Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.13, No.13, 2022

The Impact of the Number of Distractors in Multiple-choice Test Items on the Psychometric Characteristics of the Items and Item Information Function According to the Two-Parameter Logistic Model in the Item Response Theory

Dr. Habis Sa'ad Al-zboon
Al Hussein bin Talal University, Faculty of Educational Sciences
Maan, Jordan, habis.s.alzboon@ahu.edu.jo

Abstract
The aim of the current study was to identify the impact of the number of distractors in multiple-choice test items on the psychometric characteristics of the test items and the item information function according to the two-parameter logistic model in the Item Response Theory (IRT). To answer the study questions, an achievement test was built covering all school tests course that measures teachers' degree of knowledge to build school tests. The test consisted of (42) items and was of 3 different forms varying in the number of options. The study sample consisted of 356 male and female teachers from Ma'an governorate and Wadi al-Sair district and was selected in a simple random manner. The statistical program (SPSS) and the (Bilog-mg3) program were used to analyse the responses of the study sample. The results of the study showed that there were statistically significant differences between the means of the values of the test items information function in favor of the four-distractors test form compared to the three-and-five- distractors test form. The results, furthermore, showed that there were statistically significant differences between the means of the values of the test items information function in favor of the five-distractors test form compared to the three-distractors test form.The results, further, showed that there were no differences between the means of the estimate of the items difficulty parameter due to the number of distractors of the item. It was indicated that there were differences between the means of the estimate of the items discrimination parameter concerning the three-and-four-distractors tests in favor of the four-distractors test. It was, besides, shown that there were statistically significant differences between the means of the estimate of the items discrimination parameter concerning the three-distractors and five- distractors tests' forms in favor of the for the five- distractors test.
Keywords: number of distractors, multiple-choice test, item response theory, items' parameters, information function
DOI: 10.7176/JEP/13-13-07
Publication date: May 31st 2022

Study Background:
Achievement tests occupy a prominent position in the educational process, and a key element thereof. It is not restricted to the learner, nonetheless, it extends to the teacher and everyone involved in the educational process. Tests are one of the most important means of assessing and evaluating students and knowing their achievement level. Therefore, educators and researchers showed interest in ensuring that such tests have a high level of efficiency and that this efficiency comes through the preparation of accurate and objective tests (Al-Tarawneh, 2018).

Allam (2011) pointed out that achievement tests assess students' assimilation of some of the knowledge, concepts, and skills related to the subject matter, as well as the educational achievement points to the status of an individual's performance, learning, or educational program already acquired.

Tests improve decision-making on the teaching process, increase the level of memorizing information, the improvement of students' level of education, improve students' motivation, increase students' self-recognition, and provide feedback on the effectiveness of the teaching process. Henceforth, a range of factors and considerations must be taken into account when preparing these tests. For instance, to be representative of the teaching subject, be consistent with the teaching objectives, and that the test should be formulated in a manner that commensurate with the objectives for which the results are used. Another factor is that the test has to have an acceptable degree of validity and reliability, and to have a test of the quality of questions that are more appropriate than others to assess the desired educational outcomes. It is the educational outcomes of a particular module that determine the types of behavior that can be accepted as evidence of achieving teaching objectives. The achievement test is only a means of invoking specific behavior through which judgements can be given about the extent to which the expected study objectives are achieved. Thus, effective measurement of achievement is affected by the choice of the quality of questions that will elicit the required answer and exclude other answers that are not relevant to the correct answers (Odeh, 2010).

Achievement tests have several varied forms, including written and non-written tests; written tests include,

short (restricted) answer tests, long (open) free answer tests, matching tests, true or false tests, multiple-choice tests, and multiple right-or-false tests (Abdul Hadi, 2001; Thorndike, 1982).

Frisbie and Sweeney (1982) claimed that multiple-choice items to be the most commonly used to assess a student's achievement in several educational purposes. They outweigh items of true or false tests, which assess the students' achievement, and that their results are highly valid and reliable.

The multiple-choice questions consist of a question or a major problem involving the main idea and a number of options or distractors put forward to stimulate the examiner's thinking so that these options include the correct answer and distractors or options. Therefore, the design, formulation, and number of these distractors directly affect the choice of the correct answer and thus the item discrimination coefficient, which accordingly affects the reliability and validity of the test as a whole (Al-zayat, 1989).

Although there are many features of the multiple-choice test items, a number of criticisms have been made, including:

The problem of the number of distractors, the problem of biasness to the place of the option and the correct answer among the strong distractor and the rest of other distractors. This, in turn, affects the response of the examiners and the test items. Therefore, it should be noted that the choice of examiners for the correct option depends on the place of the correct option in relation to other options, and on the content of the test items (Blunch, 1984).

The Item Response Theory (IRT) or Latent Traits Theory (LLT) is of the modern trends of measurement and evaluation. They have received the attention of many researchers, overcoming many traditional measurement problems. Psychological and educational tests assume that certain all individuals share traits or characteristics but they vary in their amount. Although one could not observe these traits, their existence could be inferred from observable behavioral manifestations or changes, which justifies their designation as latent traits (Abu Awwad, 2018).

Like the classical theory, the IRT assumes that the response to test items is attributed to latent traits. The IRT is defined as a theory of statistical estimation that uses the underlying traits of individuals and the item as predictors of seen responses, and is essentially a decline (two or multi-logistic decline) of the responses observed in the item on the order of individuals on the latent variable and the traits of the item (Ayala, 2009).

