The Effect of Multiple-Choice Test Items' Difficulty Degree on the Reliability Coefficient and the Standard Error of Measurement Depending on the Item Response Theory (IRT)

Dr. Habis Saad Al-zboon[1], Dr. Amjad Farhan Alrekebat[1]

& Prof. Mahmoud Sulaiman Bani Abdelrahman[1]

[1] Department of Curriculum and Instruction, College of Education, Al-Hussein Bin Talal University, Jordan

Correspondence: Dr. Habis Saad Al-zboon, Department of Curriculum and Instruction, College of Education, Al-Hussein Bin Talal University, Jordan.

Received: April 13, 2021 Accepted: May 16, 2021 Online Published: June 2, 2021

doi:10.5430/ijhe.v10n6p22 URL: https://doi.org/10.5430/ijhe.v10n6p22

Abstract

This study aims at identifying the effect of multiple-choice test items' difficulty degree on the reliability coefficient and the standard error of measurement depending on the item response theory IRT. To achieve the objectives of the study, (WinGen3) software was used to generate the IRT parameters (difficulty, discrimination, guessing) for four forms of the test. Each form consisted of (30) items with different difficulty coefficients averages (-0.24, 0.24, 0.42, 0.93). The resulting items parameters were utilized to generate the ability and responses of (3000) examinees based on the three-parameter model. These data were converted into a readable file using the (SPSS) and the (BILOG-MG3) software. Then the reliability coefficients for the four test forms, the items parameters, and the items information function were calculated, and dependence on the information function values to calculate the standard error of measurement for each item.

The results of the study showed that there are statistically significant differences at the level of significance ($\alpha \leq 0.05$) between the averages of the values of the standard error of measurement attributed to the difference in the difficulty degree of the items in favor of the test with the higher difficulty coefficient. The results also found that there are apparent differences between the test reliability parameters attributed to the difficulty degree of the test according to the three-parameter model in favor of the form with the average difficulty degree.

Keywords: test reliability, standard error of measurement, item response theory (IRT)

1. Introduction

Achievement tests are considered to be one of the most important various assessment methods that are relied upon when making important decisions concerning the individual and society. The use of tests has spread widely in many areas. They are designed for various purposes, among which: choosing a person for a job; or for classification purposes such as determining the path of learners in proportion to their abilities and skills; and in evaluating students' achievement through the grades they obtain in class tests. Thus, it is possible to work on improving and developing the educational and learning process and moving it forward for the better through developing these tests, whether verbal or performative, and improving their ability to measure learning outcomes (Haladyna, 2004).

Moreover, tests are one of the most important educational methods through which students' performance is assessed, as they provide the final output of the educational process, so they must be prepared carefully, taking into account the availability of the objectivity, honesty and reliability factors, so that the tests yield the desired result (Al-Hariri, 2007).

(Baghaei & Amrahi, 2011) believe that multiple-choice items are the best types of objective items, and the most common and widespread in achievement tests. They are easy to correct and provide good coverage of the subject matter. The student's score on them is characterized by a high degree of reliability. In addition, they determine the intended learning outcomes to a high degree, although their preparation requires a long time, great effort, and great skill on the part of their authors. Furthermore, the multiple-choice items are able to measure learning outcomes at the higher mental levels of the cognitive domain to a degree that exceeds the matching items, true or false items, filling the blanks items, and short answers.

(Crocker and Algina, 2009) see that when building any test, its user prefers to obtain the same results when applying it to the same individuals in similar circumstances. This consistency in test scores is called reliability.

Reliability is one of the most important characteristics of a good test, as a test is supposed to be reliable and gives the same order to students almost every time it is applied to the group of students. Nevertheless, the high or low marks of the student group does not mean that the test is not reliable, especially in the field of achievement (Odeh, 2014).

Reliability is related to the accuracy of the measuring instrument when the test is repeated more than once in the same conditions, as the absence of measurement errors leads to an increase in the reliability of test scores, and thus the ease of interpretation (Moss, 1995).

There are many sources of measurement errors related to the test. Among these sources is the extreme difficulty of the test items compared to the level of students. This extreme difficulty encourages students to random guesses, and thus not obtaining true scores, which increases measurement errors and thus affects the value of the test reliability coefficient.

