Ed Process Int J

www.edupij.com

ISSN 2147-0901 (Print), 2564-8020 (Online)

EDUCATIONAL PROCESS
INTERNATIONAL JOURNAL

Research Article

Cite this article: Freihat, A. M., & Yassin, O. S. B. (2025). The Accuracy of Estimating Parameters of Multiple-Choice Test Items, Following Item-Response Theory: A Simulation Study. *Educational Process: International Journal, 14*, e2025054. https://doi.org/10.22521/edupij.2025.14.54

Received December 11, 2024
Accepted February 10, 2025
Published Online February 14, 2025

Keywords:
Estimation, item parameter, item-response theory, models, multiple-choice test.

Authors for correspondence:
Aiman Mohammad Freihat
aiman.freihat@bau.edu.jo
Ajloun University College, Al-Balqa Applied University, Jordan
Omar Saleh Bani Yassin
omarsa@bau.edu.jo
Irbid University College, Al-Balqa Applied University

The Accuracy of Estimating Parameters of Multiple-Choice Test Items, Following Item-Response Theory: A Simulation Study

Aiman Mohammad Freihat, Omar Saleh Bani Yassin

Abstract

Background/purpose. This study aimed to reveal the accuracy of estimation of multiple-choice test items parameters following the models of the item-response theory in measurement.

Materials/methods. The researchers depended on the measurement accuracy indicators, which express the absolute difference between the estimated and actual values of the parameters of the items. The researchers depended on the square root of the error's mean squares and their relative efficiency (RE). (1500) responses were generated under the assumption of a normal distribution, following the ability parameter. Several tests comprising (50) items each were generated under the assumption of distributions (normal for difficulty, regular for discrimination, regular for guessing), assuming that the tests are multiple-choice, using the Wingen V data generation V.3 program. The BILOG-MG software was used to estimate the item's parameters using the marginal maximum likelihood method. Then, the estimated parameters were compared to the actual parameters using two indicators (absolute difference, the square root of the squares mean of the error, and the relative efficiency index of the variances of the estimated parameters).

Results. The study results showed that the three-parameter model was more accurate in estimating the difficulty parameter, followed by the single-parameter model and then the two-parameter model.

Conclusion. The results showed that the three-parameter model was more accurate than the two-parameter model. Also, the results showed the guessing parameter is only related to the three-parameter model. The estimated guessing parameter was more accurate in the five-alternative tests, followed by the three-alternative tests and then the four-alternative tests.

1. Introduction

Evaluation is considered an essential element in the educational process because it plays an active role in its development and progress. Its purpose is to determine the extent of progress in achieving the educational goals. Teachers continuously observe students' behavior and collect information using various instruments and methods. Accurate tools are necessary to measure this development and progress. Based on the results, appropriate decisions are made regarding student advancement to higher levels, classification, and other educational actions.

In the educational process, testing is one of the most important evaluation instruments used in schools. Additionally, various psychological scales, as well as tests of readiness, intelligence, and abilities, rely primarily on this tool. Therefore, the measurement instrument must be accurate and possess multiple characteristics and specifications. One of the most important specifications is objectivity, which is achieved by eliminating examiner bias (Odeh, 2014).

The primary purpose of a test is to provide scores that reflect an examinee's skill level or the degree to which they possess the trait under study in accordance with test measurement principles (Hambleton & Swaminathan, 1991). Multiple-choice tests are among the most widely used formats for achievement testing due to their efficiency in reliably and validly measuring academic performance (Aiken, 1982; Gay, 1980; Frisbie & Sweeney, 1987).

A multiple-choice test consists of two main parts. The first part (the stem) is a question or an incomplete sentence that presents a problem requiring a response. The second part includes a set of possible answers, with one being the correct answer and the rest serving as distractors. These distractors are designed to appear plausible to individuals who do not know the correct answer. The number of answer choices typically ranges from two to five (Nitko, 2001). Although multiple-choice test items have many advantages, they face criticisms related to examinee response behavior and item design. Notably, the examinee's selection of the correct answer can be influenced by the placement of the correct alternative among the options and by the content of the test items (Blunch, 1984).

The optimal number of answer choices in test items has been a focus of psychometric research, as researchers aim to understand its impact on item and test characteristics. Brewer and Haladyna's findings (as cited in Crehan, 1993) indicated that items with three alternatives were more difficult than those with four alternatives. However, the number of alternatives did not significantly affect the discrimination power of the items.

