Research Article

Development and Validation of the Mathematics Test for Tenth Grade Jordanian Students, According to the Partial Credit Model

Anwar Bani Hani[1], Rohaya Talib[2] Mutasem Zrekat[3], Mohd Alfouzii Nasir[1,4] Areej Al_Ahmad[5], Nariman Wedian[6]

[1]Faculty of Social Science and Humanities, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[2]Senior Lecturer, School of Education, Universiti Teknologi, Malaysia, Johor Bahru, Malaysia.
[3]Faculty of Engineering, Universiti Teknologi Malaysia, Skudai, 81310, Johor, Malaysia
[5]Faculty of Social Science and Humanities, the University of Jordan, Amman, Jordan
[6]Faculty of Social Science and Humanities, Jadara University, Jordan

Corresponding Author: Anwar Bani Hani, [1]Faculty of Social Science and Humanities, Universiti Teknologi Malaysia, Malaysia.
Email: anwar.hani91@yahoo.com **ID: https://orcid.org/0000-0002-7165-3477**

Article History: Received: 10 November 2020; Revised 12 January 2021; Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: The study aimed at developing and validating the mathematics test for 10th –grade students according to the Rasch partial credit model (PCM) by using the descriptive approach as it is appropriate for the study aims. To achieve the study's objective, what constructed the essay type test consisted of 25 items based on the (IRT) according to the Rasch PCM. what conducted the first administration of the test to verify the validity and reliability of the test. To verify the "face validity" of the test's objectives, they were presented to a group of 12 arbitrators who work as teachers and educational supervisors. They found that the contents are representative of the level of the goal, which is pursuing in theory. The empirical reliability was calculated for the test, where the value of person reliability reached 0.91. Moreover, the items reached 0.93. The study population consisted of all 10th-grade students at the schools belonging to the "Directorate of Education of Irbid District," whose numbers were 7365, represented by 3612 male students, and 3753 female students. According to the class regarding their sex (gender), a sample according to a cluster as the test unit was the class section. The sample size of the study was 250 male and female students. According to the PCM, this study's findings have brought several issues concerning the mathematics subject achievement by verifying the tests and reliability and accomplishing the IRT's suppositions.

Keywords: Mathematics Test (MT), Item Response Theory (IRT), Partial Credit Model (PCM), Tenth Grade Jordanian Students(TGJS)

1. Introduction

Construction and validation of tests, especially academic achievement measures, contain complicated steps, procedures, the interrelationship of various ideas and latent variables. Subsequently, confirmed what must follow guidelines to develop a firmly identified test with the expected outcomes. The two most essential steps in test development as spelled out by (Haladyna & Downing, 2011) are; (i) Item development, which includes content definition, preparation of test specifications, preparation of the item pool, content validation/experts judgment, pilot testing of the items, data analysis, and revision of test items. (ii) Item validation through item analysis. All these explained processes are closely linked with others. Additionally, these processes are carefully accomplished to ensure the instrument's validity and reliability developed and used to estimate items and a person's ability. Validity is concerns about how assessment systems are built. Whether the assessment tool (Test) is standardized or locally-designed, the aim is to use an instrument that produces a true estimate of the examinee's ability that could support valid inferences. The purpose of assessing student's learning includes licensing, certification, diagnosis, and placement.

The field of educational and psychological assessment and evaluation has received increased research attention by psychologists and educators. This field's primary objective was to reveal individual differences of all kinds, whether inter-individual differences between groups or intra-individual differences. Measuring methods and instruments have been varied to achieve this goal. The assessment quality depends on the quality

of performance and the quality of the measurement process in Classical Test Theory (CTT). These efforts have led to the transition from the CTT used in the design of the tests, which have been used for a long time in the educational and psychological evaluation, to the modern approach, IRT, or the Latent Trait Theory (LTT) (Newton & Shaw, 2014).

CTT being a traditional theory, still attract the measurement community in test development and analysis due to its theoretical and practical simplicity. The continuous application of CTT in item analysis is because of its "weak assumptions," which can easily be met by test data (De Champlain, 2010; Hambleton & Jones, 1993). Although, as a result of its continuous utilization, researchers have questioned it's in the present-day measurement community (Zaman, Kashmiri, Mubarak, & Ali, 2008). The PCM could be a one-dimensional model for the analysis of responses recorded in two or more ordered classes, as well as Samejima's graded response model (GRM) (Samejima, 1969). The PCM differs from the GRM. However, therein it belongs to the Rasch family of models, then share the identifying characteristics of that family: separable person and item parameters, adequate statistics, hence, integrated additively. These characteristics modify "specifically objective" comparisons of persons and items (Rasch, 1977) and permit every set of model parameters to be conditioned out of others' estimation procedures.

