Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 1209 – 1224

The extent of compatibility between the classical theory and the two-parameter logistic model in constructing an achievement test in mathematics for the eighth basic grade (A comparative study)

[1]Dr. Mohamad Talib Dabous

[1]Associate Professor of Measurement and Evaluation, Al-Istiqlal University- Jericho—Palestine, mohammad.dabous@pass.ps

Abstract

This study aimed at detecting the compatibility between the classical theory and the two-parameter logistic model in conformity the items of an achievement test in mathematics. To achieve the aim of the study, an achievement test was built in the mathematics subject of the eighth basic grade. The test consisted of (50) multiple-choice items, and it was applied to a purposive sample of (580) male and female students of the eighth basic grade in the Nablus Governorate. The researcher relied on the SPSS software to analyses the data according to the classical theory, while the MULTILOG.7 software was used to analyse the data according to the two-parameter logistic model. The results showed that (49) items conformity the classical theory, while all the items conformity the Item Response Theory according to the two-parameter logistic model. The value of the reliability coefficient of the test according to the classical theory was (0.921), while it was according to the two-parameter logistic model (0.953), while the validity coefficient according to the classical theory was (0.945) and according to the two-parameter logistic model (0.895).

Keywords: Achievement test, Classical Theory, Item Response Theory, Two -Parameter Logistic Model.

INTRODUCTION

Measurement is considered a basic element and an important element in the elements of the educational process in general and the teaching process in particular (Wang et al., 2021). The teachers in schools and the professors in university cannot undertake their basic role as evaluators without the availability of the minimum limit of the information and the basic skills in the domain of measurement and evaluation in general and the achievement test in particular (Jamalzadeh et al., 2021). Thus, interest is clear by the decision makers in rehabilitating the teachers in this domain pre-service and in-service. Moreover, rehabilitating the professors at the universities through programs which are directed at this purpose (Shephard et al., 2006).

Tests have occupied a big space in the psychological and educational researches. Since the movement of psychological measurement was found, psychologists have been interested in achieving the validity of the tests and psychological measurements and their reliability (Hambleton & Pitoniak, 2006). They sought to achieve the highest degree of objectivity in these instruments when using them in the measurement process (Abu-Hashim, 2006). English reference

Tests are considered one of the varied measurement means which can be relied upon in taking the important decisions which pertain

the individual and the society (Reference). The use of tests was widely spread in many domains and purposes such as choosing a person for a certain job, or for the purposes of classification such as specifying the course of the learners in a way which suits their abilities and their skills, and in evaluating the achievement of the learners through the grades which they obtain in the class tests (Reference). By this work can done to improve and develop the educational and the learning process and marching with them for the best by developing these tests and improving their ability to measure the results of learning, whether these tests are written or performative (Allen & Yen, 1979).

Since tests are an effective means for measuring educational achievement among the learners, the matter requires the necessity of being interested in them in order to help the teachers in improving their efficiency through acquiring concepts, information and skills which enable them to design and prepare tests which suit the abilities of the learners in order to evaluate their educational achievement and to judge the extent of their readiness and ability (Kathem, 2001).

Concerning interpreting the grades of the students' Allam (2001) pointed out that there are two varying directions in measurement to evaluate the performance of the students. The first direction is criterion -referenced in which the student is classified in the light of some statistics which are derived from the grades of the sample of the test which measures the trait in which he is classified with the aim of measuring the individual differences among the students. In this case the distribution of the grade must be equinoctial, and then the one who prepares the test omits the extremely difficult items no matter what the aims which the tests measure are which affects the validity of the test. The second direction is criterion-referenced which appeared because of the criticisms which were directed against the standard-referenced tests, whereby the criterion-referenced tests aimed to estimate the performance of the student in relation to a group of previously specified objectives regardless of the performance of his peers.

'Allam (2001) also sees that the criterion-referenced test are considered the most suitable kinds of achievement tests for measuring the achievement of the students because they specify the skills the mastery of which is required in extreme precision which enables the teacher to measure them and notice them directly, and then to estimate the extent of what the student achieved in these skills based on a specified performance level, which helps the teacher in the diagnosis process and to put the suitable treatment programs.

Tiratira (2009) pointed out that the process of classifying the learners in the criterion-referenced tests into two groups, one of the them is the mastery of the teaching content and the other is not mastery based on comparing the performance of the teacher by a degree which is called the cutting degree which separates between those who master and those who do not master the study content. Klein et al, (2009, p. 163) define the cutting degree by being a point on the measured of the observed degree which distinguishes between those who master and those who do not master.

Measurement was subjected to two theories, one of them is known as the traditional or the classical or the old theory and the other is known as the modern theory or the Item Response Theory (Reference). The classical theory is considered one of the most important and oldest methods which scientists reached to use it in the psychological measurement and the educational measurement, whereby it was used on a wide scope in developing many different psychological tests and many educational tests. This theory was founded by scientist Spearman in the year 1927, then this theory witnessed a wide development by scientist Melkisina in the year 1950.

Allam (2005) pointed out that the classical theory offered many solutions for the problems which faced educators in building the tests and developing them. However, it was unable to solve other problems in measurement such as it supposes that the standard error in measurement is equal for all the examinees, this lacks precision. Also expressing the ability of the individual is done through the real

degree which becomes clear through his performance in the test as a whole, and not on the level of the individual. Consequently, the position of the individual's ability will change according to changing the level of the test, in addition to this the characteristics of the test and the items will change with the change of the characteristics of the individuals. Also, the characteristics of the individuals change with the change of the characteristics of the tests in terms of difficulty and ease. This theory is not fit for building criterion-referenced test.