As a result, a new approach to measurement emerged based on mathematical models and probability theory established by Lord. Lord formulated the foundations and assumptions of what became known as "Item Response Theory" (IRT). In 1960, the Danish mathematician George Rasch introduced the first models of this theory, focusing on a single parameter: item difficulty. Rasch named this model the "Rasch Model." This development brought psychometric measurement closer to the objectivity of physical measurement, characterized by the principle that the measurement results should not be affected by the instrument or the individuals using it as long as it is appropriate for measuring the intended phenomenon (Lord, 1980).

The second model, known as the Lord Model, allows test items to vary in both difficulty and discrimination, which is often observed in test construction. Finding a set of items that equally distinguish between different levels of a trait or ability measured by a test is challenging—an assumption upon which Rasch's model was based (Allam, 2000).

The third model, known as the Birnbaum Model, introduced a third parameter: the guessing parameter. This parameter accounts for the probability that a test-taker with very low ability might

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

answer some items correctly by guessing. The significance of this parameter becomes evident when evaluating data from multiple-choice test items (Allam, 2005).

1.1. Study Questions

This study aimed to examine the impact of the number of answer choices on the accuracy of estimating test item parameters using item response theory models (one-parameter, two-parameter, and three-parameter). Specifically, the study sought to answer the following research questions:

1. Are there any significant differences (α=0.05\alpha = 0.05) in estimating the difficulty parameter due to the study variables (the model and the number of alternatives)?

2. Are there any significant differences (α=0.05\alpha = 0.05) in the estimation of the discrimination parameter due to the study variables (the model and the number of alternatives)?

3. Are there any significant differences (α=0.05\alpha = 0.05) in the estimation of the guessing parameter due to the study variable (the number of alternatives)?

1.2. Study Significance

Given the widespread use of multiple-choice tests in assessing academic achievement, some psychometricians advocate increasing the number of answer choices, while others argue for reducing them. Increasing the number of alternatives may limit the ability to construct strong distractors, thereby affecting item difficulty. Therefore, the significance of this study lies in examining the impact of the number of alternatives on the estimation of item parameters using item response theory models. The findings aim to improve test quality by optimizing item characteristics based on these models.

1.3. Definition of Terms

- Item parameters: The parameters of difficulty, discrimination, and guessing as derived from item response theory models.

- Number of alternatives: The number of answer choices in a multiple-choice item, which in this study are three, four, and five.

- Difficulty parameter: A point on the ability continuum corresponding to the probability (1+ci)/2(1 + c_i) / 2 of answering item ii correctly, where cic_i represents the guessing parameter.

- Discrimination parameter: The slope of the item characteristic curve at the point where the ability level equals the item difficulty.

- Guessing parameter: The probability that examinees with low ability will answer the item correctly by guessing.

- Item response theory models: Statistical models (one-parameter, two-parameter, and three-parameter) used in data analysis to estimate item parameters through mathematical functions and specialized computer programs.

- Estimation: The process of determining the accuracy of parameter estimation, characterized by a high probability that the estimated value is close to the true value. This is assessed using an unbiased estimator (absolute difference between the estimated and actual values) and its variance, with accuracy measured by the root mean square error (RMSE).

1.4. Study Limitations

- This study was conducted using a simulation method with generated data.

- The study was limited to comparing items with three, four, and five answer choices.

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

- The study focused on item response theory models, specifically the one-parameter, two-parameter, and three-parameter models.

2. Literature Review

Many psychometricians and educational scientists have shown interest in multiple-choice tests. Both Arab and international studies have examined the impact of the number of alternatives in multiple-choice tests on test characteristics. These studies have yielded differing opinions regarding the optimal number of answer choices for multiple-choice items. Additionally, numerous studies have investigated methods for aligning test items with item response theory (IRT) models. Below are some key studies:

Rickase and Mark (1978) conducted a study comparing the estimation of ability and item parameters using the Rasch model and the three-parameter model through simulation. The results indicated that the three-parameter model provided a better fit for the test data than the Rasch model. Additionally, the ability parameter estimated by the Rasch model was lower than that estimated by the three-parameter model. A notable finding was the high correlation between ability estimates derived from both models across most datasets. The study concluded that the Rasch model is preferable for small samples, while the three-parameter model requires larger samples.

Macdonald and Paunonen (2002) generated data for 100 tests using simulation, varying the number of items (20, 40, and 60). They employed the Monte Carlo method and applied the tests to two randomly generated samples of 1,000 individuals each. To answer the study questions, item difficulty, and discrimination indices were calculated, along with individuals' ability scores using both classical test theory and IRT (Rasch and Birnbaum models). Statistical analyses were performed using SPSS and the IRT-based PARASCALE software. The correlation coefficients between ability scores from classical test theory and the Rasch model were found to be at least 0.97. This high correlation suggested that decisions regarding examinee ability levels remained consistent regardless of the theoretical framework used. The study also found high reliability in IRT-based estimations, particularly in probabilistic sampling. While both theories provided high estimates for item difficulty and ability, discrimination estimates were more accurate under IRT.

