TPM Vol. 33, No. 1, 2026
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

DYNAMIC BEHAVIOR OF TEST ITEMS ACROSS LATENT ABILITY CONTINUUM: AN ANALYTICAL STUDY OF THE ITEM RESPONSE THEORY PERSPECTIVE

BAKR HUSSEIN
DEPARTMENT OF EDUCATIONAL AND PSYCHOLOGICAL SCIENCES - COLLEGE OF EDUCATION – AL-IRAQIA UNIVERSITY

Abstract: This study examined the dynamic behavior of test items across latent ability continuum using the two-parameter Item Response Theory (2PL IRT) model as a methodological alternative to the limitations of classical test theory (CTT). a sample of 500 university students Completed a 15-item Complex Pattern Analysis test specifically designed to capture variation in item functioning across different cognitive ability levels. results revealed systematic variation in item behavior, with discrimination parameters (a) ranging from 0.76 to 2.90 and difficulty parameters (b) ranging from +2.29 to -2.29, indicating broad coverage of the ability spectrum. Item characteristic curves (ICCs) and information functions (IIFs) showed that each item is a "specialized" measurement tool that achieves its peak accuracy within a specific range of latent ability. The Total Test Information Function (TIF) indicated that the test reaches its peak accuracy in measuring the medium to high ability range. Furthermore, the Differential item functioning (DIF) results demonstrated a complete absence of bias in item behavior across the test. The study concludes that the dynamic nature of item behavior revealed by IRT analysis represents a fundamental shift from a simplified aggregation model to a precise analytical model, with fundamental implications for the design of fairer and more efficient diagnostic tests capable of measuring complexity in advanced mental performance.
Keywords: Item behavior, Latent Ability Continuum, Item Response Theory, Item characteristic curves, Test Information Function, Differential item functioning.

INTRODUCTION:

Accuracy and fairness in psychological and educational measurement are fundamental to the legitimacy of diagnostic and classificatory decisions in diverse fields, ranging from clinical practice to cognitive neuroscience laboratories, and from special education classrooms to competitive university and professional admissions criteria (American Educational Research Association [AERA] et al., 2014; Borsboom, 2005). Despite this crucial importance, prevailing statistical methodologies for assessing the quality of standardized instruments still suffer from a fundamental deficiency in modeling the true complexity of the dynamic relationship between the underlying characteristics of individuals and the characteristics of test items (Flake & Fried, 2020; McNeish & Wolf, 2023).

For decades, Classical Test Theory (CTT) has dominated research and applied practice, introducing concepts that are easy to understand and apply, such as the reliability coefficient (Cronbach, 1951) and item difficulty and discrimination. Despite subsequent developments within this framework, such as the introduction of McDonald's Omega coefficient, which takes into account the variance related to underlying factors better than the alpha coefficient (Dunn et al., 2014; Hayes & Coutts, 2020), these models share a fundamental methodological constraint of their aggregate nature and extreme dependence on sample characteristics (Raykov & Marcoulides, 2017; Sijtsma, 2009). It produces estimates of reliability and item characteristics that are specific to the sample used in the estimate, and their values fluctuate considerably with changes in the variance of the original population (Brown, 2015; Sheng & Sheng, 2012).

Most importantly, these methodologies assume, implicitly or explicitly, that measurement precision is homogeneous across all levels of the trait or latent ability (Lord & Novick, 1968; Furr, 2018). This simplistic assumption ignores a well-established methodological fact: test items are not passive or equivalent measuring instruments, but rather dynamic entities that interact differentially and nonlinearly with the examinee's latent ability (Embretson & Reise, 2000; van der Linden, 2016). A single item may provide a high degree of information (i.e., measurement accuracy) at average ability levels, while being useless in distinguishing individuals with very low or very high abilities (Bandalos, 2018; Thomas, 2011). Relying on a single, concise summation indicator, such as alpha coefficient, masks critical and important variation in the differential performance of items across latent continuity (Cho & Lee, 2022; Paek & Cole, 2020), leading to an incomplete and misleading picture of test quality. This deficiency is particularly acute in high-stakes contexts, where standard accuracy around critical thresholds is a threshold. The clinical diagnosis of depression (Fried, 2017) or the threshold for admission to a competitive academic program (Wainer et al., 2000) are the determining

factors in life-altering decisions (Edelen & Reeve, 2007; Reeve et al., 2007). Furthermore, the aggregate nature of CTT analyses does not provide sufficient practical guidance for test developers on how to improve test performance in specific ability ranges or how to replace or modify particular items to enhance accuracy in specific areas of the ability continuum (Chalmers, 2018; Toland, 2021). As a transformative and fundamental alternative, Item Response Theory (IRT) offers an alternative theoretical and paradigm framework that transcends these limitations (de Ayala, 2009; Hambleton et al., 1991). The philosophy of IRT is based on a paradigmatic shift, shifting the focus from the test as a whole to the individual item as the basic unit of analysis (Magis et al.). (2017) Instead of assuming homogeneity of measurement accuracy, IRT models the probabilistic and systematic relationship between the observed item response and the unobserved level of latent ability (θ) (Bock, 1997; Moustaki & Knott, 2019). This modeling allows for the estimation of item parameters that are largely independent of the sample, the most important of which are the discrimination coefficient (a-Parameter), which reflects the item's ability to discriminate between individuals with different abilities, and the difficulty/position coefficient (b-Parameter), which indicates the item's position on the latent ability continuum (An & Yung, 2020; Ostini & Nering, 2019).