Based on the problems which were found in the classical theory and which led to building non-flexible tests, there appeared a new orientation in measurement which is based on mathematical models which depend on the theory of probabilities. Scientist Lord in the year 1952 put the foundations and hypotheses of the modern theory and which he named as the Item Response Theory. Then scientist Rasch in the year 1960 published the first model of this theory which concentrates on one parameter of the item which is the Item Difficulty. This model was called the Rasch model. By this, the psychological measurement of behavioural phenomena began to approach the physics measurement which is characterized by that the results of the measurement are not affected by the used instrument and by the elements which used this instrument as long as it is an instrument which is suitable for measuring that phenomenon (Lord, 1980).

The modern theory has helped in offering many solutions to problems related to buildings tests and developing them, and building questions banks and disclosing the bias of items, equivalence of tests and building criterion-referenced tests (Embreston & Reise, 2000).

Hambleton & Swaminathan (1985), pointed out that the modern theory is based on strong hypotheses which must be realized in the data in order to lead results which can be reliable. The first of these hypotheses is the Unidemensionality by which it is meant that the items measure one ability. The second hypothesis is Local Independence by which it is meant that the response of the examinee to a certain item does not affect his response to another item. The third hypothesis is the Item Characteristic Curve hypothesis which represents a mathematical significant which tie between the probability of the correct response to the item with the ability of the examinee which is measured through a group of items for a test which has been prepared for that purpose.

It is worth mentioning that from the modern theory there emanated from it several models all of which specify the relationship between the observed performance of the individual on the test and the trait or the ability which lies behind this performance and it interprets it and that the difference among these models is the number of parameters by which the item is described.

One of the most famous of these models is the Rasch Model, One-Parameter Logistic Model whereby it appears from this model that guessing equals nearly zero and that the discrimination is constant for all the items of the test, and that the difficulty of the item takes changing values (Reference).

As for the second model, it is the Burnoum Model, Two-Parameter Logistic Model in which this study is interested. This model supposes that both the parameter of difficulty and the parameter of discrimination are Variables, and that the guessing for all the items equals zero. Its arithmetical operations are distinguished by being more difficult than the Rasch Model. Its equation takes the following form (Hambleton & Swaminathan, 1985):

$$Pi(\Theta) = \frac{eDai(\theta-bi)}{1+ eDai(\theta-bi)}$$

i=1,2,3……..n

Where:

Pi(Θ): The probability that a randomly selected examinee with ability θ answers item i correctly.

Pi: item difficulty

ai: item discrimination

D= a constant which can be arbitrarily set. It is customary to set D= 1.7

The third model is the (Lord Model) Three-Parameter Logistic Model. This model is based on three parameters which are; the difficulty, the distinction and the estimation. This model is distinguished from the Two-Parameter Model in that it added the parameter of estimation which is the low approach line of the curve of the characteristics of the item.

Problem Statement

With the appearance of the modern theory in measurement and evaluation, specialists in measurement and evaluation began to do comparisons between the statistical indicators between the classical theory (the traditional theory) and the modern theory (the Item Response Theory), and this was in several attempts specially in preparing the achievement tests and the criterion-referenced tests. This was to indicate the extent of compatibility between the two theories in choosing the items of the test, and in indicating which of the two theories offers a value for the teaching process in its different stages.

The opinions of specialists and researchers in the domain of measurement and evaluation varied. Some studies pointed out to the surpassing of the modern theories in the figured out statistical indicators, and this was based on the classical theory such as the study of Salem (2011) and the study of Adedoyin (2010), while some studies indicated that there are no differences between the two theories in the figured out statistical indicators such as the study of Silvestre & Jimelo (2009). Many studies pointed out to the existence of contradictions in the psychometric characteristics of the tests such as validity and reliability.

Despite the existence of differences between the two theories, however there are no sufficient justifications for preferring one theory over the other theory. Thus this study came to compare between the two theories through building an achievement test in mathematics for the eighth basic grade, and to be acquainted with the psychometric characteristics of the test according to the classical theory and according to the Two-Parameter Logistic Model in the modern theory, and to make a comparison between the results of the two theories in selecting the items of the achievement test in mathematics in terms of difficulty, Item Discrimination the items, reliability and validity of the test.

Study questions:

Specifically, this study attempted to answer the following questions:

1. What is the extent of the conformity of the items of the achievement test in mathematics with the hypotheses of the classical theory in measurement and evaluation?

2. What is the extent of the conformity of the items of the mathematics test with the modern theory in measurement and evaluation (represented in the Two-Parameter Logistic Model)?

3. What are the psychometric characteristics (validity and reliability) of the mathematics test according to the classical theory and the modern theory (represented in the Two-Parameter Logistic Model)?

4. What is the extent of the compatibility between the classical theory and the modern theory (represented in the Two-Parameter Logistic Model) in choosing the items of the mathematics test?

Significance of the Study

Theoretical significance: The theoretical significance of this study lies in that it might contribute in supporting the theoretical basis of researches and studies related to comparing between the classical theory and the modern theory in measurement and evaluation (according to the Two-Parameter Logistic Model). Also, its theoretical importance lies in that it contributes in evaluating the extent of the quality of the psychometric characteristics of the items of the mathematics test through disclosing the extent of the conformity of the test items with the hypotheses of the classical

theory and the hypotheses of the Two-Parameter Logistic Model.

Practical significance: The practical importance of this study lies in developing a mathematics test having psychometric characteristics which can be used in disclosing the aspects of strength and weakness among the students in the mathematics subject. It is hoped that the results of this study will contribute in shedding more light on the problem of comparing between the two theories and how to complement each other which helps those who prepare the tests to use the best and the more reliable methods in analysing the items of the tests and which increase their reliability and validity.

The practical significance of this study lies in that the items of the mathematics test may become a nucleus for a questions bank in mathematics whereby this bank enjoys suitable psychometric characteristics according to the traditional theory and the modern theory in measurement and evaluation.

Aims of the Study:

This study aims at the following:

1. Knowing the extent of the conformity of the items of the mathematics test which is was prepared in this study with the classical theory and the modern theory according to the Two-Parameter Logistic Model in the modern theory.