Al-Sharifin and Taamneh (2009) examined the impact of the number of alternatives in multiple-choice tests on the estimation of individuals' abilities and item difficulty parameters. The results indicated no statistically significant differences in the mean standard error of item difficulty based on the number of alternatives. However, significant differences were observed in the standard errors of ability parameter estimates, favoring three alternative tests in terms of estimation accuracy.

Fu (2010) explored the estimation of ability and difficulty parameters using five different IRT models, varying the guessing parameter, sample size, and test length. The study generated 50 datasets under different test conditions. The results revealed differences in parameter estimations depending on the guessing value, sample size, and test length. Additionally, the accuracy of ability and item parameter estimation depended on the specific evaluation criteria used within each IRT model.

Yaman (2011) investigated the optimal number of alternatives in multiple-choice tests by comparing their psychometric properties. The findings indicated that tests with three and five alternatives had higher reliability than those with four alternatives. However, no significant differences were observed among the three test formats in terms of item difficulty and discrimination. The study recommended using three alternative items due to their ease of development and analysis.

Al-Rabba'i (2012) studied the impact of the number of alternatives and the positioning of strong distractors on item and test characteristics based on IRT. The study developed a 54-item multiple-

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

choice achievement test for tenth-grade students, administered to a sample of 2,123 students from Ramtha and First Irbid Directorates. The results showed no significant differences in the estimation of difficulty and guessing parameters due to the number of alternatives, the position of the strong alternative, or their interaction. However, a significant difference was found in the discrimination parameter based on the number of alternatives. No significant differences were observed in discrimination estimates due to the positioning of the strong alternative or its interaction with the number of alternatives.

3. Methods

3.1. Data Generation

This study employed a simulation method to generate the necessary data. The WinGen V3 program was used to simulate data under varying levels of the two study variables. The goal was to compare the mean absolute differences between estimated and actual item parameters (a, b, c) and evaluate the efficiency of parameter estimation across different conditions using the marginal maximum likelihood (MML) method in Bilog-MG v3. The actual ability distribution was generated based on a normal distribution with a mean of 0 and a standard deviation of 1, as shown in Table 1.

Table 1. Means and standard deviations of the actual ability parameter following the two study variables

Alternatives Number	Statistical	Models		
		One	Binary	Triple
3	Mean	0.12	000	0.10
	Standard deviation	1.32	1.04	1.04
4	Mean	0.25	0.41	0.04
	Standard deviation	0.96	1.00	1.12
5	Mean	0.00	-0.12	0.02
	Standard deviation	1.12	0.65	1.42

The actual difficulty parameter was also generated based on the normal distribution with a mean of (0) and a standard deviation of (1).

Table 2. Means and standard deviations of the actual item difficulty parameter following the two study variables

Alternatives Number	Statistical	Models		
		One	Binary	Triple
3	Mean	0.06	-0.066	0.325
	Standard deviation	0.82	1.041	0.785
4	Mean	0.324	0.041	0.754
	Standard deviation	1.011	1.022	1.062
5	Mean	0.231	-0.011	0.024
	Standard deviation	1.021	1.036	1.035

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

In addition, the actual discrimination parameter was generated based on the regular distribution in table (3).

Table 3. Means and standard deviations of the values of the discrimination parameter of the actual item following the two variables of the study

Alternatives Number	Statistical	Models	
		Binary	Triple
3	Mean	0.654	0.574
	Standard deviation	0.321	0.421
	Minimum Value	0.324	0.421
	Maximum Value	1.321	1.012
4	Mean	0.741	0.652
	Standard deviation	0.248	0.213
	Minimum Value	0.414	0.413
	Maximum Value	1.254	1.145
5	Mean	0.754	0.654
	Standard deviation	0.111	0.321
	Minimum Value	0.212	0.302
	Maximum Value	1.199	1.193

Finally, the actual guessing parameter was generated based on a normal distribution of values ranging from 0.32 to 0.34 when the number of alternatives was three, from 0.24 to 0.26 when the number of alternatives was four, and from 0.19 to 0.21 when the number of alternatives was five.

