parameter model, originally developed by Rasch, provides an unbiased, efficient, sufficient, and consistent estimate of separate person and item calibrations, primarily relying on item difficulty [8].

To assess and determine the most appropriate IRT model among various options, four widely accepted model fit indices were employed: the Akaike information criterion (AIC) [30], the Bayesian information criterion (BIC) [31], root mean square standard errors of estimates (RMSE), and the index of the values of the information function (average information). These four indices were used to evaluate the goodness of fit of the statistical models and facilitate the selection of the model that best aligns with the data. Generally, smaller values of the AIC, BIC, and RMSE indicate a superior fit of the IRT model. The model comparison and selection procedures were conducted using the multidimensional item response theory (MIRT) R package [32] and BILOG-MG.

### 2.2.5. Validity and reliability in item response theory

The data collected from school students via GMMAS were subjected to analysis in terms of validity and reliability, with a focus on IRT. In the context of IRT, the scale's validity was assessed by examining the levels of item discrimination and item difficulty [33]. Furthermore, the reliability, as evaluated within the framework of IRT, was assessed using the marginal confidence coefficient [34], [35].

In this study, the R program package MIRT version 1.24 [32] was utilized to calculate item parameters. The software employs the expectation a posteriori (EAP) method, which is grounded in Bayes Estimation principles. By employing the 3PL model, the test items underwent analysis to derive parameter estimates encompassing item difficulty, discrimination, and the pseudo-guessing parameter.

## 3. RESULTS AND DISCUSSION

### 3.1. Psychometric evaluation of the numerical ability item bank

#### 3.1.1. Unidimensionality

To verify the UD assumption of the test, we assessed the adequacy of the sample size using the Kaiser-Mayer-Olkin (KMO) and Bartlett's tests. The results yielded a chi-square value of 10426.066 with a significance level of 0.001 and 435 degrees of freedom, indicating that our sample size is suitable for conducting exploratory factor analysis. We then proceeded with the exploratory factor analysis, focusing on the principal components of the correlation matrix for the 30 items measuring numeric ability within the scale.

The analysis revealed four latent root factors with eigenvalues greater than one, collectively explaining 34.99% of the variance. Notably, the ratio of the eigenvalue of the first factor (5.48) to the eigenvalue of the second factor (1.57) amounted to 3.49, surpassing the threshold of two, which is indicative of UD [21]. Furthermore, the proportion of the first factor's explanatory variance in relation to the total variance stood at 52.17%, comfortably meeting Reckase's criterion of 20% for a unidimensional test. In addition, Cattell's scree plot test performed for the 30-item factor analysis provided further confirmation of the test's UD. The first factor is distinctly isolated from the remaining factors.

We employed the AMOS program to compute the RMSEA and the GFI as additional indicators to assess the data's alignment with the assumption of UD. The study presented the loadings of observed variables with a single latent parameter and the residual error values in the CFA. Our findings indicate that the RMSEA stands at 0.036, which complies with the criterion outlined by Browne and Cudeck [36].

An RMSEA of 0.05 or less suggests a favorable fit. Furthermore, the value of the GFI is 0.95, aligning with the criteria established by Tanaka and Huba [25]. These results contribute to the evidence supporting the UD of the data.

### 3.1.2. Local independence

Within the framework of the 3PL model, an examination of local independence was conducted using the Q3 statistics. Table 1 provides a concise summary of the Q3 values obtained for the test. The findings reveal that the average value of Q3 is 0.038, which is below the critical threshold of 0.2. Additionally, the results show that 100% of the pairs of items in the numerical ability test achieved local independence. This indicates that the items in the numerical ability test have successfully demonstrated local independence.

Table 1. Local independence indicators according to the IRT

Ability	No. of test items	No. of items pairs	Maximum	Minimum	Mean of Q3
Numerical	30	435	0.144	0.0004	0.038

### 3.1.3. Item response theory model comparison

Table 2 presents a compilation of the model fit indices, assisting us in the selection of the most suitable model for the numerical ability test data. The results indicate that the most suitable model for the numerical test data is the 3PL. This model accounts for difficulty, discrimination, and guessing parameters, making it the best-fitting choice. This result aligns better with the test questions in this study, which are multiple choice.