2. Specifying the aspects of similarity and difference between the classical theory and the modern theory according to the Two-Parameter Logistic Model in terms of the difficulty of the items and their distinction.

3. Specifying the aspects of similarity and difference between the classical theory and the modern theory according to the Two-Parameter Logistic Model in terms of the reliability of the test and its validity.

Limitations of the Study:

-Human Limitation; this study is restricted to a sample from the students of the eighth basic grade

-Space Limitation: This study is restricted to a sample from the schools of the Nablus Governorate in Palestine

-Time limitation: This study was conducted in the first study semester of the school year 2021-2022.

Concepts and Terms of the Study

Achievement Test

It is an organized way to know the level of achievement of the students of information and skills in a certain study subject which was previously learned through their responses to a group of examination items which represent the content of the study material in a true representation (Abu-Judeh, 2018).

The Classical Theory in Measurement

It is a traditional theory in measurement which represents a simplified model for describing the manner in which the mistakes of the measurement affect the observed grades. This theory has several names including the traditional theory in measurement, or the classical theory in measurement (Hambleton & Jonse, 1993).

Item Response Theory

It is a modern theory in psychological and educational measurement in which the relationship is specified between the examinee's performance and the latent trait which is the subject of measurement according to a specified mathematical indicator. This theory depends on a number of models which are called the models of the latent traits through which connection is done between the performance on the item and the examinee's ability (Hambleton & Swaminathan, 1985).

The Two-Parameter Logistic Model

It is one of the models of the modern theory in measurement which supposes that both parameters of difficulty and distinction are changing and that the estimation for all the items equals zero (Allam, 2001).

Item Difficulty Parameter According to the Traditional Theory:

It is the proportion of the students who answered the item correctly with the students who attempted to answer this paragraph (Crocker & Algina, 1986).

Difficulty Parameter According to the Modern Theory:

It is the estimation of the ability of the observer for the probability of the correct answer (0.5) when the point of the curve crossing of the characteristics of the ability with the x-axis equals nearly zero (that is the parameter of guessing here equals zero).

Item Discrimination Parameter According to the Traditional Theory:

That is the ability of the item to discriminate between the high category and the low category. It is calculated for the two-scale items through finding the Biserial Correlation Coefficient between the grade of the item and the total grade of the test for each examinee (Crocker & Algina, 1986).

Item Discrimination parameter According to the Modern Theory:

It is the inclination curve of the characteristics of the item which occurs in it a change in the direction of the curve which corresponds to the difficulty on the ability connector and in it there is a probability of the correct answer for the item (i) equals 0.5.

Psychometric Characteristics:

It is meant by them the coefficients of validity and reliability for the mathematics achievement test which was prepared in this study.

Related Studies

2016-2022?

The study of Abu-Fodeh (2016) aimed to uncover the compatibility between the psychometric characteristics of the test and its items according to the traditional theory in measurement and the Two-Parameter Logistic Model in the compatibility of the items of the criterion-referenced test. In order to achieve the aims of the study, a criterion-referenced test in mathematics was built in the analytical geometry unit for the students of the tenth basic grade. The test in its final form consisted of (30) items of the multiple-choice test kind the validity and reliability of which were verified. The test was applied on a sample consisting of (140) students who were randomly chosen from the schools of the Jarash Directorate of Education. The results showed the conformity of (29) items with the traditional theory of measurement and the conformity of (28) items with the Two-parameter Logistic Model. (28) items were conformity with both theories. The value of the validity coefficient according to the Two-Parameter Logistic Model was (0.925), while the value of the validity coefficient according to the traditional theory in measurement was (0.936). The results indicated that there was a statistically significant difference between the two coefficients of validity in favour of the traditional theory in measurement. The value of the experimental reliability coefficient according to the Two-Parameter Logistic Model was (0.9549). The results indicated the existence of a statistically significant difference in the estimation of the two coefficients of reliability and in favour of the Two-Parameter Logistic Model.

Whereas the study of Abu-Jarad (2014) aimed at comparing between the model of the scale of estimation emanating from the Rasch Model and the traditional theory in measurement in terms of the precision of predicting the state of anger from the trait of anger among university students through their estimations on the measure of trait and a state of anger. In order to achieve the aim of the study, , the items of the two measures of trait and a state of anger were put on a criterion according to the model of the estimation scale emanating from the Rasch model and this was through applying the two measures on a sample consisting of (125) male and female students from Al-Quds Open University. For the purpose of comparison, the two measures of trait and a state of anger were applied after putting their items on a criterion on a sample consisting of (80) male and female

students from Al-Quds Open University, (45) male students and (35) female students were chosen from outside the graduation sample. The results of the study indicated that there is a correlation which is statistically significant between trait and a state of anger in both styles and that the precision of prediction by using the model of the estimation scale as one of the models of the modern theory is higher than that in the traditional theory.

The study of Hijazi and Al-Khateeb (2014) aimed to uncover the compatibility between the classical theory and the Two-Parameter Logistic Model in the compatibility of the items of a criterion-referenced test in theoretical subject of the rulings of recital of the holy Quran and reading it with intonation. In order to achieve this, a criterion-referenced test was built in the theoretical subject for the rulings of the recital of the holy Quran and reading it with intonation and it consisted of (41) items of the multiple-choice test kind of four alternatives. It was applied on a sample consisting of (404) male and female students who were chosen by the random stratified method from (16) Quranic centres in Jordanian capital 'Amman for the study year 2011/2012. The results of the study indicated the conformity of (4) items with the classical theory, (39) items were conformity with the Two-Parameter Logistic Model, and (39) items were compatible with each of the classical theory and the Two-Parameter Logistic Model. The value of the reliability coefficient of the test according to the classical theory was (0.927) while according to the Two-Parameter Logistic Model it was (0.943). As for the validity coefficient, its value according to the classical theory was (0.73) while according to the modern theory it was (0.65).