Table 4. Means and standard deviations of the values of the actual item-guess parameter following the levels of the study variable

Alternatives Number	Mean	Standard Deviation	Minimum Value	Maximum Value
3	0.214	0.1003	0.321	0.321
4	0.321	0.1004	0.012	0.012
5	0.122	0.2100	0.012	0.124

3.2. Detection of one-dimensionality using exploratory factor analysis:

One-dimensionality was verified using exploratory factor analysis to process the generated data related to the responses of 1,010 individuals in each condition, based on the study variables for 50 items. Table 5 presents the results of the exploratory factor analysis of the generated data.

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 5. Results of the factorial analysis of the data generated following the (model, alternatives number) study variables

Model	Alternatives Number	Component	Latent Root	Ratio of Explained Variance	Ratio of Explained Accumulated Variance	First Component / Second Component	(First Component-Second Component) / (Second Component-Third Component)
One Parameter	3	1	29.762	59.523	59.523	21.37	334.08
		2	1.393	2.786	62.309		
		3	1.308	2.616	64.925		
		4	1.265				
	4	1	29.902	59.805	59.805	22.46	594.52
		2	1.332	2.663	62.468		
		3	1.284	2.567	65.035		
		4	1.253				
	5	1	26.555	53.109	53.109	19.45	322.17
		2	1.365	2.730	55.840		
		3	1.287	2.574	58.414		
		4	1.250				
Binary Parameter	3	1	24.987	49.973	49.973	17.94	277.85
		2	1.393	2.786	52.759		
		3	1.308	2.616	55.375		
		4	1.265				
	4	1	28.044	56.089	56.089	21.06	555.86
		2	1.332	2.663	58.752		
		3	1.284	2.567	61.319		
		4	1.253				
	5	1	24.589	49.178	49.178	18.01	297.03
		2	1.365	2.730	51.909		
		3	1.287	2.574	54.483		
		4	1.250				
Triple Parameter	3	1	25.970	51.940	51.940	18.64	289.43
		2	1.393	2.786	54.726		
		3	1.308	2.616	57.342		
		4	1.265				
	4	1	29.379	58.758	58.758	21.52	358.30
		2	1.365	2.730	61.488		
		3	1.214	2.574	64.321		
		4	1.324				
	5	1	26.112	52.282	52.282	19.63	421.26
		2	1.452	2.634	62.946		
		3	1.284	2.425	57.513		
		4	1.321				