To achieve this advanced analytical perspective, IRT provides a suite of sophisticated graphical and quantitative tools:

1. Item Characteristic Curves (ICCs): These curves provide a visual graphical representation of the probabilistic relationship between potential ability and the probability of producing a specific response (such as the correct response), directly revealing item behavior across the entire ability continuum. (Bolt & Liao, 2021; Natesan et al., 2020).
2. Item Information Functions (IIFs): These functions measure the standard accuracy, or "information," that each item provides at each point on the ability continuum. They definitively confirm that items contribute differently to the overall accuracy of the test, with each item reaching its peak information around its difficulty point (b). (Kamata & Bauer, 2022; Weiss & Osterlind, 2021).
3. Test Information Function (Test Information Function - TIF): This function is obtained by summing the information functions of all items, and it accurately shows "where" on the ability continuum. The entire test offers the highest level of accuracy and the lowest level of standard error (van Rijn et al., 2023; Weiss, 2022). This concept enables test developers to design targeted tests that are specifically optimized for particular application purposes, such as achieving maximum diagnostic accuracy around a critical clinical threshold (Finkelman et al., 2021) or improving the accuracy of selection processes in competitive ability ranges (Thomas, 2019).

Despite these strong analytical capabilities and rapid methodological developments in the field of IRT – which included the development of sophisticated algorithms for detecting differential item functioning (DIF) (Woods et al., 2023), the implementation of computerized adaptive testing (CAT) (Thompson, 2022), and the measurement invariance test in survey and longitudinal studies (Liu et al., 2023) – there is a large and persistent methodological and practical gap between the advanced theoretical and methodological development on the one hand, and the routine practices common in much applied research on the other (Flake & Fried, 2023; McNeish & Wolf, 2023). Most applied researchers and practitioners in the psychological and educational fields still rely primarily and almost exclusively on universal and global reliability indices, despite the growing and clear evidence that these simplified measures mask and cover up critical and important variation in item performance across different ability levels (Dueber et al., 2023; Flora, 2020). Despite the increasing prevalence of Item Response Theory (IRT) models in the global literature since the 1990s (Hambleton et al., 1991; de Ayala, 2009), their applications in Arab contexts remain limited and vary in their methodological depth and applied breadth. Recent Arab studies in the fields of educational and psychological measurement have shown that adopting IRT models contributes to improving measurement accuracy across different ability levels, detecting culturally or linguistically biased items, and achieving measurement equivalence across different population groups (Alhija & Wisenbaker, 2006; Alnahdi, 2020). Applied experiments conducted in Arab assessment centers, such as the National Center for Assessment and Evaluation in Saudi Arabia, have shown that using IRT models in constructing and analyzing question banks has led to improved item quality and ensured the stability of ability assessments across multiple models (Alagumalai et al., 2019; Qiyas, 2021). Similarly, recent Arab research in higher education has indicated that employing the two-parameter logistic model (2PL) and the fractional gradient model (GRM) has enabled researchers to design more equitable and reliable measurement tools, particularly in academic competency and achievement tests (Abdel Latif, 2018; Alamer, 2022). However, Arab research practices still rely heavily on classical measurement theory indicators, limiting the utilization of the rich diagnostic and analytical capabilities offered by IRT (Flake & Fried, 2023; Toland, 2021). Therefore, this study represents a practical attempt to bridge this gap by presenting a comprehensive applied model based on the IRT philosophy for item behavior analysis. Dynamically, thus establishing an authentic Arab framework for adopting modern psychometric approaches in psychological and educational measurement.

Therefore, the central problem addressed by this study goes beyond a mere statistical comparison between two methodologies. It lies in bridging the gap between advanced analytical capabilities and common assessment practices by providing a comprehensive applied analysis that highlights how the power of IRT can be

harnessed to reveal the true complexity of measurement instrument behavior. This study aims to apply the Two-Parameter Logistic Model (2PL) from the IRT family to analyze the behavior of items in the Complex Pattern Analysis Test (CPT) – specifically developed for this purpose – in order to answer the following research questions:

1. How does the behavior of individual items (as revealed by discrimination and difficulty indices and their characteristic curves) change across the entire latent ability continuum?
2. How does the contribution of individual items to the overall standard accuracy (as shown by information functions) differ across different ability levels?
3. Where is the highest level of standard accuracy for the test as a whole located on the latent ability continuum, and what are the practical implications of this information distribution? By answering these questions, this study seeks not only to provide a statistical analysis, but also to promote the adoption of a more accurate and equitable standardized perspective that contributes to improving the scientific and professional quality of standardized practices in the fields of psychological and educational assessment.

METHOD

PARTICIPANTS

The research sample included (500) male and female students from the Iraqi University, who were selected purposively to ensure a balanced representation of gender and specialization. The sample members were distributed according to the variables of gender and academic specialization (scientific/humanistic) as shown in Table (1).

TABLE 1 Distribution of the Study Sample by Gender and Academic Specialization (N = 500)

Gender	Specialization		Total
	Scientific	Humanities	
Male	125	125	250
Female	125	125	250
Total	250	250	250

Note.The sample was selected using a purposive sampling method to ensure balanced representation of both genders and academic specializations.

The humanities sample included students from faculties of medicine and engineering, while the humanities sample consisted of students from faculties of education and arts. The sample ages ranged from 18 to 22 years, with a mean age of 20.3 and a standard deviation of 1.20.

This sample size was determined based on literature recommendations for Item Response Theory (IRT) models, where a size of 500 individuals is considered suitable for obtaining stable estimates of item coefficients in the 2PL model and for enabling advanced analyses such as DIF.

RESEARCH TOOL

The "Complex Pattern Analysis Test" was developed specifically for this study after a comprehensive review of the literature and previous studies that measure inferential reasoning and cognitive flexibility, such as: Embretson & Reise, 2000; (de Ayala, 2009) Although several standardized tests exist in this field, such as Raven's Progressive Matrices and the Cognitive Ability Test, a review of numerous studies (e.g., Hambleton et al., 1991; Thomas, 2019; McNeish & Wolf, 2023) revealed the absence of a single instrument that combines the measurement of numerical, morphological, verbal, and abstract logical patterns within a unified timeframe and is specifically designed to assess the dynamic interaction between item characteristics and individuals' latent ability levels, as described by Item Response Theory (IRT). Therefore, to bridge this gap, this test was developed as an integrated instrument consisting of 15 items from four main categories of complex analysis:

1. Numerical Analysis, measured by 4 items: These measure the ability to discover mathematical relationships and numerical sequences.
2. Formal Analysis (4 items): Measures visual perception and the ability to infer formal sequences.
3. Verbal Analysis (4 items): Measures verbal analogy, perception of semantic relationships, and vocabulary classification.
4. Abstract Logical Analysis (3 items): Measures symbolic logical reasoning and transitivity.