In education, assessment is an important matter to identify educational success. The results of the educational evaluation have a significant function that will be useful in other academic processes. Two primary tasks measure the students' achievements and motivate and direct students' learning (Frisbie & Becker, 1991). One of the significant functions in mathematics education is identifying how far the students have possessed specific subjects such as mathematics. Besides, to determine the students' knowledge or understanding, the assessment results also provide certain concepts, such as mathematics concepts that the students have not mastered. The teachers or school might improve the learning process through the assessment results, and students can change their strategy for studying. The educational assessment results in Jordan, especially in mathematics, have not satisfied many parties over the years. It can be seen in the situation that might be found both using students' average score in the Mathematics in some research and studies which investigated in this subject and the results provided by the international studies. Based on the results of these studies, the students' achievements have not been satisfying (Khasawneh, 2009; Rabab'h & Veloo, 2015).

The researchers and mathematics teachers noticed the students' fear of essay questions at the expense of the multiple-choice questions, as they prefer the multiple-choice questions much more of the essay/structural questions. Also, they have noticed the lack of studies that have addressed the mathematics curriculum in particular according to the PCM, which aims to determine the difficulty coefficient for each step while answering the items of polytomous responses, which is considered as a generalization of the Rasch model in the dichotomous responses item (Tuckman, 1993). As a consequence, this paper identified the problem in general in an attempt to select items from achievement test in the subject of mathematics, specifically for students in the 10th grade, because this stage is the transition from the primary stage to the secondary stage, also to demonstrate the importance and effectiveness of the PCM in achievement tests or other tests. The test has psychometric characteristics so that it can be applied and used in public and private schools. According to the PCM, this study was designed to provide the mathematics test for students in the 10th grade.

The primary purpose of this study was to develop and validate the mathematics test (MT) for 10th-grade Jordanian students, according to the PCM, and depend on the Rasch model IRT. The two models used to obtain valid and reliable test items relevant to measuring students' true ability from traditional and modern measurement perspectives. Besides, the analysis determined the appropriate items that satisfied specific criteria for item quality. In light of these and many concerns, this study was conducted to investigate the nature of the IRT item parameters for Jordan's mathematics test. Overall, IRT models can be divided into two categories: uni-dimensional and multidimensional. Uni-dimensional models require a single trait (ability) dimension θ. Multidimensional IRT models model response data hypothesized to emerge from multiple traits. However, because of the significantly increased complexity, most IRT research and applications utilize a uni-dimensional model. IRT models can also be categorized depending on the number of scored responses. The typical multiple-choice item is dichotomous; although there maybe four or five options, it is still scored only as correct/incorrect (right/wrong).

1.1 Number of IRT parameters

Dichotomous IRT models are described by the number of parameters they make use of (Thissen & Orlando, 2001). The 3PL is named so because it employs three item parameters. The two-parameter model (2PL) assumes that the data have no guessing but that items can vary in location bi and discrimination ai. The one-parameter model (1PL) assumes that guessing is a part of the ability. All items that fit the model have equivalent discriminations so that a single parameter bi only describes items.

1.2 The Rasch model

The Rasch model is often considered to be the 1PL IRT model. However, Rasch modeling proponents prefer to view it as a completely different approach to conceptualizing the relationship between data and theory (Andrich, 1989). Like other statistical modeling approaches, IRT emphasizes the importance of a model's fit to observed data (Steinberg, 2000). In contrast, the Rasch model emphasizes the priority of fundamental measurement requirements, with good data-model fit being an essential but secondary requirement to be met before a test or research instrument can be claimed to measure a trait (Andrich, 2004). There are some IRT models with polytomous responses: many different models of the IRT appeared (Nering & Ostini, 2011; Ostini & Nering, 2006). Each of these models had a specific purpose. These models were mentioned as follows, with some supported studies for them:

1.2.1 Normal Ogive Model (NOM):

The NOM was the first IRT model for measuring psychological and educational latent traits (Ferguson, 1942; Lawley, 1943; Mosier, 1940; Richardson, 1936). The NOM was refined later by (Lord & Novick, 1968). An item characteristic curve (ICC) is derived from the cumulative density function (CDF) of a normal distribution in the model. Besides, some studies that applied this model (Tran & Formann, 2009)

1.2.1 Partial Credit Model (PCM):

The PCM is an extension of the 1PLM Rasch model (Masters, 1982). The study of (Choi & Swartz, 2009) applied this model.