The study of Onn (2013) aimed at comparing between the traditional theory and the modern theory in measurement in terms of the number of the selected items and the reliability coefficient. In order to achieve the aims of the study, a test in the Physics subject was prepared consisting of (50) items of the multiple-choice test kind. It was applied on a sample consisting of (69) male and female students of the students of the schools in Nigeria. The items were analysed by using the SPSS program for analysing the items according to the traditional theory in measurement and using the X-Calibreprograme for analysing the items according to the Two-Parameter Logistic Model in the Item Response Theory. The results of the analysis showed the conformity of (29) items with the traditional theory in measurement and the conformity of (38) items with the Two-Parameter Logistic Model. The results indicated the lowering of the reliability coefficient of the test in both theories, whereby the value of the reliability coefficient in the traditional theory of measurement was (0.49), whereas the value of the reliability coefficient according to the Two-Parameter Logistic model was (0.67).

As for the study of Hussein (2011) it aimed at knowing the psychometric characteristics for the questionnaire of managing time among a sample of university students in Egypt and Saudi Arabia according to the traditional and modern theories of measurement. The Egyptian sample consisted of (466) male and female students of whom (107) male students and (359) female students in the Faculty of Arts at Al-Manoufiyyeh University while the Saudi Arabian sample consisted of (553) students who were distributed to (167) male students and (386) female students in the King Khalid College. The results showed a similarity in the psychometric characteristics of measurement between the two samples in the framework of the two theories. The results also revealed a similarity in the psychometric characteristics of the questionnaire derived from the traditional theory of measurement and the Item Response Theory.

In a study which was conducted by Hernandez (2009), it aimed at comparing between the discriminating the items and their difficulty in order to test the readiness of the mental speed by using the classical theory and the Item Response Theory. The test of the readiness of the mental speed was applied in its two parts: the verbal and the non-verbal and which consisted of (40) items on a sample consisting of (229) male and female students in the faculties of the Manila city, then analysing the data according to the classical theory and the Two-Parmeter Logistic Model. The results

revealed that there were no statistically significant differences between the two theories, and this was concerning the two mediums of difficulty, and also the two mediums of discrimination for the verbal part and also the non-verbal part of the test of the readiness of the mental speed.

Methodology of the Study

The researcher used the descriptive analytical methodology as a methodology for the study, and this is because it is suitable to the nature of the study, whereby in this methodology all the data and conducting the statistical analysis for figuring out the required results are done.

Population of the Study:

The population of the study consisted of the students of the eighth basic grade in the Nablus Governorate in the study year 2021/2022. whose number is (5252) male and female students distributed among (100) governmental schools of whom (2563) were male students and (2689) were female students according to the statistics of the Palestinian Ministry of Education.

Sample of the Study:

The sample of the present study consisted of (580) male and female students of the eighth basic grade of whom (291) were male students and (328) female students distributed on (20) schools with (10) schools for males and (10) schools for the females. The schools were chosen in purposive manner.

The researcher also chose four specialists in the methods of teaching mathematics and four teachers of the mathematics subject for the eighth basic grade in order to act as referees for the test items and to specifying the cutting degree.

Instrument of the Study:

In order to achieve the aims of the study, the researcher analysed the content of the mathematics book for the eighth basic grade for the first study semester and to specify the knowledge aims of the study units which are the components of the book. The aims were classified according to the abilities which measure them, and they are the conceptual ability, the procedural ability and the ability to solve problems. Also, a table of the specifications of the test which was intended to be built was done whereby the test in its final form consists of (50) items.

After this the researcher wrote (60) items of the multiple-choice test kind with four alternatives for each item, one of them represents the correct answer, whereby the grade of (1) was given of the response was correct and the grade (zero) was given if the response was wrong. In formulating the test items, the scientific bases in building the test and their comprehensiveness of the conceptual and procedural abilities and problem solving were taken into consideration.

The content analysis, the list of aims, the table of specifications and the test items were presented to a group of referees in order to give their opinions concerning them and to judge the items in terms of their formulation, and the extent to which each item is connected to the aim which it measures. The observations of the referees were taken into consideration. The items which were in need for modification were modified, and (10) items were omitted so that the final number of the items were (50) items distributed on the three abilities: the conceptual, the procedural and problem solving and in harmony with the table of specifications of the test. The referees specified the cutting degree of the test and they are (25), and this was by the Nadelski method which is an old method in estimating the cutting degree of the test the items of which are of the multiple-choice test kind, and it depends on the estimation of the referees for these items whereby the referee specifies the alternatives of the response which the examinee having less than an acceptable level of sufficiency because they are not correct can exclude. The cutting degree specified for each item is the inverted number of the remaining alternatives, and which represent the probability of the success of the individual having less than an acceptable level of sufficiency in answering the item in a correct way if he chooses any of the remaining

alternatives in a random way (Allam, 1991, p. 87).

After finishing preparing the test and preparing it in its final way, it was applied on the individuals of the sample of the study, each in his class and his school, and this was by the help of the teachers of the mathematics subject in those schools. The researcher corrected the test in a manual way.

Statistical Treatments:

The researcher entered the data in the computer and he used the Statistical Packages for Social Sciences (SPSS) program and the MULTILOG.7 program in order to do the statistical analyses.

Results of the Study and Their Discussion

Q1: What is the extent of the conformity of the mathematics achievement test with the hypotheses of the classical theory in measurement and evaluation?

To answer the first question, the researcher analysed the data by using the Statistical Packages for Social Sciences (SPSS) program, and this was in order to calculate the coefficients of difficulty and discrimination for the test items. The coefficient of difficulty was calculated for each item by finding the proportion of the students who responded on the item correctly among the examinees who attempted to answer this item. As for the coefficient of discrimination for each item, it was calculated through finding the dual biserial (rbis) correlation coefficient between the grade of the item and the total grade on the test. Table (1) indicates the Difficulty and discrimination coefficients for each item of the test items.