Table 2. The values of the indicators for choosing the appropriate model for the numerical ability test data

S	Indicators	Model		
		1PL	2PL	3PL
1.	-2 log likelihood	97253.2598	96800.9758	96539.1533
	Model differences		452.284*	261.822*
2.	Akaike's information criterion	97280.4	96886.8	96590.5
3.	Bayesian information criterion	97463.2	97240.6	97121.2
4.	Average test information	5.192	5.61	6.275
5.	Root mean square errors	0.4071	0.3955	0.4355

Note: 1PL=one parameter logarithmic model; 2PL=two parameter logarithmic model; 3PL=three parameter logarithmic model; -2LL=-2 log-likelihood

### 3.1.4. Validity and reliability in item response theory

In Table 3, we observe the item difficulty parameters covering a range from -0.695 to 2.039, with item numbers 18 and 10 exhibiting the lowest and highest difficulty values, respectively. The mean of the item difficulty parameter is 0.501, with a standard deviation of 0.599, signifying that the majority of the test items are situated within the realm of moderate difficulty. Figure 1 illustrates the characteristic curves for item 18, which possesses the lowest difficulty value, and item 10, which has the highest difficulty value.

Table 3 provides insights into the item discrimination parameters, which vary between 0.532 for item number 26 and 3.032 for item number 28. The mean of the item discrimination parameter is 1.552, and the standard deviation is 0.586, with item 28 possessing the highest discrimination value. Figure 2 displays the characteristic curves for item 26, characterized by the lowest discrimination value, and item 28, noted for having the highest discrimination value.

In addition, Table 3 illustrates the item pseudo-guessing parameters, with values ranging from 0.000 for item 17 to 0.476 for item 1. The mean of the item guessing parameter is 0.161, and the standard deviation is 0.132, indicating that the use of guesswork when responding to the test items is quite minimal. Figure 3 depicts the characteristic curves for item 17, which exhibits the lowest guessing value, and item 1, which has the highest guessing value.

In the context of the item information function, a higher peak on the curve indicates that the item provides more information and is better for assessing the latent trait or construct being measured. Table 3, it becomes evident that the test items vary in the extent of information they provide, with values ranging from 0.071 for item 26 to 1.719 for item 28. Item 26 contributes the least amount of information, whereas item 28 offers the most substantial information. Figure 4 illustrates the item information function for items 26 and 28, further emphasizing the contrast in information provided by these two items.

Table 3. Item statistics based on IRT

Three-parameter logistic (3PL) model									
Item	a	b	c	IIC	Item	a	b	c	IIC
1	1.244	-0.052	0.476	0.147	16	1.106	-0.025	0.000	0.306
2	1.627	0.7527	0.308	0.365	17	1.262	-0.282	0.000	0.397
3	1.806	0.6814	0.27	0.484	18	1.336	-0.695	0.000	0.444
4	1.476	-0.189	0.337	0.282	19	1.811	0.7628	0.151	0.609
5	2.013	0.459	0.348	0.513	20	2.563	0.7931	0.192	1.115
6	1.711	0.5122	0.324	0.389	21	1.02	1.0894	0.079	0.223
7	1.541	0.1076	0.304	0.329	22	1.052	-0.325	0.000	0.276
8	1.695	-0.124	0.205	0.485	23	1.325	1.2045	0.164	0.32
9	1.811	0.0331	0.278	0.48	24	0.902	0.5725	0.000	0.203
10	2.256	2.0386	0.12	1	25	2.311	0.9561	0.19	0.92
11	1.293	0.1536	0.171	0.3	26	0.532	0.9514	0.000	0.071
12	2.42	0.5098	0.24	0.922	27	0.645	1.2828	0.036	0.097
13	1.848	0.4491	0.147	0.643	28	3.032	1.2898	0.149	1.72
14	1.428	0.3161	0.217	0.335	29	1.752	1.1917	0.137	0.586
15	0.826	0.1574	0.000	0.171	30	0.908	0.4466	0.000	0.206

a: discrimination parameter; b: difficulty parameter; IIC: maximum item information curve