1.2.1.1 Generalized Partial Credit Model (GPCM):

The GPCM (Muraki, 1992) is a generalization of the PCM with a parameter for item discrimination added to the model. The study of (Chen, 2010) used this model.

1.2.2 Rating Scale Model (RSM):

There are two different approaches to the RSM (Andersen, 1997) proposed a response function, in which the values of the category scores are directly used as a part of the function. Another form of the RSM was proposed by (Andrich, 1978), which can be seen as a PCM modification. The recent studies that used this model were (Dehqan, Yadegari, Asgari, Scherer, & Dabirmoghadam, 2017; Gómez, Arias, Verdugo, & Navas, 2012).

1.2.3 Graded Response Model (GRM):

The GRM was introduced by (Samejima, 1969) to handle ordered polytomous categories such as letter grading, A, B, C, D, and F, also polytomous responses to attitudinal statements such as a Likert scale. The study of (LaHuis, Clark, & O'Brien, 2011) adopted this model.

1.2.4 Nominal Response Model (NRM):

The NRM, also called the Nominal Categories Model (NCM), was introduced by (Bock, 1972). Unlike the other polytomous IRT models introduced above, NRM's polytomous responses are unordered or at least not assumed to be ordered. Even though responses are often coded numerically (for example, 0,1, 2…, m), the responses' values do not represent some scores on items but just nominal indications for response categories. There are some applications of the NRM found in uses with multiple-choice items. As for models of polytomous responses, it is used when the response consists of many scores, and each score has a Difficulty Coefficient (DC) according to the modal used. One of these models is the (PCM) which identifies DC each step (k) while answering item (i) of polytomous responses, as well as determining the latent ability of the person and his performance. There is also the (GRM), each item has a Discrimination Index, and each section of the response has DC (Mislevy & Verhelst, 1990). Some recent studies that adopted this model were (Huggins-Manley & Algina, 2015).

2. Research Questions

The following research questions were raised to guide the study:

- Do the mathematics test data for tenth-grade students achieved the assumptions of Item Response Theory (IRT)?
- To what degree does the mathematics test's data conformity to the tenth-grade students with the Partial Credit Model (PCM)?
- What are the estimates of the values of the parameter of the items according to the Partial Credit Model?

- What are the estimates of the values of the person's ability depending on the model used?

3. **Materials and Method**

3.1 Research Design

The study aimed at the development and validation of the mathematics test for 10th-grade students according to Rasch (PCM) by using a survey design as it is appropriate for the study aims.

3.2. Participants

3.2.1Population of the Study

The study population consisted of all 10th-grade students at the schools belonging to the "Directorate of Education of Irbid District," whose numbers were 7365, distributed to 3612 male students and 3753 female students.

3.2.2 Sample and Sampling Techniques

The sample in this study was drawn using stratified random sampling technique that chosen according to their gender and then the cluster, as the unit of choice was the classroom division, where 14 schools were divided into 7 male schools and 7 female schools, where the study sample size reached 250 male and female students. The instrument is a mathematics test constructed for 10th-grade students and estimated the Difficulty Coefficient (DC) following Rasch (PCM). The test consisted of (25 items) of essay type, and each item has multiple answers as each item needs. The test items covered the whole mathematics subject. The 25 items of the essay type constructed based on the (IRT), according to the Rasch PCM, was administered on the sample of 10th-grade students, which are mainly under the control of their respective schools. Each Item had four answers following the steps of the achievement test. After receiving specific instruction for the test by teachers under the monitoring and evaluation unit's supervision, responsible for the regulation of primary education in the ministries, in the north area "Irbid district" in March 2021. After Coordinating with the school, were set a management date for a visit to set a date for applying the study tool. They informed that which used the information obtained for scientific research after applying the study tool in its final form on the targeted sample study. Then Collected the questionnaires, auditing and analyzing them statistically, to answer the questions of the study, and came up with appropriate recommendations in the light of the results. The data collected were analyzed by using SPSS V 23. For factor analysis, estimated abilities of the 10th-grade students, mean, standard deviation (SD), standard error (SE), and correlation coefficient bi-serial. Moreover, it was used (winsteps V 3.72.3) for students' conformity ability on the test.