Table (1): *The Difficulty and Discrimination Coefficients According to the Classical Theory for the Test Items*

Item	Difficulty Coefficient	Discrimination Coefficient	Item	Difficulty Coefficient	Discrimination Coefficient
1	0.640	0.388	**26**	0.355	0.325
2	0.507	0.366	**27**	0.513	0.557
3	0.553	0.329	**28**	0.565	0.417
4	0.588	0.348	**29**	0.465	0.381
5	0.672	0.353	**30**	0.396	0.435
6	0.610	0.379	**31**	0.573	0.412
7	0.517	0.256	**32**	0.610	0.394
8	0.606	0.447	**33**	0.404	0.458
9	0.740	0.402	**34**	0.642	0.371
10	0.493	0.406	**35**	0.559	0.447
11	0.626	0.441	**36**	0.421	0.411
12	0.543	0.436	**37**	0.616	0.470
13	0.636	0.468	**38**	0.348	0.222
14	0.571	0.411	**39**	0.467	0.352
15	0.527	0.460	**40**	0.636	0.277
16	0.509	0.432	**41**	0.624	0.324
17	0.316	0.527	**42**	0.664	0.293
18	0.527	0.483	**43**	0.475	0.319
19	0.509	0.273	**44**	0.469	0.330
20	0.654	0.458	**45**	0.748	0.388

21	0.563	0.388	**46**	0.543	0.443
22	0.569	0.320	**47**	0.636	0.441
23	0.596	0.456	**48**	0.600	0.391
24	0.495	0.325	**49**	0.638	0.465
25	0.559	0.485	**50**	0.549	0.339
The arithmetic means of Difficulty Coefficient			**0.553**		
The arithmetic means of Discrimination Coefficient			**0.394**		

It is noticed from the results of table (1) that the values of the difficulty coefficients for the items of the mathematics test ranged between (0.316 - 0.748) and with an arithmetic mean of (0.553), and they are considered suitable difficulty coefficients, whereby the Doran (1980) indicator was adopted in judging the difficulty coefficients for the items and in which he pointed out that the item in which the level of its difficulty lies between (0.22 - 0.80) is suitable and acceptable.

It is also noticed from the results of table (1) that the values of the discrimination coefficients for the items ranged between (0.222 - 0.2557) and with an arithmetic mean of (0.394). When taking the discrimination criterion according to Allen & Yen (1979) which is pointed out to in Odeh (2010), that any item enjoys a discriminating ability of (0.25) and more. It enjoys an acceptable degree of discriminating ability. On this basis, item number (38) in which its discriminating coefficient is (0.222) is nonconformity with the criteria of the classical theory. Thus, it was omitted from the test so that the number of the final items which are conformity with the classical theory is (49) items, that is a proportion of (98%) of the items which were conformity with the classical theory in measurement. This result agreed with the study of Abu-Fodeh (2016) and the study of Hijazi and Al-Khateeb (2014) in the conformity of the items with the classical theory. The proportion of the items which were conformity with the traditional theory in these two studies was (96.67%, 97.56%). However, this study did not agree with some studies such as the study of Salem (2011), the study of Onn (2013) and the study of Jamhawi (2000) whereby the proportion of the items which were conformity with the traditional theory in these studies was successively (72%, 58%, 84%). The researcher interprets the cause of disagreement in that there was a difference in the number of the items used in these studies and the difference in the volumes of the samples whereby the difficulty and discrimination coefficients differ with the difference of the volume of the sample. Perhaps the reason may be due to the aspects of deficiency in the classical theory.

Q2: What is the extent of the conformity of the items of the mathematics test with the modern theory in measurement and evaluation (represented in the Two-Parameter Logistic Model)?

To answer the second question, the researcher verified that the data achieve the hypotheses of the modern theory. The first hypothesis which is the Unidimensionality hypothesis was verified by conducting the Factor Analysis of the first degree for the items of the mathematics test and the Promax oblique rotation, whereby the results of the Factor Analysis indicated the existence of (16) factors of the eigen value for each of them is more than one, and it interpreted the sum of (54.132%) of the total variance. Table (2) indicates the Eigen values, the proportion of the Explained Variance and the cumulative proportion of the Explained Variance.

Table (2): *Results of the Factor Analysis of the First Degree for the Items of the Mathematics Test*

component	Eigen Value	Proportion of Explained Variance	Variance Cumulative Explained Variance
1	8.045	16.090	16.090
2	1.865	3.730	19.820
3	1.582	3.164	22.984
4	1.421	2.841	25.825
5	1.368	2.736	28.561
6	1.338	2.676	31.238
7	1.282	2.563	33.801
8	1.268	2.537	36.338
9	1.259	2.519	38.857
10	1.198	2.396	41.253
11	1.173	2.345	43.598
12	1.105	2.211	45.809
13	1.068	2.136	47.945
14	1.059	2.117	50.063
15	1.029	2.058	52.121
16	1.005	2.011	54.132

It is clear from the results of Table (2) that the first hypothesis of the modern theory in measurement and evaluation which is the unidimensionality was verified. It can be inferred from the value of the eigen value of the first factor which is (8.045) which interprets a proportion of (16.090%) of the total variance, whereby the proportion of the eigen value for the first factor to the eigen value of the second factor is 4.13 which is a big proportion and it is considered an indicator of unidimensionality. Hambleton & Swaminathan (1985) pointed out that if the proportion of the Eigen value of the first factor to the Eigen value of the second factor is less than 2, then this is an indicator of unidimensionality. By observing the graphic representation (Score plot) for the factors with their eigen value in Figure (1), it is indicated that there is a change in the inclination of the curve at the second factor and the inclination remains inverted for the rest of the factors which gives greater weight also to the existence of a prevailing factor from which the unidimensionality can be inferred for the purposes of estimating the parameters of items and persons.

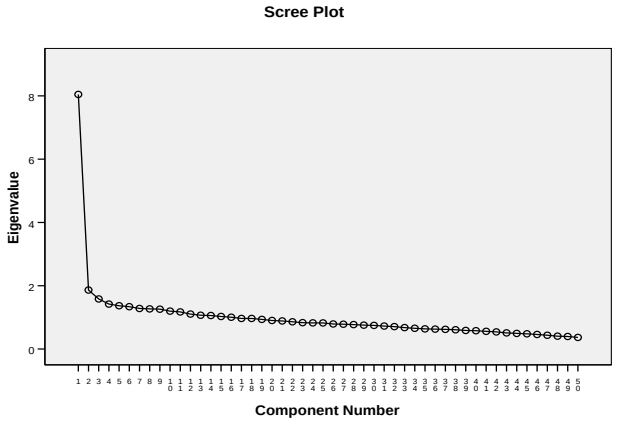

Figure (1): *Graphic representation of the values of the Eigen value for the first-degree factors for the items of the mathematics test*

As for verifying the second hypothesis of the modern theory which is Local Independence, Warm (1978) views that the hypothesis of the unidimensionality includes presupposing Local Independence.

The third hypothesis was also verified, and it is the hypothesis of Speed in Performance, through making sure that all the students answered the items of the test within the specified time. This is considered an indicator of achieving the hypothesis of being liberated from speed in performance.

After making sure of the achievement of the hypotheses of the modern theory, the researcher used the MULTILOG.7 program to analyse the responses of the sample of the study with the aim of uncovering the individuals who are not compatible with the Two-Parameter Logistic Model by means of the statistic Chi Square. The results of the analysis produce the nonconformity of three individuals with the program. So, the researcher repeated the analysis of the test of the conformity of the mathematics test items with the Two-Parameter

Logistic Model by means of the statistic Chi-Square and calculating each of the two parameters of difficulty and discrimination for each item. Table (3) shows the two parameters of difficulty ad discrimination and the conformity test for every item of the mathematics test items.

Table (3): *The Two Parameters of Difficult and Discrimination and testing the conformity of every item of the Mathematics Test Items*

Item	Item Difficulty Parameter	Discrimination Parameter	χ^2	Value Significance Level	Item	Item Difficulty Parameter	Discrimination Parameter	χ^2	Value Significance Level
1	0.98-	0.53	0.0000085	0.998	26	1.20	0.44	0.0000143	0.997
2	0.31-	0.49	0.0000068	0.998	27	0.09-	1.28	0.0001083	0.992
3	0.26-	0.69	0.0000117	0.997	28	0.5-	1.013	0.0000004	1
4	0.45-	0.83	0.0000093	0.998	29	0.07	0.96	0.0000040	0.998
5	0.94-	0.67	0.0000181	0.997	30	0.66	1.02	0	1
6	0.66-	1.02	0.0000060	0.998	31	0.24-	0.96	0.0000010	0.999
7	0.13-	0.99	0.0000145	0.997	32	0.83-	0.73	0.0000042	0.998
8	0.58-	1.31	0.0000010	0.999	33	0.54	1	0.0000004	1
9	0.99-	1.18	0.0000297	0.996	34	1.11-	0.5	0.0000111	0.997
10	0.32-	0.77	0.0000002	1	35	0.43	1.03	0.0000020	0.999
11	0.82-	0.78	0	1	36	0.48	0.9	0.0000015	0.999
12	0.47-	0.88	0.0000006	0.999	37	0.68	0.62	0.0000042	0.998
13	0.49-	1.34	0.0000035	0.999	38	1.41	0.38	0.0000143	0.997
14	0.43-	0.75	0.0000004	1	39	0.43	0.86	0.0000068	0.998
15	0.28-	1.44	0.0000058	0.998	40	1.11-	0.66	0.0000190	0.997
16	0.15	0.86	0.0000004	1	41	0.59-	0.78	0.0000154	0.997
17	0.86	1	0.0000046	0.998	42	0.67-	1	0.0000197	0.996
18	0.27-	1.4	0.0000145	0.997	43	0.27-	0.82	0.0000103	0.997
19	0.25-	0.79	0.0000130	0.997	44	0.56	0.75	0.0000116	0.997
20	0.67-	1.34	0.0000004	0.999	45	0.88-	1.01	0.0000356	0.995
21	0.20-	0.88	0.0000033	0.999	46	0.18-	0.64	0.0000004	1
22	0.24-	0.77	0.0000118	0.997	47	0.34-	1.07	0	1
23	0.41-	1.14	0.0000020	0.999	48	0.60-	1.01	0.0000050	0.998
24	0.05-	1.03	0.0000102	0.997	49	0.60-	2.17	0.0000011	0.999
25	0.19-	1.11	0.0000147	0.997	50	0.26-	1.64	0.0000103	0.997
Arithmetic Mean of the Estimations of the Difficulty Parameter					**- 0.227**				
Arithmetic Mean of the Estimations of the Discrimination Parameter					**0.925**				

It is noticed from Table (3) the values of the estimations of the difficulty for the items of the mathematics test according to the Two-Parameter Logistic Model in the modern theory ranged between (-1.21 - 1.54) Logit, and with an arithmetic mean of (0.227-) Logit. The study of Jamhawi (2000) pointed out that the items in which the values of their difficulty coefficients are between (-1.5 - 1.5) Logit are considered within the medium extent of the difficulty coefficients. Based on this, all the items of the mathematics test are considered as medium difficulty items, whereas the estimations of the discrimination parameter ranged between (1.57-0.43) and with an arithmetic mean of (0.925).

It is also noticed from Table (3) that all the items and according to the Chi- Square test for the good compatibility were compatible with the Two-Parameter Logistic Model.

It is noticed from the results of the second question that there is a difference in the proportion of the items which are conformity with the Two-Parameter Logistic Model with the results of some previous studies. The proportion of the items which were conformity in this study was 100%, whereas the proportion of the items which were conformity with the Two-Parameter Logistic Model in the study of Hijazi and Al-Khateeb (2014) was (95%), and with the study of Ab-Fodeh, it was with a percentage of (93.33%), with the study of Onn

(2013) the percentage was (89.7%), and with the study of Jimelo and Silvestre (2009) the percentage was (55%). The researcher interprets this difference in that it may be due to the difference in the software programs used in the analysis, the difference in the content of the tests, and the number of the items of each test and the difference in the samples of each study.

Q3: What are the psychometric characteristics (validity and reliability) of the mathematics test according to the classical theory and the modern theory (represented in the Two-Parameter Logistic Model)?

To answer the third question, the researcher found the psychometric characteristics of the items of the mathematics test by using the Statistical Packages for Social Sciences program (SPSS according to the classical theory, while the software program MULTILOG.7 was used in calculating the psychometric characteristics of the items of the mathematics test according to the Two-Parameter Logistic Model.

The reliability coefficient was calculated according to the classical theory by using Cronbach Alpha for internal consistency, while the coefficient of the experimental reliability (the empirical) was estimated according to the Two-Parameter Logistic Model in the modern theory in measurement and evaluation. Table (4) indicates the values of the reliability coefficients according to the classical theory and according to the Two-Parameter Logistic Model.

Table 4: *Values of the reliability coefficients according to the classical theory and the modern theory (the Two-Parameter Logistic Model)*

Reliability coefficients Cronbach Alpha according to the classical theory			
Before omitting the items		After Omitting the item	
Number of items	Value of Reliability coefficients	Number of items	Value of Reliability coefficients
50	0.887	49	0.921
Experimental reliability coefficients according to the modern theory (the Two-Parameter Logistic Model			
Number of items		Value of Reliability coefficients	
50		0.953	

It is noticed from Table (4) the values of the reliability coefficients for the test according to the Two-Parameter Logistic Model (the modern theory) were higher than those according to the traditional theory. This result agreed with the study of Hijazi and Al-Khateeb (20140 and with the study of Abu-Fodeh (2016), while the result differed with the study of Onn (2013) which showed the lowering of the reliability coefficients in both theories. The researcher interprets the result by that the number of items which were conformity with the traditional theory in the study of Onn was (29) items according to the traditional theory and (38) items according to the Two-Parameter Logistic Model out of (50) items, while all the items of the mathematics test which was prepared in the present study the number of which is (50) items were conformity with the Two-Parameter Logistic Model. This affirms that whenever the number of items increase, this leads to a rise in the reliability coefficients.

As for the validity coefficient, validity was used with a criterion-significance whereby it was inferred by means of calculating the correlation coefficients between the performance of the individuals of the sample of the study in the test and their school grades in

the mathematics subject for the first semester which obtained from the rosters of the school grades. Table (5) indicates the validity coefficients with the criterion-significance of the test.

Table 5: *Values of the validity coefficients according to the classical and modern theory*

Validity coefficients according to the classical theory			
Before omitting the items		After Omitting the item	
Number of items	Value of validity coefficients	Number of items	Value of validity coefficients
50	0.903	49	0.945
Validity coefficient according to the modern theory (the Two-Parameter Logistic Model)			
Number of items		Value of validity coefficients	
50		0.895	

It is noticed from Table (5) that the values of the criterion-validity coefficients for the test raised after omitting the nonconformity item according to the classical theory, and that the validity coefficients according to the traditional theory are higher than the validity coefficients according to the Two-Parameter Logistic Model. The researcher interprets the rise of the validity coefficients after omitting the nonconformity item in the classical theory is due to lowering the standard error in the estimation.

Q4: What is the extent of the compatibility between the classical theory and the modern theory (the Two-Parameter Logistic Model) in choosing the items of the mathematics test?

To answer the fourth question, the researcher did a comparison between the statistical indicators which were figured out according to the classical theory and the Two-Parameter Logistic Model. Table (6) indicates the number and the proportion of the items according to their conformity with the classical theory and the modern theory represented in the Two-Parameter Logistic Model.

Table (6): *Number and proportion of the items according to their conformity with the classical theory and the modern theory*

State of the item in terms of conformity	**Number of items**	**Their proportion**
conformity items according to the two theories	49	98%
conformity items according to the two theories items in the two theories	0	0
nonconformity e items according to the classical theory and conformity According to the modern theory	1	2%
nonconformity items according to the modern theory and conformity according to the classical theory	0	0
Total Sum	**50**	**100%**

It is notice from the results of Table (6) that there is an agreement between the statistical indicators of the classical theory and the modern theory (represented in the Two-Parameter Logistic Model) in the conformity of (49) items with the two theories, while there is no agreement in one item whereby this item was compatible with the modern theory but it was incompatible with this classical theory. This was item (18). This is considered an indicator that the analysis of the items according the Two-Parameter Logistic Model in the modern theory is better than their analysis according to the classical theory (although the analysis of the items in this study according to the modern theory and according to the Two-Parameter Logistic Model is in approximation with their analysis according to the classical theory). This result agrees with the results of the study of Stege (2003), but it differed from the study of Abu-Fodeh (2016), the study of Hijazi and Al-Khateeb (2017) and the study of Salem which indicated that the items which were incompatible with the classical theory were nonconformity with the modern theory too.

Recommendations

1. Conducting studies comparing between the classical theory and the modern theory on models of the multiple graduation tests.

2. Conducting comparative studies by using the Two-Parameter Logistic Model.

3. Conducting comparative studies by using the Uni-Parameter Logistic Model (the Rasch Model).

4. Using statistical software programs other than those which were used in this study for conducting the analysis according to the modern theory.

References

[1] Allam, S.(1991). A comparative study of some methods for specifying the levels of performance in a criterion-referenced test. The Egyptian Journal for Psychological Studies, issued by the Egyptian Society for Psychological Studies, 1, pp.77-96.

[2] Allam, S. (2001). The Criterion-Referenced Diagnostic Tests in the Educational and Psychological Domains, second edition, Cairo: Daar Al-Fikr Al-'Arabi (Arab Thought House).

[3] Odeh, A. (2010). Measurement and Evaluation in the Teaching Process. Irbid: Daar Al-Amal (House of Hope).

[4] -Abu-Fodeh, B.(2016). The compatibility between the traditional theory in measurement and the item response theory in the compatibility of the items of a criterion-referenced test in the unit of analytical geometry. Arab Studies in Education and Psychology, (73), 189-215.

[5] Abu-Hisham, A.(2006). A comparative study between the traditional theory and the model of Rasch in testing the items of the measure of introductions of study among the university students, Al-Zaqazeeq University. Journal of the Faculty of Education, (5). ???

[6] -Abu-Jarad, Hamdi Younes (2014). The precision of the predicting the state of anger from the trait of anger-a comparative psychometric study between the traditional theory and the modern theory in measurement. Journal of the Islamic University for Educational and Psychological Studies, 33(2), pp.101-129.

[7] Aby-Judeh, Shatha Ibrahim (2018). The efficacy of using the model of Rasch in building a criterion-referenced achievement test in mathematics for the students of the ninth basic grade in Jordan. A published Master's thesis. The Arab University of 'Amman, Jordan.

[8] Adedoyin, O. (2010). Investigating the Invariance of Peron Parameter Estimates Based on Classical Test and Item Response Theories. International Journal of Science, 2 (2); 107-113

[9] Allen. M. j., and Yen, W.M. (1979). Introduction to measurement theory. Montery, California: Brooks/cole publishing Co.

[10] Doran, R.L. (1980). Basic Measurement and Evolution of Science Education. Washington, DC: National Science Teacher Association.

[11] Embreston, S. E & Reise, S.P. (2000). Item response theory for psychologists. New Jersey: Lawrence Erlbaum Associates, Inc.

[12] Hambleton, K., Swaminathan, H. (1985). Item Response Theory: Principles and applications. Boston: Kluwer Nijhoff publishing.

[13] Hambleton, R.K. Jonse, R, W (1993). Comparison of Classical test theory and item response theory and their application to test development. Educational Measurement, Issues and practice, 12 (3), 38-47.

[14] Hambleton, R.K. Swaminthan, H. (1985). Item Response Theory: principles and Application. Kluwer. Nijhoff Publishing, Boston.

[15] Hernandez, R. (2209). Comparison of the Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test using CTT and IRT Methods. The International Journal of Education Employment and Psychological Assessment, 1 (1), 12-18.

[16] Hijazi, T. and Al-Khateeb, A. (2014). The compatibility between the classical theory and the Two-Parameter Logistic Model in the fitting of the items of the criterion-referenced test in the rulings of the recital of the holy Quran and reading it with intonation. Journal of An-Najah University for Researches-The Humanities, 28(10), ????

[17] Hussein, Mahmoud Habashi (2011). The psychometric characteristics of the questionnaire of time management among a sample of university students in Egypt and Saudi Arabic: An evaluative study of the traditional theory in management and the item response theory. The Educational Journal-Kuwait, 25(99), 353-410.

[18] Jimelo, L. & Silvestre, T. (2009). Item Response Theory and Classical Test Theory: An Empirical Comparison of Item Person statistics in A Biological Science Test. The International Journal of Education and Psychological Assessment, 1 (1), 19-31.

[19] Kathem, 'Ali Mahdi (2001). Measurement and Evaluation in Learning and Teaching, first edition. Sultan Qabous University, Masqat: Al-Kindi House for Publishing and Distribution.

[20] Klein, M. Muijtjens, A. Habets, L. Manogue, M. Van der Vleuten, C & Van der Velden, U. (2009). Intuition Who will pass dental OSCE? Comparison of the Angoffand borderline regessionstandard setting methods. European Journal of Dental Educational, 13 (1). 162-171.

[21] Lord, F.M. (1980). Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

[22] Onn, D. (2013). Classical test theory versus item response theory: An evaluation of the comparability of item analysis results. Joint Admissions and Matriculation Board, 1 – 23.

[23] Salem, H. (2011). The extent of compatibility between the Two-Parameter Logistic Model and the traditional theory in building an achievement test in the subject of General Sciences for the sixth basic grade. An unpublished Master's thesis, Al-Yarmouk University, Jordan.

[24] Stage, C. (2003). Classical Test Theory or Item Response Theory: The Swedish Experience. Umea University, pp:1-30.

[25] Tiratira, N. (2009). Cutoff Scores- The Basic Angoff Method and the Item Response Theory Method. The International Journal of Educational and Psychological Assessment, 1 (1), 39-47.

[26] Warm, T.A. (1978). A Primer of item Response Theory. Oklahoma: U.S. Coast Guard Institute 73/69.

[27] Wang, Y., Zhao, L., Shen, S., & Chen, W. (2021). Constructing a Teaching Presence Measurement Framework Based on the Community of Inquiry Theory. Frontiers in psychology, 12, 694386. https://doi.org/10.3389/fpsyg.2021.694386 .

[28] Jamalzadeh, M., Lotfi, A.R. & Rostami, M. (2021). Assessing the validity of an IAU General English Achievement Test through hybridizing differential item functioning and differential distractor functioning. Lang Test Asia 11, 8 (2021). https://doi.org/10.1186/s40468-021-00124-7

[29] Kerry Shephard, Bill Warburton, Pat Maier & Adam Warren (2006) Development and evaluation of computer-assisted assessment in higher education in relation to BS7988, Assessment & Evaluation in Higher Education, 31:5, 583-595, DOI: 10.1080/02602930600679621.

[30] Hambleton RK, & Pitoniak MJ. Setting performance standards. Educational Measurement. 2006;4:433–470