Classical models were used in the past decades to design achievement tests. Nevertheless, their benefits were limited due to the method used in analyzing those tests which were based on the foundations of that traditional theory as well as the psychometric and statistical concepts associated with it. If we look at the difficulty and discrimination coefficients, we find that they vary according to the average and the extent of the ability of the sample members used in calculating these parameters. Thus, the benefit from these coefficients becomes limited to a community similar to the community from which the sample was chosen as the scores of examinees in a test depend on the sample of the items that the test includes. Measurement scientists have tried to benefit from the technological advancement in finding new psychometric methods and solving these problems through what is known as the latent feature theory or the item response theory (Allam, 1986).

Many believe that the test scores reflect to some extent the amount of knowledge the individual possesses but in fact it does not. These scores include a certain amount of error which can be an increase or a decrease in the score. The increase comes from obtaining some marks from other sources such as the degree of test difficulty. When the degree of the test difficulty increases, this leads students to guess or cheat. On the other hand, the decrease comes from the loss of some knowledge due to forgetfulness. Hence, the apparent score does not reflect the actual amount of knowledge the individual possesses because it includes a percentage of error which may affect the test reliability. This problem can be overcome by controlling the sources of error. Therefore, this study came to identify the effect of multiple-choice test items' difficulty degree on the reliability coefficient and the standard error of measurement depending on the item response theory IRT.

2. The Study Problem

Achievement tests, which are considered one of the most important measurement tools to determine the performance of the examinees depend on the score obtained by the examinee in the test according on the classical test theory.

The tool that is relied upon in measuring the examinee's performance must be truthful and gives results and indicators that can be relied upon when making decisions. It is known that the examinee's score on the tests is expected to be sufficient evidence of the extent to which the examinee possesses the skill or knowledge measured in the tests. This means that external variables such as the difficulty degree of the test items should not have an effect on performance. Most studies have indicated that when the test items or ordered from easy to difficult with a degree of medium difficulty, taking into account the individual differences of students, this provides reinforcement for the examinee and increases his motivation to answer the items of the test. Thus, the examinee will obtain a higher score when the test is of medium difficulty (Odeh, 2010); (Hambleton & Traub, 1974). This would affect the reliability of the test and the standard error of measurement.

By reviewing the studies that dealt with this topic, we find that they dealt with it from the viewpoint of the classical test theory. Therefore, this study came to know the effect of the test items difficulty on the test reliability and the standard error of measurement depending on the item response theory IRT. Consequently, this study seeks to answer the following questions.

The Study Questions:

- **The first question:** Are there statistically significant differences between the test reliability coefficients attributed to the difficulty degree of the items according to the three-parameter model in the item response theory?

- **The second question:** Are there statistically significant differences between the values of the standard error of measurement attributed to the difficulty degree of the items according to the three-parameter model in the item response theory?

The Importance of the Study:

The importance of this current study lies in the following:

1. The scarcity of studies that dealt with the effect of the test item’s difficulty on the test reliability, depending on the item response theory.
2. The scarcity of studies that dealt with the effect of the test item’s difficulty on the standard error of measurement in estimating the item difficulty parameter depending on the item response theory.
3. This study seeks to determine the appropriate difficulty degree of the test items which achieve the best reliability for the test.
4. Providing test authors with the necessary information that helps them build tests with a high degree of reliability by determining the best difficulty degree of the test items.

The Study Objectives:

1. Identifying the potential differences between the test reliability coefficients due to the difficulty degree of the items according to the three-parameter model in the item response theory.
2. Identifying the potential differences between the values of the standard error of measurement due to the difficulty degree of the items according to the three-parameter model in the item response theory.

Terms Definition:

- **The Test:** A measurement tool prepared according to an organized method of several steps that include a set of procedures that are subject to specific conditions and rules, with the aim of determining the degree of an individual's possession of a certain characteristic or ability through his response to a sample of stimuli that represent the characteristic or ability to be measured (Odeh, 2014).
- **Item Difficulty Parameter (Threshold):** The ability level that corresponds to the probability of 0.5 for answering the item correctly when the guessing coefficient is equal to zero. (Hambleton & Swaminathan, 1985)
- **Standard Error of Measurement:** A measure of dispersion associated with (ϴ) estimated values for some examinees about their true ability value (ϴ), which is inversely related to the square root of the test information function (Al-Taqi, 2013)
- **Reliability Coefficient:** The ratio of variance in the true score to the variance in the observed score (Al-Nabhan, 2004). It is defined as the quantity of the test information function, which indicates the accuracy of the score that reflects the examinee's ability to represent this ability.

3. The Theoretical Framework and Previous Studies

Test Reliability

Reliability is statistically defined as the ratio of true variance to the total variance, that is, how much of the total variance in scores can be true, whether or not it is related to the measured characteristic. (Odeh, 2010).