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

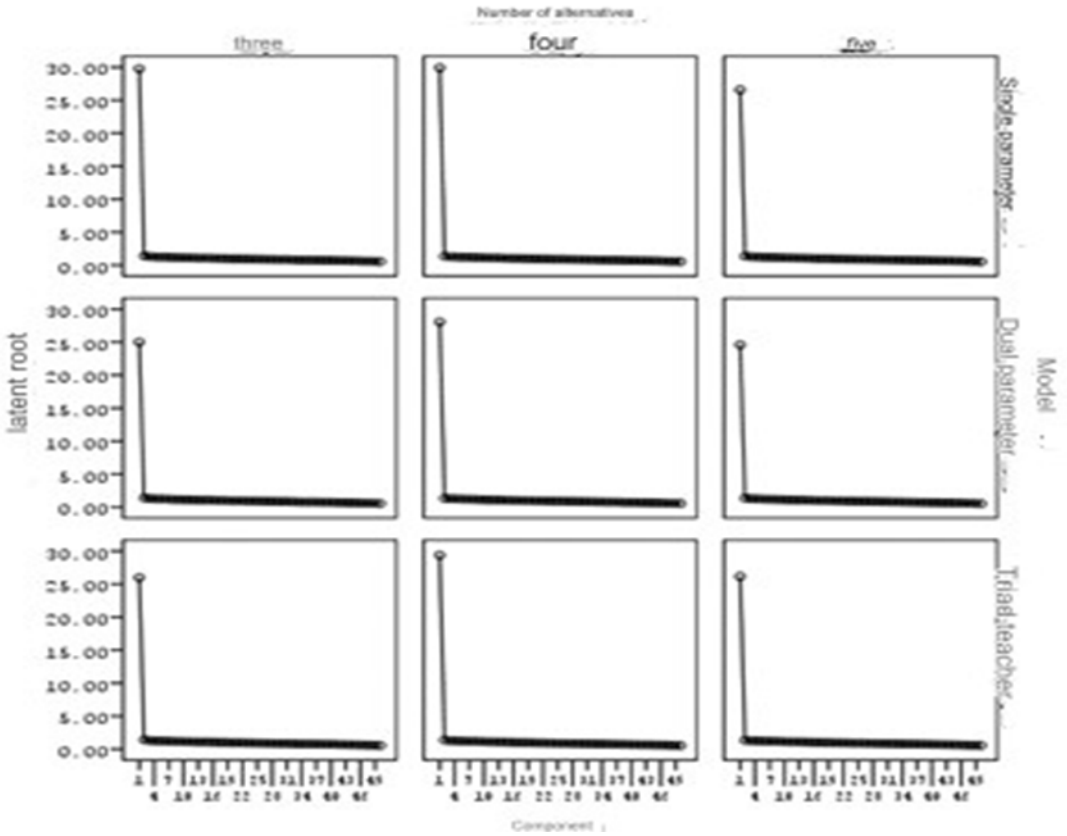

Figure 1. A Scree Plot graph that reveals the one-dimensionality of the data generated following the study variables

3.3. Detection of One-Dimensionality Using the NOHARM Program

The assumption of one-dimensionality was verified using the NOHARM program to process the generated response data for a test comprising 50 items based on the two study variables. One-dimensionality was assessed using two simultaneous indicators: the TANAKA index and the RMSR index. Table 6 presents the values for each indicator according to the two study variables.

Table 6. One-dimensional (TANAKA) and (RMSR) values synchronized together following the study variables

Model	Statistical	Alternatives Number		
		3	4	5
One	RMSR Indicator	0.12000	0.12002	0.102000
	Tanaka Indicator for Good Conformity	0.72145	0.76666	0.765421
Binary	RMSR Indicator	0.100325	0.006271	0.100023
	Tanaka Indicator for Good Conformity	0.75422	0.966244	0.76000
Triple	RMSR Indicator	0.21003	0.004378	0.100321
	Tanaka Indicator for Good Conformity	0.75600	0.963172	0.810200

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 6 shows that all values of the TANAKA index exceed 0.75, coinciding with a decrease in all RMSR index values. This result supports the assumption of one-dimensionality, as assessed by the NOHARM program. To satisfy this assumption, the TANAKA index must exceed 0.75 while the RMSR index approaches zero without exceeding its critical value in any of the study conditions (Miller, 1991).

3.4. Detection of Positional Independence

To verify the assumption of local independence, the LDID (Local Dependence Indices for Dichotomous Items) program was applied to the generated data based on the study variables. This program identified the number of residual associations between test item pairs that maintained local independence. Items that did not meet the Q3 index threshold for local independence were converted to ZQ3 G-index values corresponding to each case.

Table 7. Repetitions and percentages of the zq3 positional Independence Index following the study variables

Model	Independence Positional Case	Alternatives Number					
		3		4		5	
		frequency	percentage	frequency	percentage	frequency	percentage
One	Dependent	321	.220	265	21.6	262	21.4
	Independent	754	680	960	78.4	963	78.6
	Total	1225	100.0	1225	100.0	1225	100.0
Binary	Dependent	187	15.3	187	15.3	182	14.9
	Independent	1038	84.7	1038	84.7	1043	85.1
	Total	1225	100.0	1225	100.0	1225	100.0
Triple	Dependent	204	16.7	188	15.3	191	15.6
	Independent	1021	83.3	1037	84.7	1034	84.4
	Total	1225	100.0	1225	100.0	1225	100.0

Table 7 shows that all repetitions and percentages did not fall below 76.7%, indicating that the assumption of local independence was met. This assumption states that "there is no local dependence at the significance level ($\alpha = 0.05$)" in all study conditions based on the study variables (Jasper, 2010).

4. Results and Discussion

The first research question states: "Are there any significant differences ($\alpha = 0.05$) in the estimation of the item difficulty parameter due to the two study variables (model and number of alternatives)?"

To answer this, the mean and standard deviation of the estimated item difficulty parameter were calculated based on the study variables (model and number of alternatives). Table 8 presents the results of the analysis.

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 8. Means and standard deviations of the difficulty parameter of the estimated item following the study variables

Alternatives Number	Statistical	Models		
		One	Binary	Triple
3	Mean	0.012	-1.214	0.423
	Standard deviation	0.31	0.752	0.742
4	Mean	0.100	-0.762	0.521
	Standard deviation	0.43	1.420	1.022
5	Mean	0.012	-0.820	0.321
	Standard deviation	0.52	1.102	0.820

Table 8 indicates that the mean of the estimated item difficulty parameter varies between the two study variables. Based on the results in Table 8, the means and standard deviations of the absolute difference between the estimated and actual difficulty parameters were calculated according to the study variables.