On one hand, the importance of the IRT and its applications in analyzing the item lies in reaching the parameters of a relatively fixed item that does not change as the group of examiners used in the analysis of the item changes. If the parameters are relatively fixed, they could be estimated from one set of data, and then it could be applied with confidence to any group of examiners, including the total population of examiners. On the other hand, in the classical theory, the statistics of traditional analysis of the item change as the examiners change (Crocker & Algina, 2017).

The IRT, with its different models, overcomes the problem of selecting items according to classical methods by introducing the method of selecting items. It further overcomes the existing problem by the ability of the examiner to develop the measuring scale of selecting the most effective items in a range determined by a cut-off mark on the scale that helps separating the levels of mastery on the scale (Hambleton & Swaminathan, 1991). The dichotomous IRT models vary based on the three parameters of the item, from which a set of mathematical models known as latent trait models emerged, each based on a mathematical formula that determines the relationship of an individual's performance to latent his/her ability (Hambelton, Swaminathan, 1991; Embretson & Reise, 2000).

One Parameter Logistic Model- 1PLM

One Parameter Logistic Model, also known as the Rasch Model, is considered among one of the most common used models in the IRT which assumes that all items differ from each other only by the item difficulty parameter and assuming that the item discrimination parameter is equal to all items. While the item guessing parameter is almost equal to zero. The model takes the following equation to reflect the likelihood of the correct answer to the item:

$$p_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}},$$

$$i = 1,2,3.....,n$$

Two-Parameter Logistic Model (2-PL)

The two-parameter logistic model, which was proposed by (Birinbaum, 1968), includes item discrimination parameter in the model which could be obtained from 3-PL IRT model when the pseudo-chance parameter is assumed zero.

The 2-PL IRT model is expressed in the flowing equation.

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \quad i = 1,2,3.....,n$$

Where:

$P_i(\theta)$: the probability that an examinee with ability θ answers item i correctly.

θ : ability parameter of I th examine.
a_i : item discrimination parameter.
bi : item difficulty parameter.
D : 1.7 (scaling factor)
θ : trait level for person i
n: number of examine.
e=2.718(Napier constant)

Three Parameter Logistic Model- 3PLM

The three-parameter logistic model is an extension of the Two-parameter logistic model, derived by adding the item guessing parameter. The three-parameter logistic model represents the overall shape of the logistics models, because it contains the three possible parameters of the item, difficulty, discrimination, and guessing (bi, ai, ci) respectively, expressed through the following equation, which measures the probability of the answer of the examinee of the ability of (Θ) to item (i).

$$p_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

$P_i(\theta)$:the probability that an examinee with ability θ answers item i correctly .

b_i : item difficulty parameter , a_i : item discrimination parameter , c_i item guessing parameter, θ : trait level for person i.
e=2.718(Napier constant), D : 1.7 (scaling factor)

Assumptions of Item Response Theory:

Models of the IRT assume a number of assumptions in the data to be applied, including:

1. Unidimensionality:

It does not mean the simplicity of the variable in question, i.e., the simplicity of what the item measures, rather that the items of the scale are homogeneous and measure essentially the same trait. This means that an item of the gradual difficulty items requires the same type of procedures and behavioral processes, however they differ only in terms of their difficulty.

Hambleton and Swaminthan (1989) considered this assumption difficult to achieve, as there are some factors affecting the performance of individuals on testing, such as the motivation level, test anxiety, ability to respond quickly, test-wiseness, and guessing in answering certain test items.

2. Local Independence:

This means that an individual's responses to different items of the test are statistically independent, which means that an individual's response to one item does not affect his or her responses to other items. This is evident in:

a. The measurement shall be free from the distribution of the sample-free. This means the reliability and constancy of the measurement of the parameters of the item, although the sample of individuals used in the measuring scale varies as long as it is appropriate.

b. The measurement shall be free from the used (Item-Free) set. This means that an individual's ability and constancy are reliable, notwithstanding the different set of items used for measurement, insofar as they are appropriate, and as long as these different sets of items fall on a single scale, i.e. one variable.

3.Item characteristics curve (ICC):

The characteristic curves of the item are mathematical functions that link a person's probability of success in answering an item to the ability measured by the group of items included in the test. It could be said that it is the decline in the degree to which an individual in one item acquires his or her ability.

The item information function (IIF) is one of the most important statistics in the modern theory of measurement, whereby a standard error in estimation could be identified relying on the maximum likelihood of the ability parameter. The discrepancy in the estimation of the ability is equal to the inverse information function, and the variance covariance matrix for estimates is equal to the inverse of the information matrix of the estimates of the item's parameters (Hambleton & Swaminathan, 1991).

The test information function is an important concept in the modern theory of measurement, by which a standard error in estimation could be identified. It a represents the sum of the items' information functions at a certain level of ability, since it is not dependent on the sample of examinees. Consequently, modern theory of measurement provides additional advantages, in terms of increasing the ability of estimating the measurement errors (Brannick, 2003).

A worthwhile use of the item information function is the opportunity of identifying the extent to which each item contributes to the test information function without relying on the other items of the test. If we have a good idea of the abilities of the group of examinees, the test items, that maximize the information provided by the test, could be selected to the extent to which the test examinees' abilities are distributed.