It is defined as the quantity of the test information function, which indicates the accuracy of the score that reflects the examinee's ability to represent this ability.

(Arifij, 1987) stated that reliability means objectivity, this means that the results are not substantially affected if the examiner or the grader changes. Consistency may mean that the examinee's mark on the test part is related to his score on the test as a whole.

(Melhem, 2005) indicated that one of the meanings of reliability in measurement is consistency, so if we say that the test achieves the characteristic of reliability, this means that the test measures anything consistently. Reliability answers this question: Do we get the same score (or close to it) each time this test is performed on this individual? Hence it is possible for the scale to be consistent even though it does not measure the characteristic we wish to measure. Reliability means the consistency and harmony with which the test scores measure the characteristic or the thing that the test was prepared to measure. As for the validity of the test, it is the extent to which the test measures the characteristic that it is intended to measure.

Test reliability is one of the basic components of a good test, as it is assumed that a test gives almost the same results when it is reused at different times. For example, the meter is a reliable test because it gives the same results in measuring the length of things (Jaradat, 2011).

The concept of reliability in the item response theory is related to the item information function $I_i(\theta)$ and the test information function $I(\theta)$ and to the standard error of estimating the capabilities of the examinees SEE. (Thissen, 2000) showed that the best method for estimating the reliability coefficient is based on the test information function. The relationship between reliability and the test information function can be represented by the equation (Rxx = 1- $\frac{1}{\sum_{i=1}^{n} I(\theta)}$), where Rxx denotes the test reliability coefficient, and $I_i(\theta)$ denotes the item information function. This equation confirms that the relationship between the test information function and the test reliability is direct proportionality.

To find the value of the empirical reliability coefficients, the statistical program (BILOG-MG3) was used, and the value of the empirical reliability coefficient indicates the amount of information we obtain from the test. (Al-Sharifain, 2009).

Standard Error of Measurement

The reliability coefficient is an estimate of the correlation coefficient between the scores of a group of examinees in a particular test, and the scores of the same examinees in another test that is equivalent to the first test. The higher this coefficient, the greater the consistency of the test in measuring what it is designed to measure. Complete reliability cannot be obtained from a practical point of view, which is represented by a stability coefficient of (1.00). Although the values of the reliability coefficient such as (0.96) or higher are mentioned in reports and some research, most test designers are satisfied if their tests give a reliability coefficient of around (0.90). On the other hand, the reliability coefficient for tests that teachers prepare tends to not reach this value.

Another way to look at and interpret the reliability coefficient is by considering it as the ratio of the variance of the true scores to the variance of the observed scores obtained by the examinees. The true score of an individual in a specific test is a hypothetical score by which we mean the average of a large number of scores that the same individual can obtain on similar tests under favorable conditions. While the observed score, it is the score obtained in a specific test.

The difference between the observed score (X) and the true score (T) is called the 'error of measurement' (E). So, the relationship between these scores is:

X = T + E

That is, the observed score (X) of a given test consists of two parts, the first being the true score (T) and the second being the error (E).

An individual's score observed in a test differs in most cases from his true score, due to the fact that the observed score is affected by multiple sources of errors. If we assume that we can determine the degree of the random errors that affected the observed score for each individual, then the standard deviation of the error degrees can be found, and the resulting value is called the 'Standard Error of Measurement'. In fact, we cannot find the degree of error for each individual in a group unless the test is repeated on the same individual a large number of times, which is not possible. Therefore, we cannot find the standard error of the measurement, but we can estimate this value if we know the value of the standard deviation of the observed scores, as well as the value of the reliability coefficient of test scores, using a mathematical formula that can be directly derived from the following equation:

$$\frac{\sigma_{\mathrm{E}}^2}{\sigma_{\mathrm{X}}^2} = 1 - \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{X}}^2}$$

Since $\frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{X}}^2}$ is the reliability coefficient ρ, then: $\frac{\sigma_{\mathrm{E}}^2}{\sigma_{\mathrm{X}}^2} = 1 - \rho$

Cross multiplication results in: $\sigma_{\mathrm{E}}^2 = \sigma_{\mathrm{X}}^2(1 - \rho)$

By calculating the square root of each side, we find that: $\sigma_{\mathrm{E}} = \sigma_{\mathrm{X}}\sqrt{(1-\rho)}$

That is, the standard error of measurement equals the standard deviation of the observed degrees multiplied by the square root of 1 minus the reliability coefficient (Allam, 2002).