Table 9. Means and standard deviations of the absolute difference between the estimated and actual difficulty parameter following the two study variables

Alternatives Number	Statistical	Models		
		One	Binary	Triple
3	Mean	0.379	1.306	0.354
	Standard deviation	0.28	0.49	0.25
4	Mean	0.437	0.856	0.354
	Standard deviation	0.39	0.38	0.32
5	Mean	0.361	0.566	0.421
	Standard deviation	0.521	0.26	0.35

Table 9 shows apparent differences in the mean of the absolute difference between the estimated and actual difficulty parameters based on the two study variables. This may be because the three-parameter models account for the guessing parameter, unlike the binary model, which assumes the absence of guessing—a difficult assumption to achieve in practice, especially with multiple-choice test items. This made the triple model more accurate in estimating the difficulty parameter. To verify the nature of the differences, a binary variance analysis of the absolute difference between the estimated and actual difficulty parameters was calculated based on the two study variables.

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table10. Results of binary variance analysis of the absolute difference between the estimated and actual difficulty parameter following the two study variables

Variance Source	The Squares Sum	The calculated "F" Value	The Average of The Squares Sum	The calculated "F" Value	The Error Probability
Models	36.123	2	16.1020	139.357	0.012
Alternatives	2.310	2	1.021	8.527	0.010
Models x Alternatives	7.421	4	1.822	15.304	0.010

Table 10 shows statistically significant differences at the significance level (α=0.05) in the means of the absolute difference between the estimated and actual difficulty parameters due to the two study variables. This is because the one-parameter model assumes that discrimination is equal for all items and equals one, which means that the means of discrimination equal one.

The binary model shows means following the generation of actual data equal to (0.8). This is because the regular distribution of the discrimination parameter was assumed, with an initial value of (0.4) and a final value of (1.2). Thus, the mean of the discrimination parameter in the one-parameter model is larger than the mean of the discrimination parameter in the binary model. This made the one-parameter model more accurate in estimating the difficulty parameter of the items.

Table 11. Values of relative efficiency in estimating the parameter of item difficulty in different situations following the two study variables

Model	One			Binary			Triple			Variance of Estimated Difficulty Parameter
Alternative Numbers	3	4	5	3	4	5	3	4	5	
Variance of Estimated Difficulty Parameter	0.207	0.330	0.396	0.721	1.186	1.042	0.695	1.037	0.830	
Model / Alternative Numbers										
One Model										
3										0.207
4	0.626									0.330
5	0.522	0.833								0.396
Binary Model										
3	0.287									0.721
4	0.200	0.319		0.695	1.145		0.671			1.186
5	0.198	0.317	0.380	0.691	1.138					1.042
Triple Model										
3	0.297			1.036						0.695
4	0.200	0.319		0.695	1.145		0.671			1.037
5	0.249	0.398	0.478	0.868	1.429	1.256	0.838	1.249		0.830

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 11 notes that the relative efficiency ratio varies depending on the two study variables. When the efficiency ratio is less than one, preference is given to the numerator, and when it is greater than one, preference is given to the denominator.

The second question is: "Are there any significant differences (α=0.05) in the estimation of the discrimination parameter due to the two study variables (model and number of alternatives)?" The means and standard deviations of the estimated discrimination item parameter were calculated following the two study variables (model and number of alternatives).

Table12. Means and standard deviations of the values of the estimated discrimination item parameter following the two study variables

Alternatives Number	Statistical	Models	
		Binary	Triple
3	Mean	0.651	1.102
	Standard deviation	0.123	0.20
4	Mean	0.602	0.752
	Standard deviation	0.122	0.34
5	Mean	0.533	1.031
	Standard deviation	0.211	0.210

Table 12 notes that there are differences in the mean values of the discrimination parameter estimated following the two study variables. This is because the two-parameter model assumes the absence of guessing for all items, meaning that a non-proficient examinee does not score by guessing. Therefore, the discrimination of items will be greater between proficient and non-proficient examinees. Based on Table 12 results, the means and standard deviations of the absolute difference between the estimated and actual discrimination item parameters, following the study variables, were calculated.

Table13. Means and standard deviations of the absolute difference between the estimated and actual item marking parameter following the two study variables

Alternatives Number	Statistical	Models	
		Binary	Triple
3	Mean	0.165	0.254
	Standard deviation	0.120	0.216
4	Mean	0.321	0.012
	Standard deviation	0.174	0.210
	Mean	0.123	0.021
	Standard deviation	0.120	0.100
5	Mean	0.210	0.102
	Standard deviation	0.15	0.16

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 13 shows differences in the mean absolute difference between the estimated and actual discrimination parameters following the study variables. To verify these differences' significance, a two-way variance analysis was conducted for the absolute difference between the estimated and actual discrimination parameters based on the study variables.