2. **Results**

After analyzing the data obtained from the instrument, results were presented in table based on the research questions

4.1 First Research Question:

Do the mathematics test data for 10th- grade students achieve the assumptions of (IRT)? To answer this question, the factor analysis was conducted using SPSS V23.0 to verify uni-dimensional assumption test items, as shown in table 1.

Table.1 The result of the factor analysis of the mathematics test items for the 10th grade students.

Some of Square of Saturation			The Eigen values			
Cumulative Explanatory variance %	Explanatory variance	Total	Cumulative Explanatory variance %	Explanatory variance	Total	Component Number
51.49	51.49	12.87	51.49	51.49	12.87	1
			55.13	3.63	0.91	2
			58.38	3.26	0.81	3

Table.1 presents the results of the factor analysis of the mathematics test items indicated to a uni-dimensional investigation of three indicators: the result of dividing the first factor's Eigenvalue by the Eigenvalue of the second greater than 2. Then, the result of dividing the quotient of the root of the second Eigenvalue from the first one on the quotient of the root of the third one from the second one has a high value, the value of the variance explained of the first component is higher than 20.0% (Hattie, 1985).

Figure.1 Showing the Eigenvalues for the factors that make up the test was used with emphasis on a uni-dimensional assumption.

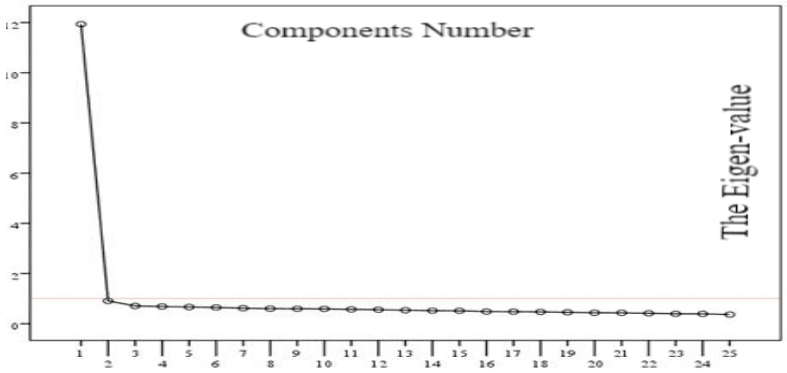

Figure.1 The Plot for Sorting Test of the Values of the Eigenvalues of the Factors of the Test.

Table.2: The

Percent %	Frequency of Correlational Pairs	Status of local independence
1.67	5	Dependent
98.33	295	Independent
100.00	300	Total

frequencies and percentages of local independance of the items of test.

Table.2:Highlighted the assumption of (LI) for the items test was verified by calculating the standard value (♀χ2) of the standardized form of the LI (Standardized LD χ2), each pair of test items (300) has a correlational pair that is calculated by multiplying (25) items by (24) and then dividing by (2), using the (IrtPro V3.1.21505.4001) software. The frequencies and percentages of both LI cases were then monitored provided that the standard LI value is greater than (10), indicating that the LI of a certain number of correlational pair has not been achieved and vice versa if 10 is lower, indicating that the LI of a certain number of correlational pair has been achieved. Moreover, table 2 also shows that LI is achieved in 295 a correlational pair of 300 a correlational pair to items of a test in percentage 98.33%.

4.2 Second Research Question:

To what degree does the mathematics test data conformity for the 10th -grade students with the (PCM)?

Table.3: The descriptive statistics of Non-matching individual's indicators based on (INFIT) and (OUTFIT).