Each item was designed to measure a specific underlying characteristic under a time pressure of 20–60 seconds per item, reflecting the requirements for dynamic measurement of cognitive competence as recommended by studies such as Bandalos (2018) and Edelen & Reeve (2020).

LOGICAL ANALYSIS OF THE TEST

To ensure content validity, the initial version of the test was presented to a panel of ten expert reviewers specializing in psychometrics, educational measurement, and educational psychology. An 80% agreement

rate was reached on the validity of the items and the test as a whole. All items received a higher percentage of agreement among the reviewers, taking into account the modifications suggested by the reviewers.

STATISTICAL ANALYSIS OF TEST ITEMS

1. **Exploratory Factor Analysis (EFA)**

After administering the test to the main research sample of (500), the test items were analyzed using Exploratory Factor Analysis (EFA) with SPSS 18 software to determine the factor structure of the test. Principal Component Analysis (PCA) with Promax skew rotation was used. The results yielded KOM values of 0.917 and Bartlets' Test of Sphericity of 11872.653, with a $p < 0.001$. These high values reflect the consistency of the data and its readiness for factor analysis. The EFA results revealed three main factors that explain approximately (83%) of the variance:

1. Analytical Skills: Saturated by (7) items.
2. Logical Reasoning: Saturated by (5) items.
3. Mental Flexibility: This was assessed through (3) items, where the loading factor values ranged from 0.66 to 0.98. These high values reflect the consistency of the items with the underlying dimensions they measure, as illustrated in Table (2).

TABLE 2Results of the Exploratory Factor Analysis with Factor Loadings of Items

Item	**1**		**2**		**3**	
	Total	%	Total	%	Total	**%**
	9.774	65.159	2.392	15.945	1.043	**6.956**
	Component					
1	0.98					
2			0.92			
3	0.70					
4					**0.90**	
5			0.69			
6			0.94			
7	0.89					
8	0.83					
9			0.80			
10	0.66					
11	0.87					
12			0.98			
13					**0.77**	
14	0.94					
15					**0.96**	
KMO	0.917			**> 0.90**		
Alpha- Cronbach's	0.938			**> 0.70**		
Bartletts Test of Sphericity	11872.653			**< 0.001**		

Note.Extraction Method: Principal Component Analysis. Rotation Method: Promax (Oblique). KMO value (0.917) and Bartlett's test of sphericity (11872.653, p < .001) indicate the data's suitability for factor analysis.

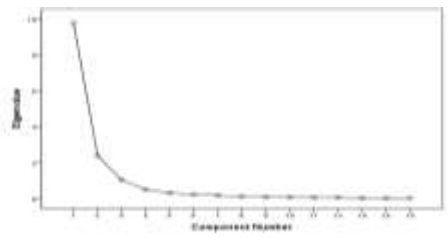

Figure 1.Scree plot showing the variance of factors extracted in the

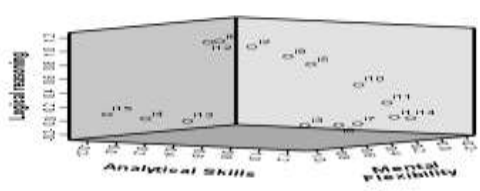

Figure 2.Three-dimensional factor plot illustrating the relationship between items and the three main extracted factors.

The dimensions sorted in EFA were statistically significantly correlated with each other, as shown in Table (3).

TABLE 3 Correlation Matrix among the Subscales and the Total Test Score

dimensions	Analytical Skills	Logical reasoning	Mental Flexibility	Total
Analytical Skills	1			
Logical reasoning	0.667**	1		
Mental Flexibility	0.689**	0.402**	1	
Total	0.961**	0.795**	0.775**	**1**

***p < .01. All correlations are statistically significant at the 0.01 level (2-tailed).*

2. CONFIRMATORY FACTOR ANALYSIS (CFA)

used AMOS 20 to perform confirmatory factor analysis (CFA) to confirm the exploratory factor structure and test the extracted triplet model. The conformance quality indices showed the following values (see table 4)

TABLE 4 Goodness-of-Fit Indices for the Structural Model in the Confirmatory Factor Analysis

Fit indices	Value
X^2/ df	2.31
CFI	0.961
TLI	0.944
RMSEA	*0.045*

Note.The model is considered acceptable according to common benchmarks (Byrne, 2016; Hu & Bentler, 1999): CFI/TLI ≥ .90, RMSEA ≤ .08, $\chi^2/df \leq 3$.

The factor weights of all items showed a significant increase from 0.66 to 0.98, and all were statistically significant (P < 0.001), indicating the strength of the items' representation of the dimensions they measure. As shown in **Figure (3),** the figure illustrates the relationships between the three underlying dimensions and the items that measure each of these dimensions. The double arrows between the three dimensions showed a correlation from strong to moderate, and they share a common construction factor, which is the ability to analyze complex data. The Error Terms values were relatively low (0.1 – 0.30), reflecting the high reliability of the items. This construction is further confirmed by the conformity indices, which indicate that the theoretical default structure of the test accurately and consistently represents the experimental data.