Table 14. Results of the two-way analysis of variance for the absolute difference between the discrimination parameter and the actual following study variables

Variance Sources	Squares Sum	Freedom Degrees	Average of Squares Sum	Calculated "F" Value	Error Probability
Model	0.124	1	0.127	3.195	0.031
Alternative	0.062	2	0.102	1.423	0.230
Model X Alternative	0.230	2	0.125	5.920	0.004
Error	7.268	294	0.0120		

Table 14 shows that there are statistically significant differences at the significance level (α=0.05) in the means of the absolute difference between the estimated and actual discrimination parameters, attributed to the variable (model). The significant differences favor the binary model over the triple model.

Table 15. Relative efficiency values for estimating the discrimination parameter in different cases following the study variables

Model	Binary			Triple			Variance of Estimated Difficulty Parameter
Alternative Numbers	3	4	5	3	4	5	
Variance of Estimated Difficulty Parameter	0.034	0.150	0.021	0.085	0.056	0.102	
Model / Alternative Numbers							
Binary Model							
3							0.049
4	0.895						0.042
5	1.652	1.266					0.032
Triple Model							
3	0.452						0.090
4	0.621	0.734		1.322			0.057
5	0.429	0.325	0.532	0.812	0.412		0.112

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

Table 15 notes that there are differences in the relative efficiency ratio, following the study variables. When the efficiency ratio is less than one, the numerator is preferred, and when the relative efficiency ratio is greater than one, the denominator is preferred.

The third question is: "Are there statistically significant differences (α=0.05) in estimating the item guessing parameter attributed to the study variable, the number of alternatives?" To answer this question, the means and standard deviations of the estimated item guessing parameter were calculated following the study variable (alternatives). Table 16 shows the results of the analysis.

Table 16. Means and standard deviations of the estimated item guessing parameter following the levels of the study variable

Alternatives	Means	Standard Deviations
3	0.321	0.042
4	0.421	0.052
5	0.321	0.051

Table 16 shows differences in the mean of the estimated item guessing parameter, following the levels of the study variable. The results of Table 16 showed that the means and standard deviations of the absolute difference between the estimated and actual item guessing parameters were calculated following the levels of the study variable.

Table17. Means and standard deviations of the absolute difference between the guessing parameter for the estimated and actual item following the levels of the study variable

Alternatives	Means	Standard Deviations
3	0.062	0.04
4	0.081	0.03
5	0.077	0.05

Table 17 shows that there are differences in the mean of the absolute difference between the estimated and actual guessing parameters following the levels of the study variable. To verify the significance of the differences, a one-way analysis of variance was conducted to detect the absolute difference between the estimated and actual guessing parameters following the levels of the study variable.

Table 18. Results of the one-way analysis of variance for the absolute difference between the estimated and actual guessing parameter following the levels of the study variable

Variance Sources	Squares Sum	Freedom Degree	Average of Squares Sum	Calculated "F" Value	Error Probability
Alternatives	0.032	3	0.021	4.854	0.006
Error	0.421	137.121*	0.013		
Total	0.256	149			

Table 18 shows that there are statistically significant differences at the significance level (α=0.05) in the means of the absolute difference between the estimated and actual guessing parameters attributed to the "number of alternatives" variable. Since the variable is multilevel, the Brown-Forsythe test was conducted to detect any violation of variance homogeneity due to inhomogeneity

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

in the means of the absolute difference between the estimated and actual guessing parameters, following the levels of the study variable.

Table 19. Results of the Games-Howell test for multiple comparisons of the mean of the absolute difference between the estimated and actual guessing parameter following the levels of the study variable

Alternatives	4	3	5	Means
Games-Howell				
Means	0.071	0.073	0.099	
3				0.071
4	0.003			0.073
5	0.028	0.025		0.099

Table 19 shows that there are statistically significant differences in the means of the absolute difference between the estimated and actual guessing parameters. In light of the above, the values of the RMSE accuracy index for the estimated guessing parameter were calculated using the equation, $RMSE_c = \sqrt{\sum_{j=1}^{n} \left(\hat{c} - c\right)^2 / n}$. This equation is relative to the actual guessing parameter following the levels of the study variable.

Table 20. RMSE values of the estimated guessing parameter relative to the true guessing parameter, following the levels of the study variable

Alternatives	Triple Model
3	0.075
4	0.074
5	0.102

Table 20 notes that the highest value of the RMSE for the estimated guessing parameter relative to the true guessing parameter was 0.102 when there were five alternatives. The lowest value of the RMSE was 0.074 when there were four alternatives.