Point biserial Correlation Coefficient	ZSTD		MSQ		The Standard error for ability	Ability	Score	Number of Items	Gender	Number
	OUTFIT	INFIT	OUTFIT	INFIT						
0.21	2.11	1.95	1.75	1.66	0.21	-0.54	38	25	Male	19
0.73	-1.69	-2.47	0.5	0.43	0.25	1.45	80	25	Female	33
0.86	-2.61	-2.88	0.41	0.4	0.21	0.42	60	25	Female	40
0.19	1.82	2.21	1.63	1.75	0.21	-0.2	46	25	Female	43
0.19	2.13	2.22	1.78	1.77	0.22	0.55	63	25	Female	59
0.44	0.35	0.21	1.21	1.20	0.22	0.34	57.40		Arithmetic Mean	

0.33	2.31	2.64	0.70	0.72	0.02	0.77	16.24	Standard Deviation
0.19	-2.61	-2.88	0.41	0.40	0.21	-0.54	38	Minimum Value
0.86	2.13	2.22	1.78	1.77	0.25	1.45	80	Maximum Value

4.3 Third Research Question

What are the estimates of the values of the parameter of the items according to the Partial Credit Model (PCM)? To answer this question, the descriptive statistics conducted to each raw score and the 10th-grade students' ability who's matching the test, the (SE) for ability according to the Rasch model, and (PCM).

Table.4: The descriptive statistics of the matched of individual's indicators based on the ZSTD and MSQ

Point biserial Correlation Coefficient	ZSTD		MSQ		The Standard Error for Ability	Ability	Raw Score	Statistical
	OUTFIT	INFIT	OUTFIT	INFIT				
0.47	-0.07	-0.12	0.99	0.98	0.23	-0.02	49.57	Arithmetic Mean
0.17	0.79	0.80	0.25	0.23	0.03	1.14	23.62	Standard Deviation
0.15	-1.95	-1.81	0.53	0.55	0.21	-1.62	17.00	Minimum Value
0.80	1.69	1.65	1.58	1.53	0.34	2.34	91.00	Maximum Value

Table 2 indicated that the values of the match statistics based on INFIT according to the standard values of the Mean-Square Residuals (MSR) for the observed frequencies of the expected frequencies ranged from (-1.81) to (1.65), and the values of OUTFIT according to the standard values for the Mean-Square Residuals (MSR) of the expected observed frequencies, the expected frequencies ranged from (-1.95) to (1.69).

4.4 Fourth Research Question:

What are the estimates of the values of the person's ability depending on the model used? To answer this question; the difficulty parameter values for the estimated the mathematics test, SE, and the Point biserial Correlation Coefficient for 10th-grade students were calculated according to the (PCM)

Table.5: The descriptive statistics of the matched of individual's indicators based on the ZSTD, MSQ, and the values of the difficulty parameter for the estimated mathematics test items, SE, and the point biserial correlation coefficient.

Point biserial Correlation Coefficient	ZSTD		MSQ		The Standard error for Parameter	Difficulty Parameter	Score	Number of Items
	OUTFIT	INFIT	OUTFIT	INFIT				
0.75	-0.95	-0.81	0.81	0.85	0.15	-0.64	140	1
0.65	0.12	-0.05	1.02	0.99	0.17	0.56	81	2
0.79	-1.87	-1.92	0.67	0.67	0.17	1.07	65	3
0.56	0.94	0.98	1.27	1.21	0.16	-1.22	161	4
0.53	1.69	1.74	1.36	1.37	0.17	0.78	74	5
0.74	0.53	0.55	1.16	1.13	0.15	-0.07	108	6
0.69	1.02	0.91	1.23	1.19	0.15	-0.26	120	7
0.56	1.66	1.49	1.33	1.29	0.17	-0.24	115	8
0.57	0.78	0.70	1.19	1.17	0.16	1.00	57	9
0.66	-0.63	-0.43	0.87	0.92	0.17	0.43	82	10
0.56	1.37	1.48	1.31	1.31	0.15	0.63	76	11
0.68	-0.75	-0.76	0.87	0.87	0.17	-0.69	137	12
0.70	0.44	0.59	1.11	1.12	0.15	0.69	76	13
0.76	-1.09	-0.97	0.80	0.84	0.15	-0.49	128	14