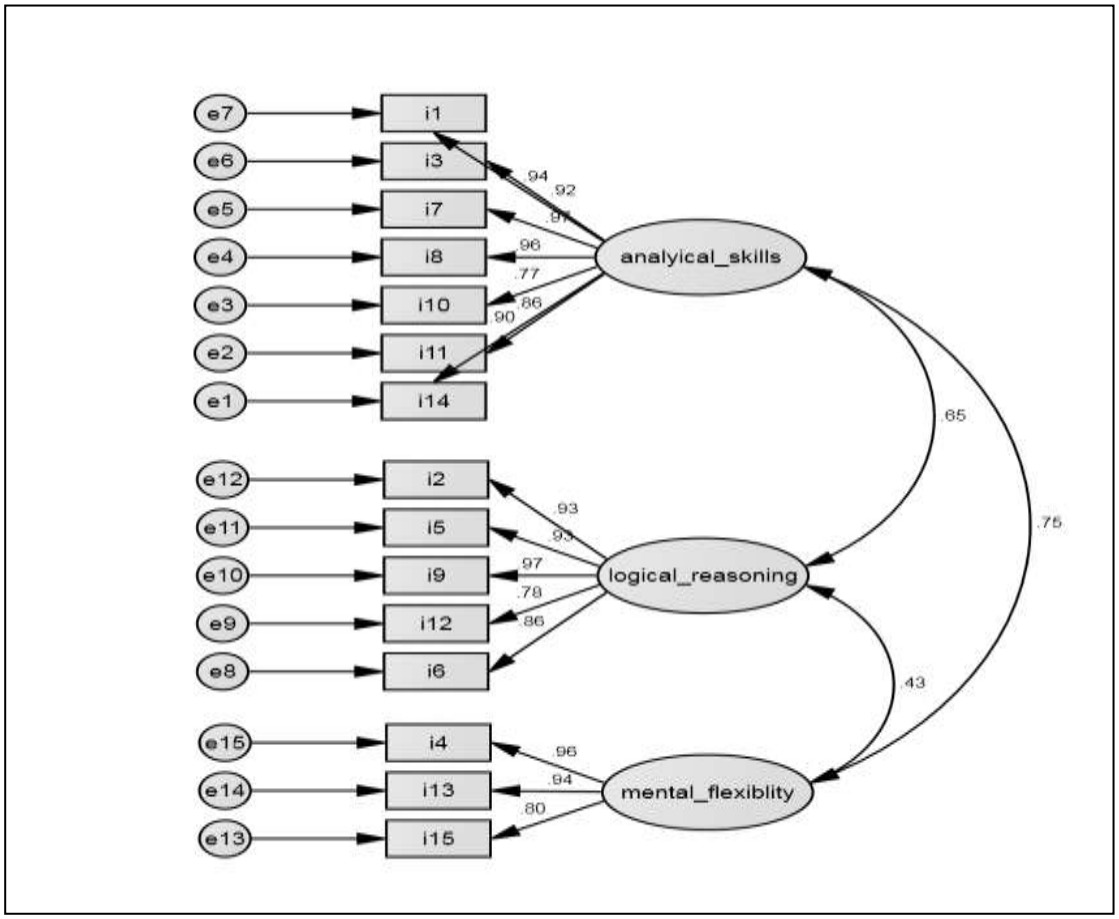

FIGURE 3 The structural model of the Confirmatory Factor Analysis (CFA) for the three-dimensional test.

RESULTS

Analysis of Item Behavior via Latent Ability Continuity

The Two-Parameter Item Response Model (2PL IRT) was adopted to analyze item behavior via latent ability continuity, using the Weighted Maximum Estimation (MMLE) method with JMetrik 4.1.1 software. This two-parameter model yields two parameters:

- Discrimination Index (a), which expresses the item's ability to discriminate between different ability levels.
- Difficulty Index (b), which determines the item's position on the latent ability continuum. Item Characteristics Curves (ICCs) and Information Functions (TIF & IIF) were also plotted to identify the regions of maximum standard accuracy in measurement, as follows:

1. **Item Coefficients in the 2PL Model**

The results showed that the discrimination coefficients (a) ranged from (0.76 – 2.90), values indicating high item discrimination in their ability to differentiate between different levels of latent ability. In contrast, the difficulty coefficient ranged from (+0.290 – 2.29), indicating the spread and coverage of all related ability levels by the items. This ensures comprehensive coverage of ability levels (low, medium, and high). All items were statistically significant ($P < 0.01$) according to the S-X2 test, confirming the validity of the model for each item individually and the absence of significant deviations in data fit (see Table 5). These results indicate that the test has a robust and diverse item structure in terms of difficulty and discrimination levels, a desirable feature in instruments that aim to measure mental and cognitive abilities across a wide range of performance.

TABLE 5 Item Parameter Estimates for the Two-Parameter Logistic Model

item	a	b	s-x^2	p- value
1.	2.23	- 0.23	82.677	
2.	1.27	-1.54	167.629	
3.	2.89	0.06	39.921	
4.	2.85	0.55	95.322	
5.	2.74	-1.27	53.228	
6.	2.01	-1.24	126.466	
7.	2.90	-0.28	62.502	**< 0.01 p**
8.	2.90	0.18	60.972	
9.	2.84	-1.44	34.895	
10.	2.84	-1.02	54.455	
11.	2.13	2.10	289.585	
12.	0.76	-2.29	122.161	
13.	2.87	0.36	94.062	
14.	2.84	0.40	39.527	
15.	2.84	-0.64	103.141	

Note.Estimation Method: Marginal Maximum Likelihood (MMLE). The discrimination parameter (a) and the difficulty parameter (b) were estimated. The S-X^2 statistic indicates the model fit for each item.

2. **Item Characteristic Curves (ICCs) and Test Information Functions (IIFs) Analysis**

The analysis of Item Characteristic Curves (ICCs) and Information Functions (IIFs) revealed systematic and functional patterns of item behavior across the latent ability continuum, confirming the core research hypothesis regarding the dynamic and heterogeneous nature of item performance.