5. Conclusion and Recommendations

The study results showed that the three-parameter model was more accurate in estimating the difficulty parameter, followed by the single-parameter model, and then the two-parameter model. Regarding the discrimination parameter, the results showed that the three-parameter model was more accurate than the two-parameter model. Additionally, the results showed that the guessing parameter is only related to the three-parameter model. The estimated guessing parameter was more accurate in the five-alternative tests, followed by the three-alternative tests, and then the four-alternative tests.

In light of this study's results, the study recommends conducting studies on the reliability of item-parameter estimations across different samples of individuals based on their abilities. Comparative studies should be conducted between the classical measurement theory and the item-response theory regarding the accuracy of estimating the psychometric properties of items, such as difficulty, discrimination, and guessing. More studies should be conducted on four-alternative tests, comparing

https://doi.org/10.22521/edupij.2025.14.54 Published online by Universitepark Press

their accuracy in estimating item difficulty for individual samples with high abilities, average abilities, and random abilities.

Declarations

Author Contributions. All authors have read and approved the published on the final version of the article.

References

Aiken, J. (1982). Testing with multiple-choice items. *Journal of Development in Education, 20*, 44–57.

Allam, S. E.-D. (2000). *Educational and psychological measurement and evaluation: Its basics, applications, and contemporary trends*. Cairo: Dar Al-Fikr Al-Arabi.

Allam, S. E.-D. M. (2005). *One-dimensional and multidimensional test-item response models and their applications in psychological and educational measurement*. Cairo: Dar Al-Fikr Al-Arabi.

Al-Rabba'i, I. M. (2012). The effect of the number of alternatives and changing the position of the strong camouflage in multiple-choice items on the psychometric properties of the test, the parameters of the items, and the abilities of individuals (Unpublished master's thesis). Yarmouk University, Jordan.

Al-Sharifain, N., & Taamneh, I. (2009). The effect of the number of alternatives in a multiple-choice test on the ability estimates of individuals and the psychometric properties of the items and the test following the Rasch model in the item response theory. *Jordanian Journal of Educational Sciences, 5*(4), 309-335.

Blunch, N. J. (1984). Positional bias in multiple-choice questions. *Journal of Marketing Research, 21*, 216–220. https://doi.org/10.1177/002224378402100210

Crehan, K., Haladyna, T., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53*, 241–247. https://doi.org/10.21449/ijate.421167

Frisbie, D. A., & Sweeney, D. S. (1987). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*, 29–35. https://doi.org/10.1111/j.1745-3984.1982.tb00112.x

Fu, Q. (2010). Comparing the accuracy of parameter estimation using IRT models in the presence of guessing (Unpublished PhD dissertation). Illinois University, USA.

Gay, L. R. (1980). *Educational evaluation and measurement: Competencies for analysis and application*. Ohio: Charles E. Merrill Publishing Company.

Hambleton, R. K., & Swaminathan, H. (1991). *Item-response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Jasper, F. (2010). Applied dimensionality and test structure assessment with the START-M mathematics test (Unpublished doctoral dissertation). *The International Journal of Educational and Psychological Assessment, 6*(1), University of Mannheim, Germany.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of items & statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921–943. https://doi.org/10.1177/0013164402238082

Miller, T. (1991). Empirical estimation of standard errors of compensatory MI model parameters obtained from the NOHARM estimation program (ACT Research Report No. onr91-2). Iowa City, IA: ACT Inc.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). New Jersey: Prentice-Hall, Inc.

Odah, A. S. (2014). *Measurement and evaluation in the teaching process*. Irbid: Dar Al-Amal for Publishing and Distribution.

Reckase, M. D. (1978). A comparison of the one and three-parameter logistic models for item calibration. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada. Available at http://eric.ed.gov

Yaman, S. (2011). The optimal number of choices in multiple-choice tests: Some evidence for science and technology education. *The New Educational Review, 23*(1), 227–241.

About the Contributor(s)

Aiman Mohammad Freihat Ajloun University College, Al-Balqa Applied University, Jordan
E-mail: aiman.freihat@bau.edu.jo
ORCID: https://orcid.org/0000-0002-4161-4143

Omar Saleh Bani Yassin Irbid University College, Al-Balqa Applied University, Department of Educational Sciences.
E-mail: omarsa@bau.edu.jo
ORCID: https://orcid.org/0000-0002-4495-1030