Point biserial Correlation Coefficient	ZSTD		MSQ		The Standard error for Parameter	Difficulty Parameter	Score	Number of Items
	OUTFIT	INFIT	OUTFIT	INFIT				
0.78	-0.35	-0.33	0.91	0.93	0.15	0.01	105	15
0.68	0.72	0.28	1.14	1.05	0.16	0.34	92	16
0.68	0.23	-0.01	1.06	1.00	0.16	0.96	69	17
0.80	-1.48	-1.58	0.73	0.73	0.15	-0.46	129	18
0.75	-0.22	-0.42	0.95	0.92	0.15	-0.31	123	19
0.70	-0.53	-1.61	0.86	0.70	0.16	-1.24	163	20
0.78	-0.48	-0.59	0.87	0.88	0.15	-0.62	134	21
0.81	-1.99	-1.99	0.69	0.67	0.16	0.17	103	22
0.80	-0.91	-1.05	0.83	0.81	0.15	0.09	105	23
0.65	0.76	1.45	1.16	1.31	0.16	-0.17	115	24
0.70	-0.40	-0.40	0.93	0.93	0.17	-0.32	119	25
0.69	-0.06	-0.11	1.01	0.99	0.16	0.00	107.08	Arithmetic Mean
0.09	1.04	1.10	0.21	0.21	0.01	0.66	29.34	Standard Deviation
0.53	-1.99	-1.99	0.67	0.67	0.15	-1.24	57	Minimum Value
0.81	1.69	1.74	1.36	1.37	0.17	1.07	163	Maximum Value

Table 5 highlighted that the (INFIT), according to Mean-Square Residuals (MSR) for observed frequencies on expected frequency range from -1.99 to 1.74. Moreover, the (OUTFIT), according to Standardized Mean-Square Residuals (SMSR) for observed frequencies on expected frequencies ranges from -1.99 to 1.69.

5. Discussions of Findings

The result of the first study question showed, achieving the results of the 10th-grade students on the mathematics test for an un-dimensionality assumption according to the IRT depending on the PCM with three indicators, which means that the performance of the students examined on the test can be attributed to a dominant trait or only one ability, as some (latent- trait models) assume the existence of a single-trait that lies behind the interpretation of the performance of the students examined on the test. Likewise, which means that the test items were homogenous among themselves and measure the same trait and that the items, despite their different difficulties, did not differ among themselves in terms of measuring the same trait.

The results for the first study question also showed achieving the second assumption of IRT. It is a local independence LI, which means that the examined students' responses to the various tests were statistically independent. In other words, the examined students' performance doesn't affect either negatively or positively on the items on the test on (his/her) response to any other items of the test. This means that there is reliability in assessing students' abilities and the difficulty and reliability of the items, despite the difference in the sample of individuals used in the measurement scale as long as it is an appropriate sample. Besides, there is reliability in estimating both the individual's ability and the item difficulty and their reliability, despite the difference in the group of items used in the measurement, as long as it is an appropriate item. Moreover, the result of the first question reveals achieving the monotonicity assumption, which means that the probability of responding correctly to the terms should increase with increasing ability. Besides, for this explanation, this means that the speed factor doesn't play a role in the response of the examined student to the test items, meaning that the reason for the examined student's failure to answer the test items correctly is due to his limited ability, and not because he was unable to reach all the test items because of the speed factor. The findings of this question were consistent with the findings of (Becker & Forsyth, 1992; Craig & Kaiser, 2003; Kimball, 1989; Sebastian & Huang, 2016; Seegers & Boekaerts, 1996; Tambychik, Meerah, & Aziz, 2010; Wu, Wu, Chang, Kong, & Zhang, 2020).

The second study question results showed matching results of the 10th-grade students in terms of ability parameter for them on the mathematics test of the IRT's assumptions, PCM, where only 3 students were deleted. Their answers did not match the expectations of the PCM. Where a non-matching student means that his/her observed responses deviate from the model's expectations, such as he/she may answer about the items incorrectly despite its difficulty level below his/her ability level, or he/she is responding about the items correctly, despite its difficulty level above his /her ability. Moreover, the results of the same question also showed matching the findings of the 10th-grade students in terms of the difficulty parameter of the items on the mathematics test, where none of the test items did match the expectations of the PCM; where is meant by the non-matching item is that the probability of answering the items is high for students with low abilities and low for students with high abilities. The findings of this question regarding matching the items to the PCM are in agreement with the findings of a study and disagreed with the findings of (Afrassa & Keeves, 1999; Bielinski & Davison, 2001; Hashway, 1977; Muraki, 1992; Muthen, 1988).