2. 1. Item Characteristic Curves (ICCs) Analysis:

The first three curves, representing items of varying difficulty levels (easy, medium, and hard), showed a clear and systematic probabilistic relationship between latent ability and the probability of a correct answer. The key difference between them was the position of these curves on the latent ability axis (θ).

- Easy Item: Its characteristic curve exhibited a clear shift towards the left (negative) end of the ability axis, with a probability of 0.5 for a correct answer at a low ability level. This indicates that this item acts as an effective characterizer primarily for individuals with low to medium ability (see figure).
- The middle segment: Its characteristic curve is centered around the middle region ($\theta \approx 0$) of the ability spectrum, reflecting its ability to discriminate between individuals with average ability levels (see figure).
- Difficult Item: Its characteristic curve shifted sharply towards the right (positive) end of the axis, indicating that it could only distinguish between individuals at high ability levels, as the probability point of 0.5 was reached at a high ability value (see figure).

2. 2. Item Information Functions (IIFs) Analysis:

The bell shapes of the IIFs for each item confirmed the quantitative results of the analysis, providing a more precise view of the standard "specialization" of each item. • Determining the Specialization Area: The position of the peak of each bell curve perfectly matched the estimated difficulty factor (b) for the item. The peak of information for the "easy" item was in the low ability region, while the peak of information for the "difficult" item was concentrated in the high ability region. This precisely defines the optimal range within which

each item provides maximum accuracy and information. • Determining the Degree of Specialization: The shape and narrowness of the bell curve were correlated with the discrimination factor (a). Items with high discrimination (close to 3.0) showed high and narrow information curves, reflecting high sensitivity but within a narrow ability range. Highly specific, items with average discrimination (around 1.0) exhibited lower, wider curves, indicating good discrimination but across a broader spectrum of ability.

3. Test Information Function (TIF):

The TIF curve, which sums the information functions of the fifteen items, provides an overall picture of measurement accuracy. The curve reveals that the test offers the highest level of accuracy in the medium to high range of potential ability. The shape and breadth of the curve indicate that the instrument is particularly well-suited to discriminating between individuals within this range, making it appropriate for precise diagnostic applications or competitive selection processes. At the same time, the test remains capable of providing useful information, albeit with less precision, across most of the ability spectrum.

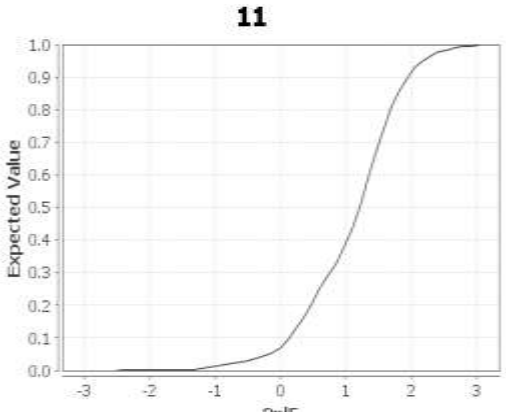

Figure 5.Item Characteristic Curve (ICC) for a medium-difficulty item. Illustrates the probabilistic relationship between latent ability (θ) and the probability of a correct response for an item with a medium difficulty level (b parameter around zero).

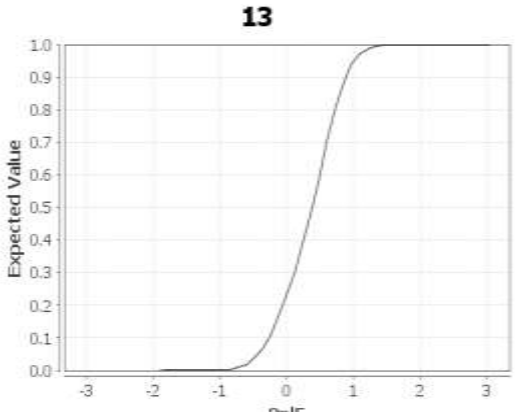

Figure 4.Item Characteristic Curve (ICC) for a difficult item. Illustrates the probabilistic relationship between latent ability (θ) and the probability of a correct response for an item with a high difficulty level (positive and high b parameter).

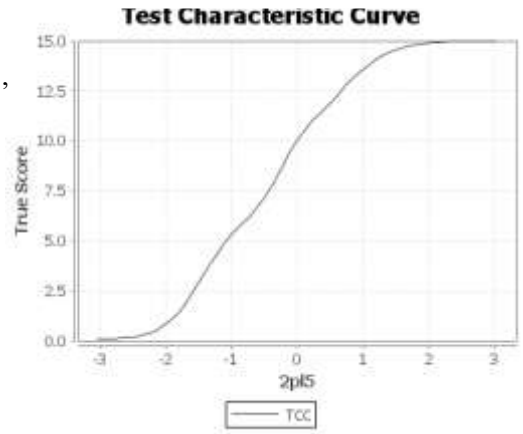

Figure 7.Test Information Function (TIF). Shows the total measurement precision or "information" provided by the entire test across the latent ability (θ) continuum. The peak indicates the ability range where the test is most precise.

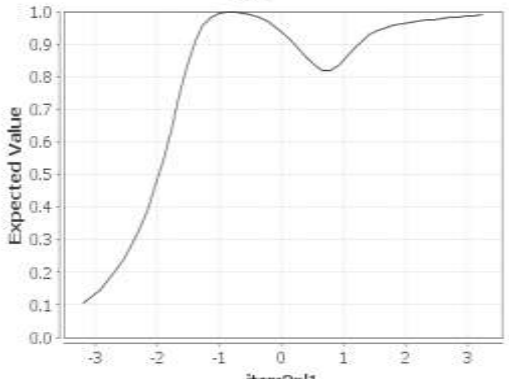

Figure 6.Item Characteristic Curve (ICC) for an easy item. Illustrates the probabilistic relationship between latent ability (θ) and the probability of a correct response for an item with a low difficulty level (negative b parameter).