The item information function is used to select items when building tests, drawing on modern theory of measurement, assuming that item information changes across different levels of traits. Thus, it is possible to select items that provide high measurement accuracy at a particular point on the trait continuum. In addition, items that contain significant discrimination parameters provide greater information on the ability of examiners, thus, the information function of the test is given through the following equation:

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}$$

$I(\theta)$: : test information function.

θ: trait level for person i .

$P_i(\theta)$: the probability that an examinee with ability θ answers item i correctly.

$$Q_i(\theta) = 1 - P_i(\theta)$$

$P_i'(\theta)$: first derivative of $P_i(\theta)$

Where the test items could therefore be selected based on the amount of information that the items contribute to the total amount of test information.

The information function is influenced by the parameters of the item. In one and two-parameter logistic models, the highest amount of information is at the difficulty parameter of the item (when ability is equal to difficulty). This is because the form of the item information function is generally close to the bell shape.

In IRT, the general interest is the estimated value of ability parameter for an examinee. The amount of information based on an item is able to be computed for any ability level. Item information function is shown as $I_i(\theta)$: $_{I = 1, \ldots\ldots, n}$ where n is the number of examinees. 2-PL is shown as:

$$I(\theta, a, b) = a^2 P(\theta) Q(\theta)$$

As it is clearly seen in the equation, discrimination parameter value has importance in computing item information function (Baker, 2001).

There are several studies on the number of options in the multiple-choice items and their impact on the psychometric properties of the test and the information function, including a study conducted by Elboni (2007). The study aimed at identifying the impact of the number and attractiveness of options in multiple-choice items in conformity with the three-parameters model, and in order to achieve the objectives study, a math test for the basic ninth grade was built in three forms each of (50) items, where the first form has five options. The second form has three options, by deleting two options randomly for the first form and deleting the least distinct options from the first form. The three forms were applied to a sample of (1656) male and female students in public and private schools of Irbid Second Directorate of Education, distributed to (20) schools selected in a simple random manner.

The results showed the compatibility of the items with the three-parameter model of the test in its three forms. They further showed that it is preferred to use multiple-choice tests with five options rather than the three-option tests regardless of the elimination method.

AL- shrifin & Taamnah (2009) conducted a study investigated the effect of multiple choice test number of alternatives on the estimation of a person's ability and the psychometric properties of a test and its items according to Rasch model in item response theory (IRT),To achieve the study objectives, a multiple choice achievement test consisting of 40 items in tenth grade maths was constructed.

The test took three different forms in terms of the number of alternatives: Form One with three alternatives, Form Two with four alternatives, and Form Three with five alternatives.

The test forms were applied on a total sample of 600 male and female students, with 200 students assigned to each form.

Data obtained for each form of the test were analyzed separately using (BIGSTEPS) and (BILOG-MG).

The study findings revealed the following: there were no statistically significant differences (α=0.05) among the standard error means of item in the estimation of difficulty parameters, and no statistically significant differences (α=0.05) among a person's reliability coefficients due to number of alternatives. However, item reliability coefficients were equal and there were statistically significant differences (α=0.05) among the standard error means of a person's ability parameters, and such estimations in a person's ability were more accurate in Form One of the test than in Form Three, and in Form Two of the test than in Form Three whereas such estimations were similar in Forms One and Two.

In addition to that, Form One of the test yielded more information at the low ability level but at the medium and high ability levels, the most information was in Form Three of the test, and there were statistically significant differences (α=0.05) among criterion validity coefficients in favor of Form Two of the test.

The study of Tarrant & Ware (2010) aimed at comparing the psychometric characteristics of multiple-choice tests (three options, four options). To achieve the study's objectives, the researchers applied a multiple-choice test to a pilot sample to ascertain the psychometric characteristics of the test. The final form of the test was prepared, consisting of 41 items in each form. The results of the study indicated that the three-option test was more effective despite the lack of distractors due to the strength of the distractors. They furthermore indicated that the distractors used in the study became highly discriminatory when weak distractors were eliminated.

Bani Atta & AL-Rabaei (2013) conducted a study aimed at verifying the effect of the number of alternatives and changing the strong distractor position on the items parameters, person's ability and information function. To achieve this aim, a 41 item multiple – choice achievement test in 10th grade mathematics was constructed. The test had four different forms according to number of alternative and the strong distractor position. The responses of 2111 examinees on the four forms were analyzed by Bilog–Mg3 programs according to the three parameter logistic model. The results of two-way ANOVA revealed no significant statistical differences between the means of item difficulty and guessing parameters. Differences were found, however, between the means of item discrimination, due to the position of the strong distractor; the differences were in favor of the second form. No differences were found in the means person's ability parameter, but the information function of the test varied due to the test forms. Significant differences were noticed in the values of empirical reliability coefficients in favor of the second form.

Abu Musallam (2016) conducted a study that aimed at identifying the impact of both the number of options and their arrangement in multiple-choice achievement test items on the test's psychometric characteristics and items. The sample of the study consisted of (113) students from the Research Institute of Cairo University, and the study tool consisted of (50) multiple-choice items with three different forms in the number of options per item.

One of the most important findings of the study was that the validity and reliability of multiple-choice test increases when the test has three options. The more difficult the test items, the more options of the item, and the more discriminatory power the item is, the more options of the item are.