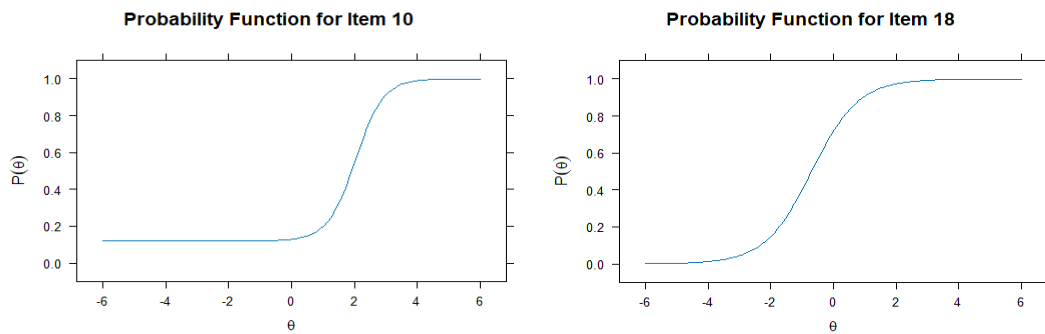


Figure 1. Item characteristic curve of the item (10) and (18)

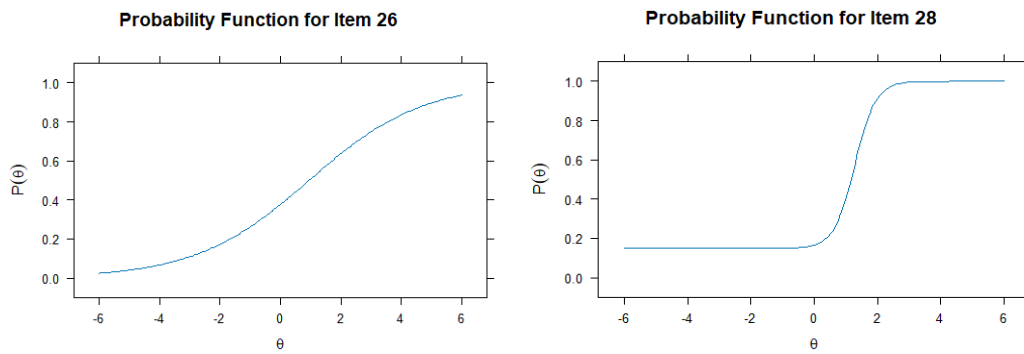


Figure 2. Item characteristic curve of the item (26) and (28)

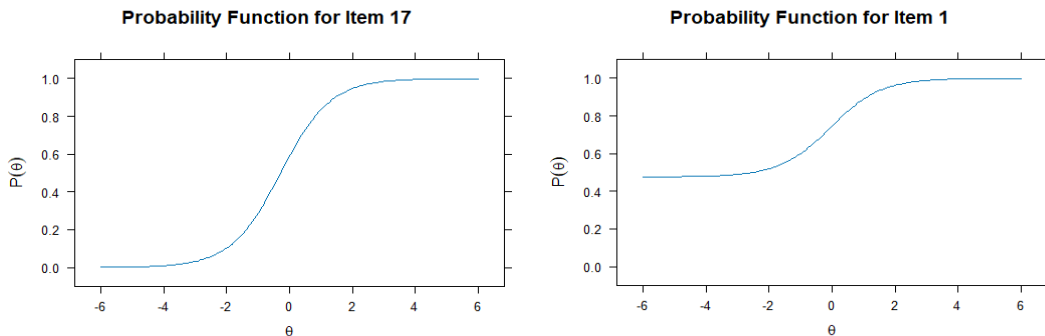


Figure 3. Item characteristic curve of the item (17) and (1)

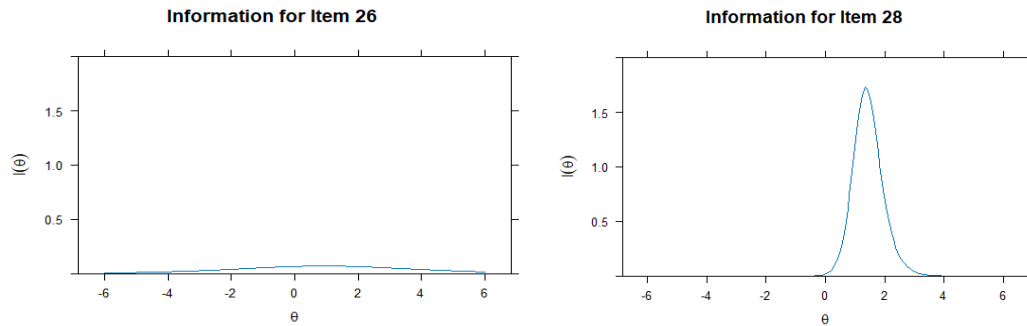


Figure 4. Item information function of the item (26) and (28)

The test information function reflects the overall information provided by the measurement tool. In this case, the test information function suggests that the scale offers the most accurate information about the items falling within the  $-2$  to  $3$  intervals. Specifically, the figure shows that the maximum value of the test information function is  $11.12$ . Also, in Figure 5, the curve's resemblance to a normal distribution indicates its ability to provide information across various levels of the measured trait. This means that the scale offers the most precise information about individuals' satisfaction levels within this specific interval. The marginal reliability coefficient of the numerical ability was calculated to be  $0.83$ . This value is quite close to the reliability values obtained with Cronbach's alpha.