The third study question results showed that the values of the abilities of the 10th-grade students matched on the mathematics test were ranged from -0.02 to 2. 34 with an AM=49.57, SD=23.62. This means that the student's abilities on the test are not distributed using the normal distribution (ND) as they are supposed to. Moreover, the results showed that the value of the arithmetic mean AM of the student's abilities becomes clear that their abilities are higher than the test level. This question's results agree with the result of the study (Abedalaziz, 2010; Bohlin, 1994) and are not in agreement with the study's result (Stone, 1992).

The fourth study question showed that the values of the difficulty parameter for the mathematics test ranged from 1.00 to -0.32, within AM=107.08, SD=29.34, which means mediating the difficulty of the mathematics test items, as it is not extreme in its difficulty. This question's results agreed with the results of the study(Gershon, 1994; Kelkar, Wightman, & Luecht, 2000; Retnawati, Kartowagiran, Arlinwibowo, & Sulistyaningsih, 2017; Yusha'u, 2013).

6.Conclusion

This study was considered unique in choosing the primary stage in public schools in the Irbid government. It is unique in its approach to construct an achievement test in the mathematics test, in particular. Despite this, it is a study in common with previous studies in its general field of achievement tests and its attempt to identify the degree of effectiveness of applying the PCM in achievement tests. The findings of this study have brought several issues concerning mathematics subject achievement by verifying the test's and reliability and its accomplishment of the suppositions of the (IRT) according to the (PCM). (i) The foreign studies that dealt with the current topic are a lot and various. In other words, there is great interest in the topic by foreign research, whereas it is limited in Arab societies and Arab studies in terms of using and employing it. (ii) All the samples were university students or secondary stage students in these studies. As a result, we will apply it to the primary stage students to demonstrate its effectiveness at this stage, compared to the older age stages. (iii) The majority of the previous research focused on using statistical methods in the light of the Classical Theory to verify psychometric characteristics. A few of them used modern forms in measurement, so it has been applied according to Rasch PCM and demonstrated its effectiveness with the achievement tests.

7. Recommendations

Based on the findings of this study and considering the significant place the mathematics in our educational system, the study made the following recommendations:

(1) Teachers and other stakeholders should pay special attention to encourage and motivate students to develop good study habits to improve their academic achievement in mathematics.
(2) Further studies should be adopted the PCM to see the contribution of this model in measuring students' achievements, especially in the mathematics subject.
(3) Adoption of the current mathematics test by 10th-grade teachers.
(4) Teachers and other stakeholders should endeavor to encourage and motivate students to learn mathematics subjects.
(5) Teachers may need to be more sensitive to the different needs of male and female students. Hence, care has to be placed when teaching both genders.
(6) Curriculum developers should develop instructions that would improve students' knowledge by emphasizing the perceived difficulty areas in mathematics subjects.

References

[1]. Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal, 5*, 101-116.
[2]. Afrassa, T. M., & Keeves, J. P. (1999). Changes in students' mathematics achievement in Australian lower secondary schools over time. *International Education Journal, 1*(1), 1-21.

[3]. Andersen, E. B. (1997). The rating scale model. In *Handbook of modern item response theory* (pp. 67-84): Springer.

[4]. Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594.

[5]. Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. *Mathematical and theoretical systems, 4*, 7-16.

[6]. Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical care*, I7-I16.

[7]. Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement, 29*(4), 341-354.

[8]. Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*(1), 51-77.

[9]. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *psychometrika, 37*(1), 29-51.

[10]. Bohlin, C. F. (1994). Learning style factors and mathematics performance: Sex-related differences. *International Journal of Educational Research, 21*(4), 387-398.

[11]. Chen, L.-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.*

[12]. Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419-440.

[13]. Craig, S. B., & Kaiser, R. B. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods, 6*(1), 44-60.

[14]. De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education, 44*(1), 109-117.

[15]. Dehqan, A., Yadegari, F., Asgari, A., Scherer, R. C., & Dabirmoghadam, P. (2017). Development and validation of an Iranian Voice Quality of Life Profile (IVQLP) based on a classic and Rasch Rating Scale Model (RSM). *Journal of Voice, 31*(1), 113. e119-113. e129.