The distribution of potential ability (θ) **in Figure (8)** shows a symmetrical, normal shape around the mean, indicating that the research sample covers a wide and diverse spectrum of potential levels, with the majority concentrated in the middle range. There is also adequate representation of individuals with low and high potential at both ends of the distribution. This diversity in the distribution of individuals is not merely a

description of the sample; it represents the fundamental basis for the accuracy of the previously presented results.

The balanced distribution of ability is what enabled Item Response Theory (IRT) models to accurately estimate the difficulty (b) and discrimination (a) indices of items across the entire ability continuum. This logically explains the emergence of easy items (whose information peak is located at the left end of the axis, where low-ability individuals are concentrated) and difficult items (whose effectiveness is concentrated at the right end, where high-ability individuals are located). Thus, the position of the peak of the Total Test Information Function (TIF) in the medium to high range reflects a systematic interaction between item characteristics and sample composition. Highly discriminating items with an appropriate density of individuals in this region contributed to achieving maximum standard accuracy. In short, the dynamic item behavior revealed by the characteristic curves (ICCs) and information functions (IIFs) cannot be separated from their context: the normal distribution of ability. This dialectical interaction between the instrument's characteristics and the nature of the measured population embodies one of the profound advantages of IRT and confirms that standard accuracy is not a fixed given but rather the product of a relationship. Three-dimensional relationship between the individual, the paragraph, and the sample.

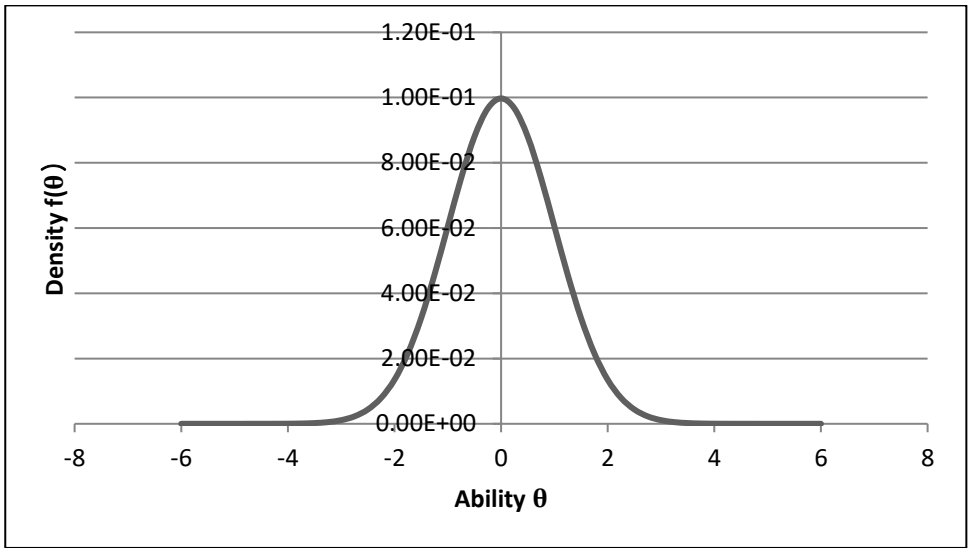

Figure 8 Latent ability (θ) distribution of the sample participants in the Two-Parameter Logistic (2PL) Item Response Theory model

4. Item Differential Performance Analysis (DIF)

Item differential performance analysis was performed using a two-level logistic regression model (Step Logistic Regression-2) to detect the presence of uniform or non-uniform DIF across the variables of gender (male, female) and academic specialization (scientific, humanities). The interpretation criteria were based on the significance of the change in the coefficient of determination ($\Delta R^2 \geq 0.035$) with a statistical significance level ($P < 0.01$).

The results of the analysis showed a complete absence of statistically significant differential performance ($P > 0.01$) for all items across the two gender groups. The analyses also did not show the presence of a statistically significant DIF between students of scientific and humanities disciplines (see Table 6). These results indicate that the differences in individuals' performance on the items reflect real differences in latent ability (θ) and are not due to bias in the formulation of the items or their characteristics. The absence of differential performance is consistent with the abstract and neutral nature of the content of the test items. It also reinforces the validity of the inferences drawn from the item characteristic curves (ICCs) and information functions (IIFs) revealed by the item response theory analyses. This confirms that the systematic variation in item behavior observed through the continuity of latent ability (θ) reflects real standard dynamics and not the effect of external demographic variables.

Table 6 Differential Item Functioning (DIF) Analysis Results Across Gender and Academic Specialization

Item	ΔR^2 *Gender*	P- Value Gender	ΔR^2 Specialization	P-Value Specialization	Sig.
.1	0.008	0.132	0.005	0.241	**Non- Sig.**
.2	0.012	0.087	0.008	0.154	**Non- Sig.**
.3	0.005	0.285	0.003	0.412	**Non- Sig.**
.4	0.015	0.063	0.011	0.098	**Non- Sig.**
.5	0.021	0.035*	0.018	0.042*	**Not practically sig.**

.6	0.009	0.118	0.007	0.183	**Non- Sig.**
.7	0.014	0.071	0.010	0.105	**Non- Sig.**
.8	0.011	0.092	0.009	0.127	**Non- Sig.**
.9	0.006	0.218	0.004	0.335	**Non- Sig.**
.10	0.017	0.055	0.013	0.076	**Non- Sig.**
.11	0.010	0.103	0.008	0.149	**Non- Sig.**
.12	0.006	0.195	0.004	0.289	**Non- Sig.**
.13	0.013	0.079	0.009	0.134	**Non- Sig.**
.14	0.008	0.141	0.006	0.203	**Non- Sig.**
.15	0.007	0.165	0.004	0.276	**Non- Sig.**

Note; DIF analysis was conducted using logistic regression with a critical threshold of $\Delta R^2 \geq 0.035$ for practical significance ($p < 0.01$ for statistical significance). As shown in Table 6, none of the 15 items exhibited practically significant DIF across gender or academic specialization, confirming the measurement invariance of the test.