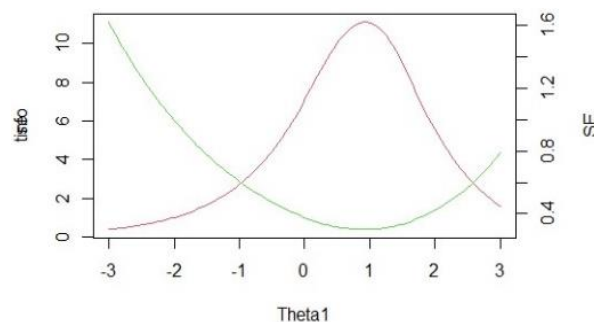


Figure 5. Test information function and standard error of numerical ability

### 3.2. Discussion

This study was conducted to assess the psychometric properties of the numerical ability test. The analysis, employing the 3PL, indicated that all 30 items in the test conform to the model's expectations, validating the assumption of local independence. Therefore, the final analysis was carried out using the complete set of 30 items that make up the scale.

The outcomes of this research highlight the superior performance of the 3PL in comparison to the 1PL and the 2PL when evaluating the numerical ability test. This advantage can be attributed to the multiple-choice format of the test questions, a format widely employed in educational settings [37]. The 3PL model, which considers difficulty, discrimination, and guessing parameters, offers a more accurate estimation of examinees' ability parameters. The inclusion of the guessing parameter in the 3PL model accounts for this behavior, resulting in a better fit for the data. This finding aligns with previous studies [38], [39]. At the same time, it faces the problem of guessing in multiple-choice questions that have been mentioned in much of the literature [40], [41].

The study also provides insights into the results obtained from calibrating items in the numerical ability test. One of the most prominent findings concerns the item difficulty parameter, which displays a wide range of difficulty levels across the test items. However, the mean difficulty parameter, at  $0.50$ , indicates that, on average, the items fall within the realm of moderate difficulty. This discovery carries significant implications for both test design and test-taker performance. It implies that the test effectively challenges students in grades 5 and 6, striking a balance between accessibility and rigor. This outcome aligns with the fundamental principles of measurement theory, emphasizing the importance of including a diversity of item difficulty levels to precisely capture test-takers' abilities.

Regarding the item discrimination parameter, the analysis reveals that the majority of test items exhibit high discrimination, with an average item discrimination parameter of 1.552. High discrimination values indicate the test items' effectiveness in distinguishing between test-takers with varying levels of ability. This feature is highly desirable in an assessment as it enhances the precision and accuracy of ability measurement, ultimately contributing to the test's validity. Furthermore, the low item pseudo-guessing parameter values suggest that test-takers rely minimally on guesswork when responding to the test items. This observation implies that the test items are thoughtfully designed to discourage random guessing. By reducing the impact of guessing, the test can more accurately reflect the true abilities of test-takers, thus enhancing the test's validity and reliability.

An analysis of the item information curves uncovers a diverse range of information provided by the numerical ability test items, with values ranging from 0.071 to 1.719. This variability reflects the items' ability to effectively differentiate among individuals with differing levels of latent traits. Equipped with this understanding, test developers and researchers can identify items that contribute the most valuable information, evaluate the overall measurement quality of the test, and make informed decisions regarding item selection, adaptation, or elimination to improve the assessment's effectiveness and reliability.

The data that illustrates the test information function and the associated standard error for the numerical ability test using the three-parameter model, indicates a significant finding that the largest value of the test information function is 11.1, which is larger than the appropriate value as indicated by Zenisky and Hambleton [42]. This maximum value of the test information function is closely associated with the lowest value of the standard error. It aligns with the idea that higher values of the test information function correspond to lower levels of measurement error, thereby enhancing the test's overall accuracy and reliability. This finding is of significance in the field of psychometrics and educational assessment, as it underscores the test's capability to offer highly reliable and precise measurements of numerical ability, further enhancing its utility in evaluating individuals' skills and capabilities.