The standard error of measurement in the item response theory is related to the information function to the degree of accuracy in estimating the information function, as the shape of the information function distribution leads to important information. The value of the standard error of measurement is related to the information function by the following equation: $SSE = \frac{1}{\sqrt{I(\theta)}}$

This means that by estimating the standard error of measurement for each ability level, the information function has a meaning that helps in understanding the accuracy of the measurement and thus the test reliability. Where it is clear from the equation that the relationship between the standard error of measurement and the information function is an inverse relationship, so the greater the standard error of measurement, the lower the values of the information function.

Item Response Theory (IRT)

IRT is a theory of measurement that is based on the probabilistic relations between responses to items in a test and the construct a test aims to measure (Schultz & Whitney, 2005). The construct that is intended to be measured with a test but is not directly observed is called a latent trait in IRT. For this reason, the Latent Trait Theory is another name for IRT (De Ayala, 2009).

Moreover, the IRT represents the contemporary trend in psychological and educational measurement. It is called item response theory and latent trait theory. The credit for presenting the foundations of the item response theory to those interested in psychometrics and education goes to Lord (Allam, 2002).

IRT with its different models overcomes the problem in selecting items according to classical methods. It provides a method of selecting the items and the ability of the examinee in a way that the developer of the test can choose the most effective items in a range determined by a cut-off mark on the ability scale that helps to separate the levels of mastery and mastery on the scale (Hambleton & Rogers, 1991). The item response theory (the modern theory of measurement) proposes a model for the relationship between an unobservable variable used to measure the abilities intended to be measured by the test, and the probability of a correct response on a given item. This is being done through logarithmic functions linking the examinee's ability and the item parameters to the probability of the correct answer to it. Multiple models have emerged from the IRT, all of which assume that a single ability measures performance on the test. The ability can be represented on an infinite continuum, but it changes in its characteristics described by the items. The difficulty and ability degrees extend theoretically over a continuum ranging from (-∞) to (+∞), but it ranges practically between (- 3 and +3). That is because it is rare to have values greater than (+3) or less than (-3) (Hambleton & Swaminathan, 1985).

Differing models were suggested throughout the historical development of IRT. The first model suggested within the framework of IRT was the Rasch model, which was developed for items rated in two categories and contains only difficulty parameters (DeVellis,2003). A two-parameter model was developed with the inclusion of a discrimination parameter in the Rasch model, and a three-parameter model was developed with the inclusion of a guessing parameter in the two-parameter model (Furr & Bacharach, 2008).

As can be understood, the first factor influential in the emergence of different models in the development process of IRT was the number of estimated item parameters. The second factor was the response categories in relation to items. IRT was first developed for items that were rated dichotomously. However, later, the use of the theory was not limited to dichotomously rated items and, thus, models for polytomous items (nominal response model, partial credit model and graded response theory) were also included in IRT (Harvey & Hammer, 1999; van der Linden, 2005, Embreston & Rise,2000)

IRT is divided into two categories, parametric and non-parametric models, in terms of approaches considered in estimating the item characteristic curve, while parametric IRT models assume that the item characteristic curve normal ogive or logistic properties, non-parametric models do not have an assumption limiting the item characteristic curve to a certain form (Takno, Tsunoda & Muraki,2015).

The models of the two-stage item response theory vary depending on the three parameters of the item. A set of mathematical models known as the "latent traits models" emerge from it. Each of those models depends on a mathematical equation that determines the relationship of the individual's performance on an item with his ability that lies behind and explains this performance. The IRT includes the logistical models (Hambleton & Rogers, 1991).

The item response theory has a set of features mentioned by (Hamilton & Swaminathan, 1985):

1. Item parameters are independent of the group of examinees used from the population of examinees for whom the test was designed (item-free).
2. Examinee ability estimates are independent of the particular choice of test items used from the population of items that were calibrated (person-free).
3. There is a statistic indicating the accuracy in estimating the ability of each individual, such as the standard error of estimation. This statistic is different from one individual to another, and it is found for each item. It is not constant for all items. The most important characteristic of this theory in psychological and educational measurement is the possibility of obtaining the item statistics that does not depend on the characteristics of the examinees; and the scores that express the ability of the subjects do not depend on the characteristics of the items.