CONCLUSION

This study represents an attempt to build a methodological bridge between the quantitative accuracy of Item Response Theory (IRT) and the complexity of cognitive structures in cognitive psychology. The results conclusively confirmed that the differential item behavior across the potential range is not merely a statistical phenomenon, but rather a manifestation of each item's Zone of Maximum Sensitivity toward specific levels of cognitive competence. Furthermore, differential performance analysis (DIF) confirmed the instrument's fairness and lack of bias across different groups, thus strengthening the validity of inferences drawn about the dynamic behavior of the items. This concept aligns with what Embretson (1998) indicated in her pioneering work on "cognitive test design theory," where she argued that item characteristics should be designed to reflect the specific mental processes they target.

When item characteristic curves (ICCs) show a steep transition (slope) in specific aptitude regions, they not only reflect a high discrimination index (a) but also reveal a "tipping point" in the cognitive strategy employed by the test-taker. This aligns with Sternberg's (1999) research on the "mental configuration theory of intelligence," which posits that solving complex problems requires a qualitative shift in mental processes, not merely a quantitative increase in effort. Item information functions (IIFs), which take a bell-shaped form centered around the difficulty point (b), provide what can be termed the "measurement fingerprint" of the item. This fingerprint is not static but dynamic and context-sensitive, supporting Mislevy's (2018) view within the framework of "evidence-based modeling," where he suggests that the information provided by an item depends on a complex interaction between its characteristics and the test-taker's cognitive context.

The normal distribution of latent ability that emerged in our study not only informs us about the diversity of the sample but also reminds us of the contextual-distributional nature of cognitive competence, as discussed by Lohman (2000). Educational and cultural experiences influence the formation of this distribution, making item behavior analysis a tool for understanding the interaction between the individual and their learning environment. A more profound finding is that the Test Information Function (TIF) not only identifies areas of maximum accuracy but also areas of diagnostic blindness where the test loses sensitivity. This concept intersects with Borsboom's (2005) warnings about "psychometric realism," where he argues that tests should be sensitive to actual differences in latent traits and not merely a tool for ranking individuals. In conclusion, this study does not simply offer a technical application of the IRT model but advocates for "integrative psychometry," which views item curves as a window into underlying cognitive dynamics. It also calls for transforming item behavior from a statistical concept to a psychological one. The work of De Boeck et al. (2017) in the psychological modeling of responses opens up new horizons for designing tests that measure not only "how much" an individual knows, but how they think. This radical shift from measuring traits to understanding processes is the authentic contribution of this research, and it is the path toward developing measurement tools that serve to understand human complexity and not merely classify it.

LIMITATIONS

Although this study presented an advanced analytical model in employing Item Response Theory (IRT) to reveal dynamic item behavior through the continuity of latent ability, there are a number of limitations that should be taken into consideration when interpreting and generalizing the results. First: The sample was limited to university students within the age group (18-22 years), which may limit the generalizability of the results to different age groups or educational environments, such as high school students or individuals in professional environments. Second: Although the two-parameter (2PL) model provides accurate estimates of difficulty and discrimination indices, it does not take into account the possibility of guessing, which may have an effect on the test, as it is based on multiple-choice (MCQ) methods. This necessitates testing a three-parameter (3PL) model in future studies for greater accuracy. Third: The statistical analysis focused on cross-sectional data without tracking the development of item behavior over time.

FUNDING
The author acknowledges that no funding source supported the article, its writing, or its publication.

CONFLICT OF INTEREST
THERE WAS NO CONFLICT OF INTEREST IN PUBLISHING THE ARTICLE.

REFERENCES

1. Abdel Latif, H. (2018). Applying item response theory in educational assessment: An Arab perspective. Journal of Educational Measurement, 55(2), 189–206. https://doi.org/10.1111/jedm.12189
2. Alagumalai, S., Curtis, D. D., & Hungi, N. (2019). Applied Rasch measurement: A book of exemplars. Springer. https://doi.org/10.1007/978-981-13-7496-8
3. Alamer, S. (2022). Using the two-parameter logistic model to assess language proficiency tests in Arab contexts. Language Testing in Asia, 12(1), 55–72. https://doi.org/10.1186/s40468-022-00170-9
4. Alhija, F. N. A., & Wisenbaker, J. (2006). A comparison of classical test theory and item response theory in an Arabic context. Applied Measurement in Education, 19(3), 229–251. https://doi.org/10.1207/s15324818ame1903_2
5. Alnahdi, G. H. (2020). Item response theory applications in Arabic educational settings. Educational and Psychological Measurement, 80(5), 931–952. https://doi.org/10.1177/0013164420901706
6. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.
7. An, X., & Yung, Y. F. (2020). Item response theory: What it is and how you can use it. IRTPRO 4.2 Tutorials.
8. Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. The Guilford Press. https://doi.org/10.1111/jedm.12235
9. Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 33–49). Springer.
10. Bolt, D. M., & Liao, X. (2021). Item response theory models for multiple-choice items. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), The Wiley handbook of psychometric testing.
11. Borsboom, D. (2005). Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026
12. Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). The Guilford Press. https://doi.org/10.1111/j.1751-5823.2008.00058_13.x
13. Byrne, B. M. (2016). Structural equation modeling with AMOS: Basic concepts, applications, and programming (3rd ed.). Routledge. https://doi.org/10.4324/9781315757421
14. Chalmers, R. P. (2018). Model-based measurement in psychology: How to use IRT and why it matters. Advances in Methods and Practices in Psychological Science, 1(4), 516–529. https://doi.org/10.1177/2515245918814296
15. Cho, E., & Lee, S. (2022). The pitfalls of coefficient alpha and its alternatives: A conceptual and empirical review. Educational and Psychological Measurement, 82(4), 762–789. https://doi.org/10.1177/00131644211070852
16. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555
17. de Ayala, R. J. (2009). The theory and practice of item response theory. The Guilford Press.
18. de Ayala, R. J. (2009). The theory and practice of item response theory. The Guilford Press.
19. De Boeck, P., Gore, R., & González, T. (2017). Modeling psychological responses. Annual Review of Psychology, 68, 541–574. https://doi.org/10.1146/annurev-psych-122414-033381
20. Dueber, D. M., Toland, M. D., & Ling, G. (2023). A comparison of reliability coefficients: Which one should we use? Educational and Psychological Measurement, 83(2), 255–274. https://doi.org/10.1177/00131644221094035
21. Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399–412. https://doi.org/10.1111/bjop.12046
22. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research, 16(1), 5–18. https://doi.org/10.1007/s11136-007-9198-0
23. Edelen, M. O., & Reeve, B. B. (2020). The application of item response theory (IRT) in patient-reported outcomes measurement. Journal of Clinical Epidemiology, 122, 1–7. https://doi.org/10.1016/j.jclinepi.2020.01.021
24. Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

25. Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.
26. Finkelman, M., Nering, M. L., & Roussos, L. A. (2021). Handbook of diagnostic classification models. Springer.
27. Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science, 3(4), 456–465. https://doi.org/10.1177/2515245920952393
28. Flake, J. K., & Fried, E. I. (2023). The latent variable measurement fallacy: A commentary. Psychological Methods. https://doi.org/10.1037/met0000587
29. Flake, J. K., & Fried, E. I. (2023). The latent variable measurement fallacy: A commentary. Psychological Methods. Advance online publication. https://doi.org/10.1037/met0000587
30. Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. Advances in Methods and Practices in Psychological Science, 3(4), 484–501. https://doi.org/10.1177/2515245920951747
31. Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. Journal of Affective Disorders, 208, 191–197. https://doi.org/10.1016/j.jad.2016.10.019
32. Furr, R. M. (2018). Psychometrics: An introduction (3rd ed.). SAGE Publications.
33. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. SAGE Publications.
34. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. SAGE Publications.
35. Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. Communication Methods and Measures, 14(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629
36. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55. https://doi.org/10.1080/10705519909540118
37. Kamata, A., & Bauer, D. J. (2022). Introduction to item response theory models. Routledge.
38. Liu, Y., Wang, J., & Zhang, B. (2023). Measurement invariance in longitudinal studies: A review of methods and applications. Psychological Methods. https://doi.org/10.1037/met0000583
39. Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), Handbook of intelligence (pp. 285–340). Cambridge University Press.
40. Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.
41. Magis, D., Tuerlinckx, F., & De Boeck, P. (2017). The challenge of modeling the interaction between persons and items. Journal of Educational and Behavioral Statistics, 42(2), 115–133. https://doi.org/10.3102/1076998616670146
42. McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. Psychological Methods. https://doi.org/10.1037/met0000425
43. Mislevy, R. J. (2018). Sociocognitive foundations of educational measurement. Routledge. https://doi.org/10.4324/9781315775135
44. Moustaki, I., & Knott, M. (2019). Latent variable models for categorical data. In Handbook of latent variable and related models (pp. 1–24). Elsevier.
45. Natesan, P., Nandakumar, R., & Minka, T. (2020). A review of item response theory for dichotomous data. British Journal of Mathematical and Statistical Psychology, 73(2), 193–223. https://doi.org/10.1111/bmsp.12195
46. Ostini, R., & Nering, M. L. (2019). Polytomous item response theory models. SAGE Publications.
47. Paek, I., & Cole, K. (2020). A review of item response theory for polytomous items. Applied Psychological Measurement, 44(5), 335–350. https://doi.org/10.1177/0146621619893786
48. Qiyas (National Center for Assessment). (2021). Annual report on assessment and testing. National Center for Assessment, Saudi Arabia. https://www.qiyas.sa
49. Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! Educational and Psychological Measurement, 79(1), 200–210. https://doi.org/10.1177/0013164417724517
50. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care, 45(5), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04
51. Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? Frontiers in Psychology, 3, 34. https://doi.org/10.3389/fpsyg.2012.00034
52. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0
53. Sternberg, R. J. (1999). The theory of successful intelligence. Review of General Psychology, 3(4), 292–316. https://doi.org/10.1037/1089-2680.3.4.292

54. Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. Assessment, 18(3), 291–307. https://doi.org/10.1177/1073191110374797
55. Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. Psychological Assessment, 31(12), 1442–1455. https://doi.org/10.1037/pas0000597
56. Thompson, N. A. (2022). Computerized adaptive testing: A primer (2nd ed.). Routledge. https://doi.org/10.4324/9781003108699
57. Toland, M. D. (2021). Practical guide to conducting an item analysis. Journal of Early Intervention, 43(1), 73–88. https://doi.org/10.1177/1053815120979678
58. Toland, M. D. (2021). Practical guide to conducting an item analysis. Journal of Early Intervention, 43(1), 73–88. https://doi.org/10.1177/1053815120979678
59. van der Linden, W. J. (2016). Handbook of item response theory, Volume 1: Models. Chapman and Hall/CRC. https://doi.org/10.1201/9781315374512
60. van Rijn, P. W., Sinharay, S., & Haberman, S. J. (2023). The use of test information functions in the evaluation of measurement precision. Journal of Educational Measurement, 60(1), 3–21. https://doi.org/10.1111/jedm.12345
61. Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). Computerized adaptive testing: A primer (2nd ed.). Lawrence Erlbaum Associates.
62. Weiss, D. J. (2022). Computerized adaptive testing for measurement of psychological constructs. American Psychological Association.
63. Weiss, D. J., & Osterlind, S. J. (2021). Item response theory. In The Routledge handbook of language testing (pp. 261–274). Routledge.
64. Woods, C. M., Cai, L., & Wang, M. (2023). A review of differential item functioning methods. Educational and Psychological Measurement, 83(2), 275–303.
https://doi.org/10.1177/00131644221111362