#### 4. CONCLUSION

This study delved into the psychometric properties of the numerical ability test and its items using IRT. The findings of the study strongly support the quality of the numerical ability test items. These items are well-crafted, covering a spectrum of medium difficulty levels, exhibiting high discriminatory power, and minimizing the reliance on guesswork by test-takers. The results further validate the scale's reliability and validity, which is attributed to the positive attributes of the difficulty, discrimination, and guessing parameters. Additionally, the item information function and test information function both contribute to the overall strength of the assessment, enhancing its precision. These findings underscore the meticulous construction and precision employed in the test's development, ultimately resulting in a valid and accurate evaluation of test-takers' numerical abilities. Consequently, it is recommended that the positive outcomes from this study be leveraged to transition the numerical ability test from a traditional paper-and-pencil format to a computerized adaptive test. This transition would likely lead to more efficient and precise assessments in the field of numerical abilities.

#### ACKNOWLEDGEMENTS

The authors thank the Arab Bureau of Education for the Gulf States (ABEGS), the Ministry of Education, and Universiti Putra Malaysia (UPM) who provided data and logistical support for this research activity.




#### REFERENCES

- [1] W. J. Popham, *Classroom assessment: what teachers need to know*, 8th ed. Pearson, 2018.
- [2] M. Scriven *et al.*, *Handbook on measurement, assessment, and evaluation in higher education*, 2nd ed. Routledge, 2017.
- [3] L. Guo, J. Huang, and Y. Zhang, "Education development in China: education return, quality, and equity," *Sustainability*, vol. 11, no. 13, p. 3750, Jul. 2019, doi: 10.3390/su11133750.
- [4] S.utama *et al.*, "Collaborative mathematics learning management: critical thinking skills in problem solving," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 11, no. 3, pp. 1015–1027, 2022, doi: 10.11591/ijere.v11i3.22193.
- [5] J. Wai and J. M. Lakin, "Finding the missing einsteins: expanding the breadth of cognitive and noncognitive measures used in academic services," *Contemporary Educational Psychology*, vol. 63, p. 101920, Oct. 2020, doi: 10.1016/j.cedpsych.2020.101920.
- [6] P. Peng and R. A. Kievit, "The development of academic achievement and cognitive abilities: a bidirectional perspective," *Child Development Perspectives*, vol. 14, no. 1, pp. 15–20, Mar. 2020, doi: 10.1111/cdep.12352.
- [7] Z. K. Szabo, P. Körtesi, J. Guncaga, D. Szabo, and R. Neag, "Examples of problem-solving strategies in mathematics education supporting the sustainability of 21st-century skills," *Sustainability*, vol. 12, no. 23, p. 10113, Dec. 2020, doi: 10.3390/su122310113.
- [8] R. Jabrayilov, W. H. M. Emons, and K. Sijtsma, "Comparison of classical test theory and item response theory in individual change assessment," *Applied Psychological Measurement*, vol. 40, no. 8, pp. 559–572, Nov. 2016, doi: 10.1177/0146621616664046.