Many studies have been conducted which examined the effect of the difficulty degree of the multiple-choice test items on the reliability coefficient and the standard error of measurement based on the item response theory. Among these studies is (Hambleton & Traub, 1974) which aimed to study the effect of ordering the test items according to their difficulty level on the performance in a math test and on the anxiety generated during the test. In order to achieve the objectives of the study, an achievement test consisting of (40) items was prepared and applied to (250) examinees in order to calculate the items difficulty coefficients. Accordingly, two forms of the test were prepared based on the difficulty coefficients. The items were arranged in ascending order in the first test form, and in descending order in the second test form. Then the two test forms were applied to (106) examinees who were subjected to the Achievement Anxiety Test (AAT). One of the most important findings of the study is that the order of the test items in ascending order leads to higher scores than the scores obtained when the test items are arranged in descending order. Moreover, it was found that the arrangement of the test items affects the examinees' scores in the (AAT).

(Crehan, K. D., Haladyna, T. M., & Brewer, B. W., 1993) aimed at determining the optimal number of alternatives in the multiple-choice test in terms of difficulty and discrimination coefficients. In order to achieve the study objectives, an achievement test consisting of (40) items was prepared. Two equivalent forms of the test were prepared. One form with three alternatives and the other with four alternatives. The two forms were applied to a random sample of (220) students. The difficulty coefficients were calculated for each model. The average difficulty coefficient was (0.80) for the three-alternative form and (0.77) for the four-alternative form. The averages of discrimination coefficients were (0.35) for the three-alternative form and (0.36) for the four-alternative form. It was found that there were statistically significant differences between the average of difficulty coefficients in favor of the three-alternative form, as it was found that its items were easier than those of the four-alternative form.

(Za'al, 2010) aimed to study test anxiety and the arrangement of test items according to their difficulty degree on the achievement of ninth-grade school students in mathematics. To achieve the study objectives, an achievement test in mathematics was prepared as well an anxiety test. The study sample consisted of (516) male and female students. Among the most important findings of the study is that there are no statistically significant differences at the level of significance ($\alpha = 0.05$) between the averages of students' performance on the mathematics test attributed to the test items arrangement according to their degree of difficulty. The study also found the presence of statistically significant differences at the level of significance ($\alpha = 0.05$) among the averages of female students’ performance attributed to the test items arrangement according to their difficulty degree in favor of both ascending and random order.

(Al-Lihyani, 2010) entitled "The effect of arranging the test items on the validity and reliability of the multiple-choice test in mathematics for high school students in Makkah. The study aimed to identify the most important methods of arranging the test items according to the sequence of course content, as well as the arrangement according to the difficulty coefficients. The study also aimed at determining the best pattern of arrangement for test vocabulary and its effect on the validity and reliability of the multiple-choice achievement tests. To achieve the objectives of the study, a multiple-choice achievement test in mathematics was prepared for the second secondary grade consisting of (30) items. One of the most important results of the study is that there are no statistically significant differences between the values of the internal consistency coefficients calculated by the

Cronbach Alpha equation attributed to the difficulty degree.

(Ma'rouf, 2013) aimed at identifying the effect of arranging the test items according to their difficulty level on the psychometric characteristics of the intelligence test that is not influenced by culture. In order to achieve the objectives of the study, four forms of the test were prepared. In the first form, the test items were arranged in ascending order; in the second form items were arranged in descending order; in the third form items were arranged randomly, while in the fourth form items were arranged in a circular order. The test was applied to (2040) male and female students. The study found that: i. there were no statistically significant differences between the Cronbach Alpha coefficients for the test scores attributed to the arrangement of the test items according to their level of difficulty. ii. there were statistically significant differences in the reliability coefficients of the half-segmentation of the test scores according to the Spearman-Brown equation in favor of the circular arrangement.

(Ilhan & Guler, 2018) aimed to compare difficulty indices calculated for open-ended items in accordance with the classical test theory (CTT) and the Many-Facet Rasch Model (MFRM). Although theoretical differences between CTT and MFRM occupy much space in the literature, the number of studies empirically comparing the two theories is quite limited. Therefore, this study is expected to be a substantial contribution to the literature. The research data were collected through three teachers rating the answers given by 375 eighth-grade students to ten open-ended questions in a mathematics test. The difficulties of the items in the test were calculated according to CTT and MFRM by using the obtained data, and the consistency between the difficulty indices estimated based on the two theories was tested. While the Microsoft Excel program was used in the analyses for CTT, the FACETS package was employed in the analyses for MFRM. Findings: The research findings showed that CTT and MFRM yielded similar results in terms of difficulty indices of open-ended questions. It was found that, according to both theories, the ten items in the achievement test were ranked as I2, I3, I1, I4, I7, I6=I8, I5, and I9, from easiest to most difficult. Implications for Research and Practice: It may be said that estimating item difficulties according to either CTT or MFRM will not cause any notable differences in terms of the items to be included or excluded in the development of an achievement test with open-ended questions.