Abdel-Aal & AL-Enezi (2019) conducted a study that aimed at identifying the effect of the item format of multiple-choice and true/false tests on psychometrics characteristics according to the IRT of the computer test for the first secondary grade in the city of Tabuk. Two forms of computer test have been created, multiple-choice and true/false tests, consisting of (24) items. The test was been applied on (421) students. The results showed that there were no statistically significant differences in the item difficulty between the test forms, however the results revealed statistically significant differences in the item discrimination and in favor of true/false test. However, the results showed that there were no statistically significant differences concerning the empirical reliability between test forms

Study Problem:

The accuracy of the test and its items' psychometric characteristics depend on the answers of the test items. Hence, one of the problems the researchers face when building multiple-choice tests is the number of answer distractors, as there is no agreement on their number even at the same school level. Through a review of previous literature, the researcher noticed that studies on the issue of the number of distractors were based on the classical theory of measurement and that studies based on modern theory relied on the three-parameter model. Consequently, this study came to identify the impact of the number of distractors in multiple-choice test items on the psychometric characteristics of test items and the function of items information function according to the two-parameter model defined in the IRT. The problem of the current study could be determined by answering the following questions:

Question 1: Are there statistically significant differences at the significance level of (α≤0.05) between the means of the items information function based on the two-parameter model defined by the number of distractors?

Question 2: Are there statistically significant differences at the significance level of (α≤0.05) between means of the items parameters estimates (difficulty, discrimination) based on the two-parameter model according to the

number of distractors?

Importance of the study:
The importance of this study stems from the topic it discusses which is about a very important topic, namely the impact of the number of distractors in multiple-choice test items on the psychometric characteristics of test items and the item information function according to the two-parameter model defined in the IRT. As the expected results would benefit researchers and test-takers in determining the appropriate number of distractors specific to each item so that the items have good psychometric characteristics, which helps to obtain a test with good psychometric characteristics and gives a high information function that increases the accuracy of measurement. It is, moreover, expected that the study would enriching knowledge in this area and be the start of new studies regarding the number of distractors in test items.

Objectives of the study:
1. Identifying the differences between the means of item information function's values based on the two-parameter model defined by the number of distractors.
2. Identifying the differences between the means of estimating the parameters of items (difficulty, discrimination) based on the bi two-parameter model as the number of distractors varies.
3. Identifying the adequate number of distractors that achieve good test characteristics.

Definition of terms:
Distractors: Number of incorrect options in multiple-choice test item.
Item parameter: It is the parameter of difficulty, discrimination, arising from the two-parameter model.
Multiple-choice Test: Tests consisting of a number of items. The item consists of two parts. The first part is called content. The second part is made up of options or distractors and serves as a solution or possible answers to the question in the content (Thorndike, 2010).
Two-parameter model: It is one of the dichotomous models of IRT, through which the two parameters of difficulty and distinction could be estimated.
Test information function: The total number of information functions of test items. This study defines the test information function as the value obtained from the application of the following formula:

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}$$

Study Approach:
Study population:
The study population is made up of all teachers of Ma'an governorate and Wadi al-Sir district

Study Sample:
The study sample consisted of 356 male and female teachers from Ma'an Governorate and Wadi El-Sir district selected in a simple random manner.

Study Tool:
In order to achieve the objectives of the study, an achievement test was built in school tests that measures teachers' degree of knowledge to build school tests. To this end, the subject of school tests taught by the researcher to students of the General Diploma in Education at the University of Hussein bin Talal was used based on the following steps provided by (Gronlund & Linn, 1990):
First: the purpose of the test is to measure students' achievement in the subject of school tests.
Second: Writing the objectives of the content and the table of specifications, in which the levels of the objectives were linked to the course material of the subject understudy.
Third: Writing 50 multiple-choice items, with five options including one as the correct answer and four distractors for each item. The technical specifications of test-building were taken into consideration in writing this type of item and its relevance to the objective it measures in terms of content and knowledge level.
Fourth: Presenting the test in its tentative form to a group of arbitrators from the faculty of education specializing in measurement and evaluation, pedagogical psychology, and general curricula, at the University of Hussein bin Talal and other universities. Arbitrators were asked to express their views on the relevance of the items to the subject of the study and the clarity of the language and wording of the items.
Fifth: Conducting an analysis of the items after application to a pilot sample consisting of 30 male and female teachers and identifying the difficulty and discrimination coefficients of the test items.
Sixth: Based on the arbitrators' comments and relying on the values of the difficulty and discrimination coefficients, (8) items of the test were eliminated. Three (3) items were eliminated based on the arbitrators'

opinion and five (5) items of low discrimination coefficients were eliminated (less than 0.19), items with low difficulty coefficients (less than 0.2), and items of high difficulty coefficients (higher than 0.8). Thus, final form of the test consisted of (42) items. Bearing in mind that the remaining items after the elimination achieve their comprehensiveness and representation of the skills of the tests and the measurements that teachers are expected to observe during their measurement of their students, and Table 1shows the values of the difficulty and discrimination coefficients of the test items.