- [9] A. A. Bichi, R. Embong, R. Talib, S. Salleh, and A. bin Ibrahim, "Comparative analysis of classical test theory and item response theory using chemistry test data," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5C, pp. 1260–1266, May 2019, doi: 10.35940/ijeat.E1179.0585C19.
- [10] B. Subali, Kumaidi, and N. S. Aminah, "The comparison of item test characteristics viewed from classic and modern test theory," *International Journal of Instruction*, vol. 14, no. 1, pp. 647–660, Jan. 2021, doi: 10.29333/iji.2021.14139a.
- [11] A. W. Alili, "Compatibility between the classical and modern theories of matching items achievement test in psychometrics," *Educational Journal*, vol. 3, no. 7, pp. 222–238, 2017.
- [12] R. Al-Saikhan and R. Al-Momani, "Comparison between classical test theory and the three parameters logistic model for item selection of English language achievement test," (in Arabic), *International Journal of Educational Psychological Studies (EPS)*, vol. 10, no. 1, pp. 136–156, Aug. 2021, doi: 10.31559/EPS2021.10.1.8.
- [13] O. A. Awopeju and E. R. I. Afolabi, "Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination," *European Scientific Journal (ESJ)*, vol. 12, no. 28, p. 263–284, Oct. 2016, doi: 10.19044/esj.2016.v12n28p263.
- [14] E. Genge, "Dichotomous IRT models in money-saving skills analysis," *Studia Ekonomiczne*, vol. 304, pp. 84–94, 2016.
- [15] A. A. Bichi and R. Talib, "Item response theory: an introduction to latent trait models to test and item development," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 7, no. 2, pp. 142–151, Jun. 2018, doi: 10.11591/ijere.v7i2.12900.
- [16] D. Ojerinde, "Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results," A guest lecture presented at the *Institute of Education, University of Ibadan*, 2013.
- [17] A. Sahin and D. Anil, "The effects of test length and sample size on item parameters in item response theory," *Educational Sciences: Theory & Practice*, vol. 17, no. 1, pp. 321–335, 2017, doi: 10.12738/estp.2017.1.0270.
- [18] F. Alzayat *et al.*, "Technical Report of the Gulf Scale for Multiple Mental Abilities (GMMAS)," Arab Gulf University, Bahrain, 2011.
- [19] R. R. Meijer and J. N. Tendeiro, "Unidimensional item response theory," in *The Wiley handbook of psychometric testing*, Wiley, 2018, pp. 413–443, doi: 10.1002/9781118489772.ch15.
- [20] D. R. Crişan, J. N. Tendeiro, and R. R. Meijer, "Investigating the practical consequences of model misfit in unidimensional IRT models," *Applied Psychological Measurement*, vol. 41, no. 6, pp. 439–455, Sep. 2017, doi: 10.1177/0146621617695522.
- [21] X. Tian and B. Dai, "Developing a computerized adaptive test to assess stress in Chinese college students," *Frontiers in Psychology*, vol. 11, Feb. 2020, doi: 10.3389/fpsyg.2020.00007.
- [22] B. B. Reeve *et al.*, "Psychometric evaluation and calibration of health-related quality of life item banks," *Medical Care*, vol. 45, no. 5, pp. S22–S31, May 2007, doi: 10.1097/01.mlr.0000250483.85507.04.
- [23] M. O. Edelen and B. B. Reeve, "Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement," *Quality of Life Research*, vol. 16, no. S1, pp. 5–18, Aug. 2007, doi: 10.1007/s11136-007-9198-0.
- [24] N. Smits, P. Cuijpers, and A. van Straten, "Applying computerized adaptive testing to the CES-D scale: a simulation study," *Psychiatry Research*, vol. 188, no. 1, pp. 147–155, Jun. 2011, doi: 10.1016/j.psychres.2010.12.001.
- [25] J. S. Tanaka and G. J. Huba, "A fit index for covariance structure models under arbitrary GLS estimation," *British Journal of Mathematical and Statistical Psychology*, vol. 38, no. 2, pp. 197–201, Nov. 1985, doi: 10.1111/j.2044-8317.1985.tb00834.x.
- [26] M. C. Edwards, C. R. Houts, and L. Cai, "A diagnostic procedure to detect departures from local independence in item response theory models," *Psychological Method*, vol. 23, no. 1, pp. 138–149, Mar. 2018, doi: 10.1037/met0000121.
- [27] W. M. Yen, "Scaling performance assessments: strategies for managing local item dependence," *Journal of Educational Measurement*, vol. 30, no. 3, pp. 187–213, Sep. 1993, doi: 10.1111/j.1745-3984.1993.tb00423.x.
- [28] W.-H. Chen and D. Thissen, "Local dependence indexes for item pairs using item response theory," *Journal of Educational and Behavioral Statistics*, vol. 22, no. 3, pp. 265–289, Sep. 1997, doi: 10.3102/10769986022003265.
- [29] L. I. Eleje, F. E. Onah, and C. C. Abanobi, "Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results," *European Journal of Educational & Social Sciences*, vol. 3, no. 1, pp. 71–89, 2018.
- [30] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974, doi: 10.1109/TAC.1974.1100705.
- [31] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.
- [32] R. P. Chalmers, "MIRT: a multidimensional item response theory package for the R environment," *Journal of Statistical Software*, vol. 48, no. 6, pp. 1–29, 2012, doi: 10.18637/jss.v048.i06.
- [33] Ç. Toraman, E. Karadağ, and M. Polat, "Validity and reliability evidence for the scale of distance education satisfaction of medical students based on item response theory (IRT)," *BMC Medical Education*, vol. 22, no. 1, p. 94, Feb. 2022, doi: 10.1186/s12909-022-03153-9.
- [34] R. K. Hambleton, "The rise and fall of criterion referenced measurement?" *Educational Measurement: Issues and Practice*, vol. 13, no. 4, pp. 21–26, Dec. 1994, doi: 10.1111/j.1745-3992.1994.tb00567.x.
- [35] D. E. Ayala, *The theory and practice of item response theory*, 1st ed. Guilford Publications, 2009.
- [36] M. W. Browne and R. Cudeck, "Alternative ways of assessing model fit," in *Testing structural equation models*, A. Bollen and J. S. Long, Eds., Newbury Park, CA: SAGE, 1993, pp. 136–162.
- [37] P. McKenna, "Multiple choice questions: answering correctly and knowing the answer," *Interactive Technology and Smart Education*, vol. 16, no. 1, pp. 59–73, Mar. 2019, doi: 10.1108/ITSE-09-2018-0071.
- [38] Q. Fu, "Comparing accuracy of parameter estimation using IRT models in the presence of guessing," Ph.D. dissertation, University of Illinois, Chicago, IL, USA, 2010.
- [39] S. Gao, "The exploration of the relationship between guessing and latent ability in IRT models," Ph.D. dissertation, Southern Illinois University, Carbondale, IL, USA, 2011.
- [40] P. Baldwin, "A problem with the bookmark procedure's correction for guessing," *Educational Measurement: Issues and Practice*, vol. 40, no. 2, pp. 7–15, Jun. 2021, doi: 10.1111/emip.12400.
- [41] T. Bond, Z. Yan, and M. Heene, *Applying the Rasch model: fundamental measurement in the human sciences*, 4th ed. Routledge, 2020.
- [42] A. L. Zenisky and R. K. Hambleton, "Effects of selected multi-stage test design alternatives on credentialing examination outcomes," in *the Annual Meeting of the National Council on Measurement in Education*, San Diego, CA, 2004.




**BIOGRAPHIES OF AUTHORS**

**Mohammed Mubarak Alajmi**    is a Ph.D. student at Universiti Putra Malaysia (UPM), he got a master's degree holder in Educational Measurement from Sultan Qaboos University (SQU). He works as Director of the Career Guidance and Counseling Department at the Ministry of Education in the Sultanate of Oman (2019-now). He has contributed to the preparation and publication of many scales such as the professional interest scale and the entrepreneurial traits scale. He has extensive experience in the field of statistics according to the classical theory and the item response theory, as well as he has extensive knowledge in dealing with various statistical programs such as SPSS, BILOG\_MG, MULILOG, M PLUS, AMMOS, WINSTEPS, and ISDL. He can be contacted at email: mohd7010@gmail.com.






**Siti Salina Mustakim**    is a Senior Lecturer at the Faculty of Educational Studies, Universiti Putra Malaysia. As a Doctoral degree holder in Educational Measurement with 18 years of proven experience in managing teaching, research, books and articles publications, and consultation with industries effectively and efficiently, she is known for her attention to detail and precision. The instantaneous progress of her role in the previous organization she was working from 2004-2017 at the Malaysia Ministry of Education, ensures good time management and prioritizing skills, along with the ability to communicate ideas clearly and concisely. She can be contacted at email: mssalina@upm.edu.my.



**Samsilah Roslan**    is a professor in educational psychology at the Faculty of Educational Studies, Universiti Putra Malaysia. She joined the university as a tutor in 1997 and became a full-time lecturer in 2001. Samsilah's intellectual curiosity fuels her research in psychosocial profiling, ecosystem dynamics, special needs education, and innovative teaching methods. For inquiries or collaboration opportunities, she can be contacted at email: samsilah@upm.edu.my.



**Rashid Almehrzi**    currently works at the Department of Psychology, Sultan Qaboos University from August 1994-Present. Rashid does research in Quantitative Psychology and Psychometrics. He has different skills and expertise in reliability, normalization, correlation coefficient, variability, reliability analysis, reliability theory, statistical analysis, linear regression, descriptive statistics, data analysis, variance, applied statistics, multivariate statistics, quantitative modelling, software reliability, statistical modeling, statistical inference, mathematical statistics, maximum likelihood, hypothesis testing, multivariate analysis, R statistical package, factor analysis, correlation analysis, frequency distribution, normal distribution, statistical testing, confidence intervals, computational statistics. He can be contacted at email: mehrzi@squ.edu.om.