4. Method and Procedures

This study used the experimental simulation approach. Data were generated using the (WinGen3) software and were studied using the (SPSS) and (BILOG-MG3) software to answer the study questions according to the following steps:

Data Generation:

First: Generating the test based on the three-parameter model:

1. Generating four test forms, each form consisted of (30) items with different difficulty factors averages (-0.24, 0.24, 0.42, 0.93) according to the three-parameter model using (WinGen3) software based on the IRT.
2. Generating the items discrimination parameter according to the Log normal distribution ~ (0,0.25) based on the three-parameter model. After generating the data, the mean of the discrimination parameter values were (1.07, 0.95, 0.73, 0.94) and the standard deviation values were (0.28, 0.19, 0.29, 0.21). These values were considered good compared with the criterion defined by (Hambleton & Swaminathan, 1985) which states that the true values of the discrimination parameter range from (0-2) Logit.
3. Generating the items difficulty parameter according to the normal distribution ~ (0,1) based on the three-parameter model. The mean of the difficulty parameter values were (-0.24, 0.24, 0.42, 0.93) and the standard deviation values were (1.14, 0.94, 0.89, 0.25) respectively.
4. Generating the items guessing parameter according to the Beta distribution ~ (8,32) based on the three-parameter model. This distribution produces values for the guessing parameter which are similar to the objective test (two-response) consisting of five alternatives. The mean of the guessing parameter values were (0.21, 0.20, 0.23, 0.28) and the standard deviation values were (0.08, 0.06, 0.07, 0.08) respectively.

Second: Generating responses:

Responses of (3000) examinees were generated using the same values for the actual item’s parameters previously generated according to the normal distribution ~ (0,1).

Data Analysis:

1. To achieve the objectives of the study, the (WinGen3) software was used to generate data. The items parameters (difficulty, discrimination, guessing) were generated for the three forms of the test. The resulting items parameters were relied upon to generate the ability of (3000) examinees according to the three-parameter model based on to the IRT.
2. Generating the examinee's responses using (WinGen3) software.
3. Using the (SPSS) software to convert these data into a readable file for the (BILOG-MG3) software.
4. Calculating the reliability coefficients for the four test forms and the item's parameters.
5. Calculating the information function of the four test forms according to the three-parameter model in the IRT.
6. Utilizing the information function to calculate the standard error of measurement for each item, depending on the equation that links the information function with the standard error of measurement $SSE = \frac{1}{\sqrt{I(\theta)}}$.

Data Goodness of Fit:

The (BILOG-MG3) software was utilized to match individuals and items to the models of the item response theory. The data of (3000) examinees were analyzed, and the results indicated that all the items were fit to the model as the value of Chi-Square test (χ2 test) is not statistically significant at the level of significance ($\alpha \leq 0.05$). Moreover, the results of the analysis showed that all the responses of the examinees matched with the expectations of the models except for sixteen examinees where the value of Chi-Square (χ2) was statistically significant at the level of significance ($\alpha \leq 0.05$).

5. Results and Discussion

The first question: Are there statistically significant differences between the test reliability coefficients attributed to the difficulty degree of the items according to the three-parameter model in the item response theory?

To answer this question, the test reliability coefficients based on the degree of difficulty of the items were found according to the three-parameter model in the item response theory using the equation ($R_{xx} = 1 - \frac{1}{\sum_{i=1}^{n} I(\theta)}$), where R_{xx} refers to the test reliability coefficient, and $I_i(\theta)$ refers to the item information function. The z-test was also used to identify the significant of differences. Table (1) shows these results:

Table 1. Reliability coefficients for each of the test forms

Test form	Average of item difficulty	Reliability coefficient
1	-0.24	0.73
2	0.24	0.85
3	0.42	0.76
4	0.93	0.71

It is evident from Table (1) that there are apparent differences between the reliability coefficients values and that the best reliability coefficient is (0.85) when the item difficulty was (0.24). This shows that the best reliability coefficient is when there is no extremism in the difficulty degree of the test items, i.e., test questions should not be very difficult or very easy. Their difficulty should be graded from easy to difficult, as the test reliability decreases when the variability of its items decreases, and vice versa. Thus, questions that are very easy or very difficult lead to a decrease in the reliability coefficient because they are not variant or graded in difficulty.

To find out the significance of the differences between the reliability coefficients, the Fisher Equation was used to convert the reliability coefficients into z-values and examine their significance using the Fisher Equation as shown in the following table:

Table 2. Reliability coefficients for each of the test forms and their corresponding z-values and significance level

Test form	Average of item difficulty	Reliability coefficient	Comparison	Z-value	Sig
1	-0.24	0.73	1*2	-12.68	0.00
2	0.24	0.85	1*3	-2.61	0.00
3	0.42	0.76	1*4	1.61	0.054
4	0.93	0.71	2*3	10.06	0.00
			2*4	14.82	0.00
			4*3	4.22	0.00

It is evident from Table (2) that the differences between the dual comparisons of the reliability coefficients for the four test forms were statistically significant at the level of significance ($\alpha \leq 0.05$). It is noticed that the second form is the most stable compared to the other forms, followed by the third one. It is also noticed that the test was less reliable when the test was extremely easy or difficult. It is clear that there are no statistically significant differences between the first and fourth forms. This confirms that the reliability coefficient will be the best in the case of arranging the test items in terms of difficulty from the easiest to the difficult. This arrangement provides students with motivation to keep on trying to answer when they receive immediate reinforcement because of their ability to answer the first questions of the test, which are called encouraging questions or shock-absorbing questions. These results are consistent with the results of (Hambleton & Traub, 1974) and (Za'al, 2010) studies.

The second question: Are there statistically significant differences between the values of the standard error of measurement attributed to the difficulty degree of the items according to the three-parameter model in the item response theory?

To answer this question, the one-way ANOVA analysis was used for the values of the standard error of measurement based on the variance in the difficulty degree of the items according to the three-parameter model in the item response theory. Table (3) shows these results:

Table 3. The One-way ANOVA analysis to examine the significance of the differences between the means of the values of the standard error of measurement according to the difficulty degree of the items

Source of variance	Sum of squares	DF	Means squares	F	Sig
Between groups	20.33	3	6.78		
Within groups	211.91	116	1.83	3.71	0.014
Total	232.24	119			

It is clear from the table (3) that there are statistically significant differences between the means of the values of the standard error of measurement according to the difficulty degree of the items.

To find out the direction of the differences and to which test form these differences belong, Scheffe test for post-hoc comparisons was used. Table (3) shows the results of the post-hoc comparisons.

Table 4. The results of the post-hoc comparisons between the averages of standard errors of measurement for the items of the four test forms

Comparisons	Mean Difference	Sig
1*2	-0.90	0.09
1*3	-0.78	0.18
1*4	-1.07*	0.03
2*3	0.13	0.98
2*4	-0.17	0.97

It is clear from table (4) that there are statistically significant differences between the averages of the standard errors of measurement for the first and fourth test forms in favor of the fourth test form -which has the higher item difficulty. This may be attributed to the fact that the test form with a higher difficulty coefficient has a higher level of items difficulty, which leads students to cheat or guess. Thus, the student's observed score will be far from his real score (T = X + E), which increases the standard error of measurement for the test items. Also, the extreme difficulty in the test items may encourage the students to guess randomly, which leads to their scores being close so that the group of students appears as a homogeneous group, meaning that the test has a weak discerning ability, which increases the standard error of measurement.

6. Conclusions and Recommendations

The results of the study showed that the standard error of measurement differs according to the items difficulty degree in favor of the test with a higher difficulty coefficient. That is, the standard error of measurement increases with the increase in test items difficulty. The results also concluded that there are apparent differences between the test reliability coefficients attributed to the difficulty degree of the test based on the three-parameter model in favor of the test form with moderate difficulty degree. This means that the best reliability coefficient was for the test with moderate items difficulty degree.

Based on these results, the study recommends the need to construct achievement tests of medium difficulty and not relying on extremely difficult tests. The study also recommends conducting further studies based on the one-parameter and two-parameter models.

References

Abdulhameed, Ezzat. (2011). *Psychological and educational statistics: applications using SPSS 18 software*. Dar Al-Fikr, Cairo, Egypt.

Al-Hariri, Rafidah (2007). *The comprehensive educational assessment of school institutions*. Dar Al-Fikr Distributors and Publishers. Amman, Jordan.

Al-Lihyani, Adnan (2010). *The effect of the items number, the location of the item in the test, the size of the sample, and the number of alternatives on the value of the scale reliability coefficient.* Unpublished master's thesis. Umm Al-Qura University, KSA.

Al-Shraifin, Nidal (2009). The effect of the item formulation pattern in a trend scale on the psychometric properties of the items, the scale, and the ability estimates of the individuals according to the item response theory. *Journal of Educational and Psychological Sciences*. Vol (15). Issue (4). https://doi.org/10.12785/JEPS/100401

Al-Taqi, Ahmad (2013). *The modern theory of measurement.* Dar Al-Maseerah for Publishing, Distribution and Printing, 2nd. ed. Amman, Jordan.

Allam, Salah (1986). *Contemporary developments in psychological and educational measurement and assessment*. College of Arts, Kuwait University. Dar Al Qabas Press.

Allam, Salah Al-Din Mahmoud (2002). *Educational and psychological measurement and assessment: basics, applications, and contemporary trends*. Egypt, Cairo: Arab Thought House.

Arifij, Sami. (1987). *In Measurement and Assessment*. 3rd ed, Rafidi Press.

Baghaei, P., & Amrahi, N. (2011). the effect of the number of options on the psychometric characteristics of multiple-choice items. *Psychological test and assessment modeling, 53*(2), 192-211. https://doi.org/10.4304/jltr.2.5.1052-1060

Crokr, L., & Algina, J. (1986). *Introduction to classical and modern test theory, new yourk*: Holt pine hart and Winston.

De Ayala, R. J. (2009). The theory and practice of item response theory. New York.

DeVellis, R. F. (2003). Scale Development: *Theory and Application*. Thousand Oaks, CA: Sage.

Embretson, S., & Reiase, S. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates. Inc.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics*: An introduction. Thousand Oaks, CA: Sage.

Granhan, K. D., Haladyana, T. M., & Brewer, B. W. (1993). Use of an inclusive option and optimal number of option for multiple–choice items. *Educational and Psychological Measurement*, 53, 241-247. https://doi.org/10.1177/0013164493053001027

Haladyna, Thomas, M. (2004). *Developing and validating multiple–choice test items*. Lawrence Erlbaum associates publishers, Mahwah, new jersey, usa.25. https://doi.org/10.4324/9780203825945

Hambleton, R. K., & Traub, R. E. (1974). The effect of item order on test performance and stress. *journal of* *experimental education*, 43, 40-46. https://doi.org/10.1080/00220973.1974.10806302

Hambleton, R. K., & Swaminathan (1991). *Fundamentals of Item Response Theory*: Sage Publications.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston MA: Kluwer-Nyjhoff. https://doi.org/10.1007/978-94-017-1988-9

Harvey, R. J., & Hammer, A. L. (1999). Item response *Theory. the counseling psychologist*, *27*(3), 353-383. https://doi.org/10.1177/0011000099273004

Ilhan, M., & Guler, N. (2018). A Comparison of Difficulty Indices Calculated for Open-Ended Items According to Classical Test Theory and Many Facet Rasch Model. *Eurasian journal of educational Research*, 75, 99-114.

Jaradat, Ezzat; Obaidat, Thouqan; Abu Ghazaleh, Haifa; & Abdullatif, Khairy (2011). *Principles of Measurement and Assessment*, Contemporary Educational Library, 3rd Edition.

Ma'rouf, Dima (2013). *The effect of arranging the test items according to their difficulty level on the psychometric characteristics of the intelligence test that is not influenced by culture: a quasi-experimental study on samples from students in basic and high education in public schools in Damascus governorate*. Unpublished master's thesis, Damascus University.

Melhem, Sami Muhammad. (2005). *Measurement and assessment in education and psychology*. 3rd ed, Dar Al Masirah, Amman, Jordan.

Moss, P. A. (1995). Themes and variation in validity theory. *Educational measurement: Issues and Practice, 14*(4), 5-13. https://doi.org/10.1111/j.1745-3992.1995.tb00854.x

Odeh, Ahmed (2010). Measurement and assessment in the teaching process. 4th ed., Dar Al-Amal. Irbid, Jordan.

Schultz, K. S., & Whiteny, D. J. (2005). *Measurement theory in action*: case studies and exercises. Thousand Oaks, CA: Sage.

Takno, Y., Tsunoda, S., & Muraki, M. (2015). Mathematical optimization models for nonparametric item response theory. *Information science and applied mathematics*, 23, 1-16.

van der Linden, W. J. (2005). Item response Theory. In K. Kempf-Leard (ED), *Encyclopedia in social measurement, 2*, 379-387, San Diego, CA: Academic Press. https://doi.org/10.1016/B0-12-369398-5/00452-7

Za'al, Iman. (2010). *The effect of test anxiety and the arrangement of test items according to their difficulty degree on the achievement of ninth-grade school students in mathematics*. Unpublished master's thesis, College of Education, Yarmouk University, Jordan.

Copyrights