Table 1: Difficulty and discrimination coefficients' values for test items

Item Number	Difficulty coefficient	Discrimination coefficient	Item Number	Difficulty coefficient	Discrimination coefficient
1	0.57	0.40	25	0.58	0.29
2	0.53	0.0.31	26	0.66	0.41
3	0.77	0.57	27	0.13	0.09
4	0.42	0.56	28	0.43	0.42
5	0.47	0.25	29	0.51	0.58
6	0.3	0.11	30	0.44	0.31
7	0.73	0.23	31	0.67	0.42
8	0.80	0.37	32	0.63	0.51
9	0.61	0.28	33	0.63	0.26
10	0.68	0.41	34	0.53	0.42
11	0.63	0.39	35	0.60	0.63
12	0.59	0.40	36	0.14	-0.11
13	0.36	0.41	37	0.55	0.46
14	0.26	0.43	38	0.45	0.35
15	0.28	0.36	39	0.52	0.24
16	0.33	0.32	40	0.23	0.29
17	0.31	0.41	41	0.54	0.23
18	0.34	0.51	42	0.36	0.41
19	0.31	0.54	43	0.33	0.41
20	0.26	0.44	44	0.41	0.42
21	0.36	0.52	45	0.16	0.01
22	0.41	0.36	46	0.51	0.63
23	0.1	-0.12	47	0.52	061
24	0.61	0.25			

Seventh: The reliability of the 42-items test has been assessed after it has been applied to members of the pilot sample using the split-half method. To get rid of the split-half method effect the Spearman Brown equation was used. Based on the split-half method, the value of the reliability coefficient was (0.84), which indicates that the test has a high degree of reliability, which assures the researcher to apply on the study sample.

Eighth: One distractor was randomly removed from the five-distractor test items to form the second form of the four-distractor test. Another distractor was randomly deleted from the four-distractor test items to form the three-distractor test (the first form of the test).

Ninth: Preparation of test instructions and model answer sheet for each test form.

Tenth: Final application of the test to the study sample and collection, and analysis of the study sample responses using (SPSS).

Eleventh: Verification of the assumption of the IRT.

1. Unidimensionality assumption:

The unidimensionality assumption was verified by relying on the factor analysis of the examinees' responses using the principal component method and relying on the varimax method of rotating factors. The Eigenvalue and the ratio of explained variance as well as the cumulative explained variance of each factor are shown in table 2.

Table 2: Eigenvalues root, ratio of explained variance, and the cumulative explained variance of the test

Factor	Eigenvalue	explained variance ratio	cumulative explained variance ratio	$\frac{\text{First Eigenvalue}}{\text{Second Eigenvalue}}$
1	6.46	15.38	15.38	2.15
2	3.01	7.31	22.69	
3	2.20	5.25	27.95	
4	2.04	4.87	32.82	
5	1.85	4.40	37.23	
6	1.79	4.27	41.5	

Table 2 shows that the first factor explains the greatest ratio of variation compared to the rest of the factors. The eigenvalue root product of dividing the first factor by the second one is also found to be greater than (2). This indicates a predominant trait of the test. Thus, it could be claimed that the assumption of the unidimensionality hypothesis of the test is verified (Hattie, 1985).

The Scree Plot was used in the graphic representation of the factors for the Eigenvalue as shown in figure 1.

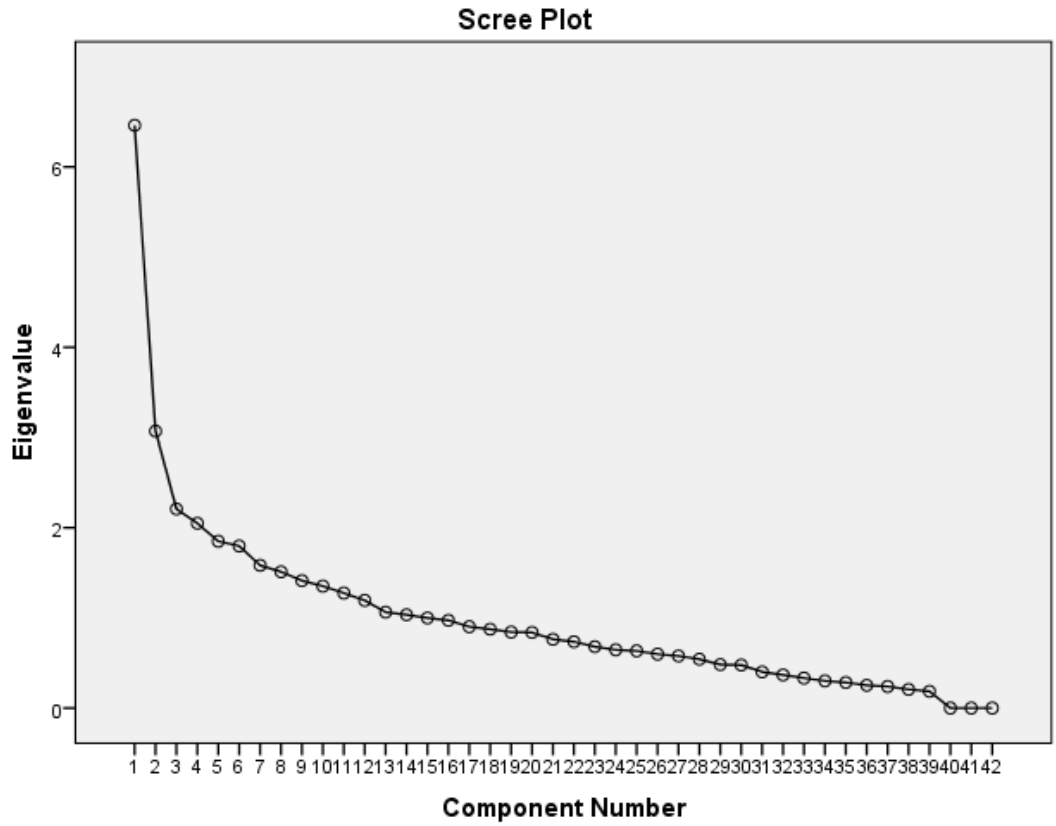

Figure (1): Graphic representation of the Eigenvalue of factors.

It is clear that the first factor explains the largest percentage of variance compared to other factors, and this is also an indicator of unidimensionality.

2. Local Independence:

This assumption states that the examinees' answer to the item is not affected by his response to another item. The only factor affecting the probability of a response to the item is the amount of (Θ) ability and the characteristics of the item. To verify this assumption, the unidimensionality assumption of test items has been adopted as evidence of the achieving the local independence assumption, as this is equivalent to a unidimensionality assumption (Hambleton & Swaminathan, 1985).

3. Speededness

It could be confirmed that the test is not a speed test by examining the percentage of examinees who completed the test, as well as examining items that students did not answer. According to Hambelton et al., (1991), if (75%) of students had answered the test, and if (80%) of the test items had been answered by students, speed would no longer be an important factor in performance on the test. Since the percentage of individuals who have completed this test is (100%), and the percentage of items answered is higher than (90%), this means that the test measures strength rather than speed.

Twelfth: Verification of the person and item fit to the two-parameter logistics model:

(Bilog-MG3) program was used to ensure that (Person-Fit) and (Item-Fit) conform to the two-parameter model using a Chi-squared test ($\chi 2$) at the significance level of (α=0.01). The results of the analysis showed that four individuals' responses did not match with the two-parameter model, with probability values below (0.01), their responses were, therefore, deleted to form the final sample study size of (352) examinees.

With regard to the examination of items' fit to the model used, it was reanalyzed using (Bilog-MG3) after the deletion of the individual's non-conforming responses based on the Chi-squared test ($\chi 2$) at the significance level of (α=0.01). The results of the analysis showed that all test items matched with the two-parameter model,

where the probability value was greater than (0.01) for all items.

Study results and discussion:

Question 1: Are there statistically significant differences at the significance level of (α≤0.05) between the means of the items information function based on the two-parameter model defined by the number of distractors?

To answer this question, the information function values and means were calculated for each item of all three-test forms using the two-parameter model marked by the number of distractors as shown in Table 3.

Table 3: Test information function for the three test forms

First form of the test (Three-option model)				Second form of the test (Four-option model)				Third form of the test (Five-option model)			
Item No.	Information Function	Item No.	Information Function	Item No.	Information Function	Item No.	Information Function	Item No.	Information Function	Item No.	Information Function
1	0.03	22	0.33	1	0.14	22	0.66	1	0.36	22	0.77
2	0.05	23	0.32	2	0.16	23	0.55	2	0.22	23	0.66
3	0.14	24	0.34	3	0.18	24	0.56	3	0.24	24	0.56
4	0.13	25	0.24	4	0.15	25	0.44	4	0.36	25	0.44
5	0.12	26	0.22	5	0.14	26	0.55	5	0.18	26	0.41
6	0.22	27	0.21	6	0.23	27	0.52	6	0.36	27	0.41
7	0.23	28	0.23	7	0.24	28	0.24	7	0.33	28	0.45
8	0.24	29	0.02	8	0.25	29	0.21	8	0.33	29	0.21
9	0.36	30	0.13	9	0.36	30	0.13	9	0.41	30	0.36
10	0.44	31	0.52	10	0.47	31	0.52	10	0.42	31	0.52
11	0.44	32	0.14	11	0.52	32	0.16	11	0.52	32	0.38
12	0.22	33	0.22	12	0.23	33	0.22	12	0.32	33	0.35
13	0.12	34	0.15	13	0.13	34	0.15	13	0.15	34	0.33
14	0.14	35	0.19	14	0.18	35	0.19	14	0.19	35	0.32
15	0.15	36	0.24	15	0.19	36	0.24	15	0.19	36	0.31
16	0.16	37	0.23	16	0.22	37	0.36	16	0.26	37	0.32
17	0.14	38	0.24	17	0.24	38	0.24	17	0.24	38	0.21
18	0.16	39	0.16	18	0.18	39	0.22	18	0.42	39	0.33
19	0.43	40	0.44	19	0.44	40	0.55	19	0.44	40	0.34
20	0.42	41	0.03	20	0.42	41	0.13	20	0.42	41	0.32
21	0.36	42	0.12	21	0.36	42	0.44	21	0.52	42	0.31
sum		9.4				12.7				15.2	
Means		0.22				0.30				0.36	

It is evident from Table 3 that there are apparent differences among the means of the test items information function values. One-way ANOVA analysis of variance was used to validate the significance of the differences among the means of test items information function values based on the two-way parameter model depending on the number of distractors of the test items (Three distractors, four distractors, five distractors) as shown in Table 4.

Table 4: One-way ANOVA analysis of variance of means of test items information function values according to the variable of the number of distractors for test items

Source of variance	Sum of squares	df	Mean Square	F-Value	sig
Between groups	0.39	2	0.199	10.72	0.000
Within groups	2.28	123	0.019		
Total	2.68	125			

Table 4 shows statistically significant differences at the level of (α≤0.05) among the means of the test items information function values on the three test form.

In order to determine the direction of these differences in, the multiple comparisons among the means of the test items information function on the three test forms using Scheffe's post-hoc comparisons test. The results were as shown in Table 5.

Table 5: Results of post-hoc comparisons among the means of test items information function values attributed to the variable of the number of distractors per item

Form	Form	Difference between means
1	2	-0.79*
	3	-0.14*
2	3	0.06*

Table 5 shows that there were statistically significant differences at the level of (α≤0.05) between the means of the test items information function on the three- distractor and four- distractor test forms and in favor of the four- distractor test form. The results likewise showed that there were statistically significant differences among the means of test items information function values on the three-distractor and five-distractor test forms in favor of the five-distractor test form. They further indicated that there were statistically significant differences at the significance level of (α≤0.05) among the means of test items information function values on the four-distractor and five-distractor test form in favor of the four-distractor test form. This confirms that the best test information function is with the four-option multiple-choice test. This could be attributed to the fact that by increasing the number of distractor of the test item this item increases the ability to distinguish between examinees and consequently to obtain high discrimination factors and thus the item will provide greater information on the ability of examinees that leads to a greater accuracy.

Question 2: Are there statistically significant differences at the significance level of (α≤0.05) between means of the items parameters estimates (difficulty, discrimination) based on the two-parameter model according to the number of distractors?

To answer this question, estimates of items parameters (difficulty, discrimination) for each item of test forms were calculated. The means of the items parameters estimates for the items of each of the three test forms were calculated based on the two-parameter model, according to the number of distractors as shown in Table 6.

Table 6: Means and standard deviations of the difficulty and discrimination parameters for the forms of the third test form

Test forms	Item parameters	Means	Standard deviation
1	Difficulty parameter	0.36	0.08
	Discrimination parameter	-1.86	1.68
2	Difficulty parameter	0.28	0.06
	Discrimination parameter	-1.17	1.46
3	Difficulty parameter	0.24	0.06
	Discrimination parameter	-0.86	1.34

Table 6 indicates that there are apparent differences between the estimates of the parameters of the items of each of the three test forms using the two-parameter model according to the number of distractors (difficulty, discrimination). To verify the significance of the differences, One-way ANOVA analysis of variance was used. Moreover, to identify the significance of differences between the means of the estimates of items parameters (difficulty, discrimination) for each item of the test forms were calculated using the two-parameter model according to the number of distractors (three, four, and five options) as shown in Table 7.

Table 7; One-way ANOVA analysis of variance of the means of the items parameters' estimates (difficulty, discrimination) according to the variable of the number of item distractors

Item parameters	Source of variance	Sum of squares	df	Mean Square	F-Value	sig
Difficulty parameter	Between groups	0.002	2	0.001	0.235	0.791
	Within groups	0.615	123	0.005		
	Total	0.617	125			
Discrimination parameter	Between groups	11.03	2	5.52	2.41	0.012
	Within groups	280.55	123	2.28		
	Total	291.58	125			

It is evident from Table 7 that there are no statistically significant differences at the significance level of (α≤0.05) among the means of the items parameters' estimates attributed to the number of the distractors for each item. It could also be attributed to the fact that the difficulty parameter depends on the ability of the examinees to respond to the test items keeping in mind that their abilities remain the same regardless of the number of distractors for each test item. It could further be attributed to the reliance of this study on the two-parameter model and therefore the guessing parameter remain constant among all examinees if the item difficulty parameter is estimated.

Table 7, furthermore, indicates that are statistically significant differences at the significance level of (α≤0.05) among the means of the estimates of the item discrimination parameter attributed to the number of item distractors. To determine the significance of the differences the multiple comparisons between the means of the

estimates of item discrimination parameter on the three forms of the test using Scheffe's post-hoc comparisons test. The results are shown in Table 8.

Table 8: Results of post-hoc comparisons among the means of test items information function attributed to the variable of the number of distractors per item

Form	Form	Difference between means
1	2	-0.64*
	3	-0.95*
2	3	-0.31

Table 8 shows that there are statistically significant differences at the significance level of ($\alpha \leq 0.05$) among the means of the item discrimination parameter estimate on the form of the three-distractor and four-distractor tests in favor of the four-distractor test. Similarly, the results show that there are statistically significant differences among the means of the estimates of the item discrimination parameter on the three-distractor, five-distractor test and in favor of the five-distractor test. This is because increasing number of distractors increases the test would be able to distinguish between those examined in the upper group and those examined in the lower group, i.e., the ability of the items to distinguish the individual differences of the examinees when relying on the test with four-options compared to the test with the three-options. Subsequently, increasing the number of options decreases the chance of guessing.

The results, besides, showed that there are no statistically significant differences at the significance level of ($\alpha \leq 0.05$) among the means of the estimates of the item discrimination parameter on four-and-five-distractors. This could be attributed to the reason that the appropriate number of distractors is four, as in case the number of distractors is more than four, test-makers might commit errors in the selection of inefficient and weak distractors, thus helping the examinee to exclude these distractors.

Conclusion and recommendations:

The results of the study showed that there were statistically significant differences among the means of the values of the test items information function in favor of the four-distractor test form compared to the three-and-five-distractor test forms.

They, furthermore, showed that there were statistically significant differences among the means of the values of the test items information function in favor of the five-distractor test form compared to the three-distractor test form.

However, the results showed no statistically significant differences among the means of the estimate of the test items difficulty attributed to the number of distractors of the item. Nonetheless, there were differences among the means of the estimates of the item discrimination parameter related to the three-distractor and four-distractor tests' forms in favor of the four-distractor test. Further, the results showed significant differences among the estimates of the item discrimination parameter related to the three-distractor and five-distractor tests' forms in favor of the five-distractor test. Therefore, the researcher recommended adopting the four-distractor test items in the preparation of the achievement tests, and conducting further studies based on the Rasch Model and the three-parameter model.

References

Abdel-Aal, S & AL Enezi,M (2019). The effect items format the multiple choice and true – false on psychometrics properties according to the (IRT) of the computer test for the first secondary Grade in The city of Tabuk, International journal of Educational and psychological studies,6(1), pp1-17.

Abdul Hadi, N. (2001). Low level of school achievement and achievement - causes and treatment, second edition , Amman: Dar Wael for printing, publishing and distribution.

Abu Awwad, F. (2018). Exploration of item parameter estimates, ability, and information function to test cognitive abilities using the three-parameter logistic model. Journal of Psychological and Educational Studies, University of Qasdi Marbah, 11 (1), 1-17.

Abu Musallam, M(2016). The Effect of The Number of Alternatives and Their Arrangement for The Vocabulary of Educational Achievement of The Multiple Choice Test on The Psychometric Characteristics of The Test and Its Vocabulary, *Universities Journal for Education and Psychology,* 14(2),154-187.

AL- shrifin,N & Taamnah, E.(2009) conducted a study investigated the effect of multiple choice test number of alternatives on the estimation of a person's ability and the psychometric properties of a test and its items according to Rasch model in item response theory (IRT), Jordan Journal of Educational Sciences,5(4),309-335.

Allam, S, (2011). Books of educational and psychological measurement and evaluation: its basics, applications and contemporary trends: 5th ed. Cairo: Dar Al-Fikr for printing, publishing and distribution.

AL-Tarawneh, S (2018). The Impact of the Number of Distracters in Multiple Choice Item Tests on the Estimation of the Psychometric Characteristics of the Items Using Item Response Theory, Al Hussein Bin

Talal University Research Journal, 4(1), p. 152-172.
AL-Zayat, F. (1989). The effect of the dispersal of alternatives in multiple selection questions on the validity and reliability of the test: analytical study, journal of the Faculty of Education, Mansoura University, Volume 11, p. 86-108.
Baker, F .(2001). The Basics of Item Response Theory , ERIC Clearinghouse on Assessment and Evaluation.
Bani Atta, Z &Al-Rabaie, I. (2013). The Effect of Alternatives Number and Changing the Position of the Strong Distractor on Items Parameters, Person Ability and Information Function, The Jordanian Journal of Educational Sciences,9(8), 319-333.
Bani Atta, Z, & AL- shrifin, A. (2017). The impact of Difficult Item's Location in Multiple-choice Test on its Psychometric Properties and student's Performance, Association of Arab Universities Journal for Education and Psychology, 15(3), p. 93-129.
Blunch, N.J. (1984). positional Bias in multiple – choice Question. Journal of Marketing Research, 21, 216-220.
Brannick, M. (2003). Basics of IRT one – linefile : //a:/ item response theory.htm.
Crocker, L.& Algina , J.(1987), Introduction to Classical and Modern Test Theory. University of Florida, U.S.A.
Crocker, Linda and Algina, James (2017). Introduction to classical and modern test theory. (Hind al-Hammouri and Zeinat Let's, translation;).
Elboni, S. (2007). Effect of Number of Alternatives and Distractors Discrimination in Multiple Choice Items on Fit in Three Parameter Model. Unpublished PhD, Yarmouk University, Jordan.
Embretson ,S. & Reiase , S. (2000). Item Response Theory for Psychologists.New jersey: Lawrence Erlbaum Associates,Inc.
Frisbie, D., & Sweeney, D. (1982). The Relative Merits of Multiple True–False achievement tests. Journal of Educational Measurement. Vol. 19(1), PP. 29–35.
Gronlund, N.E. and Linn, R.L. (1990) *Measurement and Evaluation in Teaching*. McMillan Company, New York.
Hambleton, K.R., & Swaminathan .(1991). Fundamentals of Item Response Theory: Sage Publications.
Hambleton, R., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Kluwer· Nijh off Publishing.
Hattie, J (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. Psychological Measurement, 9(2), 139-164.
Odeh, Ahmed (2010). Measurement and assessment in the teaching process. 4th ed., Dar Al-Amal. Irbid, Jordan.
Tarrant, M & Ware, J. (2010). A comparison of the psychometric of three and four – option multiple choice question in nursing assessment, Nurese education today,30, 359-543.
Thorndike, R. (1982). Applied Psychometrics. London: Houghton Mifflin Company Boston.
Thorndike, R. (2010). Measurement and evaluation in psychology and education. (8th ED). Upper saddle Rever.NJ: Pearson/Merril Prentice Hall.