[16]. Ferguson, L. W. (1942). The isolation and measurement of nationalism. *The Journal of Social Psychology, 16*(2), 215-228.

[17]. Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4*(1), 67-83.

[18]. Gershon, R. C. (1994). Analyzing Multiple Choice Tests with the Rasch Model: Improving Item Calibrations by Deleting Person-Item Mismatches.

[19]. Gómez, L. E., Arias, B., Verdugo, M. Á., & Navas, P. (2012). Application of the Rasch Rating Scale Model to the assessment of quality of life of persons with intellectual disability. *Journal of Intellectual and Developmental Disability, 37*(2), 141-150.

[20]. Haladyna, T. M., & Downing, S. M. (2011). Twelve steps for effective test development. In *Handbook of test development* (pp. 17-40): Routledge.

[21]. Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice, 12*(3), 38-47.

[22]. Hashway, R. M. (1977). *A comparison of tests derived using Rasch and traditional psychometric paradigms.* ProQuest Information & Learning,

[23]. Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and ltenls. *Applied Psychological Measurement, 9*(2), 139-164.

[24]. Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in M Plus. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 308-318.

[25]. Kelkar, V., Wightman, L. F., & Luecht, R. M. (2000). Evaluation of the IRT Parameter Invariance Property for the MCAT.

[26]. Khasawneh, A. A. (2009). Assessing Logo programming among Jordanian seventh grade students through turtle geometry. *International Journal of Mathematical Education in Science and Technology, 40*(5), 619-639.

[27]. Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological bulletin, 105*(2), 198.

[28]. LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational research methods, 14*(1), 10-23.

[29]. Lawley, D. N. (1943). XXIII.—On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 61*(3), 273-287.
[30]. Lord, F., & Novick, M. (1968). Statistical Theories of Mental Test Scores. Addison Wesley Publ & Co. *Reading, Mass*.
[31]. Masters, G. N. (1982). A Rasch model for partial credit scoring. *psychometrika, 47*(2), 149-174.
[32]. Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *psychometrika, 55*(2), 195-215.
[33]. Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological review, 47*(4), 355.
[34]. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30.
[35]. Muthen, B. O. (1988). Instructional Sensitivity in Mathematics Achievement Test Items: Application of a New IRT-Based Detection Technique.
[36]. Nering, M. L., & Ostini, R. (2011). *Handbook of polytomous item response theory models*: Taylor & Francis.
[37]. Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*: Sage.
[38]. Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*: Sage.
[39]. Rabab'h, B. S. H., & Veloo, A. (2015). Spatial visualization as mediating between mathematics learning strategy and mathematics achievement among 8th grade students. *International Education Studies, 8*(5), 1-11.
[40]. Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements in symposium on scientific objectivity, Vedbaek, Mau 14-16, 1976. *Danish Year-Book of Philosophy Kobenhavn, 14*, 58-94.
[41]. Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why Are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It? *International Journal of Instruction, 10*(3), 257-276.
[42]. Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *psychometrika, 1*(2), 33-49.
[43]. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
[44]. Sebastian, J., & Huang, H. (2016). Examining the relationship of a survey based measure of math creativity with math achievement: Cross-national evidence from PISA 2012. *International Journal of Educational Research, 80*, 74-92.
[45]. Seegers, G., & Boekaerts, M. (1996). Gender-related differences in self-referenced cognitions in relation to mathematics. *Journal for Research in Mathematics Education, 27*(2), 215-240.
[46]. Steinberg, J. (2000). Frederic lord, who devised testing yardstick, dies at 87. *New York Times*.
[47]. Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16.
[48]. Tambychik, T., Meerah, T. S. M., & Aziz, Z. (2010). Mathematics skills difficulties: A mixture of intricacies. *Procedia-Social and Behavioral Sciences, 7*, 171-180.
[49]. Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories.
[50]. Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement, 69*(1), 50-61.
[51]. Tuckman, B. W. (1993). The essay test: A look at the advantages and disadvantages. *Nassp Bulletin, 77*(555), 20-26.
[52]. Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in psychology, 11*, 2230.
[53]. Yusha'u, M. (2013). Difficult topics in junior secondary school mathematics: Practical aspect of teaching and learning trigonometry. *Scientific Journal of Pure and Applied Sciences, 2*(4), 161-174.
[54]. Zaman, A., Kashmiri, A.-U.-R., Mubarak, M., & Ali, A. (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT.