The Effect of the Multiple-Choice Test Length on Estimating the Item Parameters and the Test Information Function According to the Three-Parameter Logistic Model in the Item Response Theory

Dr. Habis Sa'ad Al-Zboon
Associate Professor of Measurement and Evaluation
Al - Hussein Bin Talal University - Faculty of Educational Sciences

Abstract

This study aimed at identifying the effect of the multiple-choice test length on estimating the item parameters and the test information function according to the three-parameter logistic model in the item response theory. To achieve the objectives of the study, the researcher adopted (WINGEN3) program to generate data. The items parameters (difficulty, discrimination and guessing) were generated for three forms of tests consisting of (20, 40, 60) items and the resulting items parameters were used to generate abilities for (4000) examinees using the three-parameter logistic model.The (BILOG-MG3) program was also used to finding items parameters and information function of the three test forms.The study results showed that there are statistically significant differences at the level of ($\alpha = 0.05$) between the arithmetic averages of the items parameters estimates attributed to the test length and in favor of the longest test in the number of items. The results also found that the test information function increases with the increase in the number of items.

Keywords: the test length, the information function, the three-parameter logistic model
DOI: 10.7176/JEP/11-24-05
Publication date: August 31st 2020

Introduction:

Tests are considered one of the most important means used in measuring and evaluating students' abilities in order to know their level of achievement; the level of achieving behavioral goals or educational outcomes; and to know the various teaching activities provided by the teacher that help raise students' achievement competencies. Therefore, the educational supervisors made sure that these tests are highly efficient in the process of measurement and evaluation (Al-Najjar, 2010).

Tests vary according to the objectives for which they are set. What works for one topic may not work for another; what works for one goal may not work for another; and what works for one level of students may not work for another level. Thus, teacher must choose the right type. The most important of these types is the Achievement Tests which aim to measure students' achievement (Al-Kilani and Adas, 1986).

The question facing the test maker is: What is the most appropriate length for the test? Does the number of test items play a key role in determining the psychometric characteristics of the items and thus the level of students' performance on the test, which leads to achieving the accuracy of the measurement process?

The length of the test or the number of questions is considered one of the main factors influencing the psychometric characteristics of the test and its items. The reliability factor is expected to increase as the number of test items increases. Thus, the longer the test is, the more reliable it is, provided that the added questions are homogeneous with the original test questions. Moreover, the length of the test provides an opportunity to cover the content of the topic and the objectives of teaching and thus increases the validity, which in turn increases the coefficient of reliability of the test (Gabrenya & Arkin, 1980).

Psychological and educational measurement distinguishes between two main approaches in designing tests, measures and analyzing of the data derived from them, namely:

a. The traditional approach of Classical Test Theory (CTT) and its concepts and principles, some of which relate to the characteristics of test items: difficulty and discrimination; and others relate to the characteristics of a good test as a whole: validity, reliability, and standards (Allam, 2001).
b. The contemporary approach of the Item Response Theory (IRT), Modern Measurement Theory or Response Theory of the Item as some measurement scientists - such as Lord and Hambleton - call it, due to its interest in linking the individual's response to the test item with the characteristics of this item (Allam, 2005).

In the light of the researcher's review of previous studies and to the best of his knowledge, he didn't find researches that tried to study the effect of the length of multiple-choice test on estimating item parameters and test information function according to the Three-Parameter Logistic Model in Item Response Theory, as most of the previous studies dealt with the issue according to the classical theory in measurement. Hence came the need for this study, whose problem is summarized as follows:

Study Problem and Questions:

Achievement tests are one of the main components of the learning process. Interest in these tests has a direct impact on the other components of this process: objectives, content, methods, and activities, and thus it has also a direct impact on the student who is the focus of the educational process and its primary goal.

Moreover, achievement tests are the most widely used means of measuring students' achievement in all levels of education. Who prepares this type of tests seeks to provide all the conditions that make the measurement more accurate and objective, especially with regard to the number of test items. The number of test items should commensurate with the test answering time and achieve the test content validity.

Therefore, this study came to identify the effect of the length of multiple-choice test on estimating item parameters and test information function according to the Three-Parameter Logistic Model in Item Response Theory. Specifically, the study will answer the following questions:

1. Are there any statistically significant differences between the items difficulty coefficients of the test attributed to the test length according to the Three-parameter Model in Item Response Theory?
2. Are there any statistically significant differences between the items discrimination coefficients of the test attributed to the test length according to the Three-parameter Model in Item Response Theory?
3. Are there any statistically significant differences between the items guessing coefficients of the test attributed to the test length according to the Three-parameter Model in Item Response Theory?
4. Do the estimates of the test information function vary depending on the test length according to the Three-parameter Model in the Item Response Theory?

Definition of Terms:

- Item Parameters: parameter of difficulty, parameter of discrimination, and parameter of guessing.
- Parameter of difficulty (threshold): The ability level that matches the probability of 0.5 to answering the item correctly when the guessing coefficient is equal to zero.
- Parameter of discrimination (slope): The slope ratio of the item characteristic curve matching the point at which the ability mark is equal to the item difficulty.
- Parameters of discrimination (asymptote): The lower asymptote of the item characteristic curve which represents the probability that the examinee with low ability will answer the item correctly.
- Test information function: The sum of the items' functions that make up the test.

Study Objectives:

This study aims to determine whether there are statistically significant differences between the test items coefficients of difficulty, distinction and guessing attributed to the test length according to the Three-parameter Model in Item Response Theory. It also aims to find out whether the estimates of the test information function vary depending on the test length according to the Three-parameter Model in the Item Response Theory.

The Importance of the Study:

The importance of this study stems from the use of different statistical methods based on modern theory in measurement to estimate the item parameters and information function. Furthermore, the study contributes to the development of scientific research in this context, as well as giving an indication to the test makers to be concerned about the appropriate length of the test when preparing achievement tests.

Theoretical Framework and Previous Studies:

The use of classical theory in measurement has long been popular in estimating individuals' abilities expressed in true score and is defined as the raw score obtained by applying the test to the examinee several times with new elements and under different conditions. As the average raw scores represents the closest unbiased estimate of the examinee's ability or his true score.

Since the measurement process is the result of an interaction between objects properties and measuring instruments, and since the measurement results must be independent of the measuring instrument used (Allam, 2002), the dependence of the phenomena measured according to the classical theory of measurement on the characteristics of the examinees sample and the instrument used on this sample limited achieving accurate and objective measurement results. Therefore, Measurement specialists directed their efforts to find a more objective measurement system in order to obtain measurement process values that more accurately reflect what the individual possesses of the parameter measured (individual capacity), which led to the Item Response Theory (Croker & Algina, 1986).

The Item Response Theory is based on the assumption that there is a relationship between the amount of parameter possessed by the examinee (θ), the item difficulty coefficient (bi) and the probability that the examinee will obtain the correct answer at a given ability level $p(\theta)$. This relationship takes the form of the mathematical curve which is assumed to have the form of the letter (S). The examinees' abilities and the items difficulty are scaled on a single continuum. What determines the probability of the correct or incorrect response on the item is the difference between the individual's ability (θ) and the item difficulty (bi).

This theory has a set of features indicated by Hambleton & Swaminathan (1993) such as:

1. Item parameters are independent of the group of examinees used from the population of examinees for whom the test was designed.
2. Examinee ability estimates are independent of the particular choice of test items used from the population of items which were calibrated.
3. Precision of ability estimates are known.

The most important characteristic of this theory is objectivity in the psycho-educational measurement, which enables us to obtain item statistics that depend on the characteristics of the examinees, and scores that reflect the ability of the examinees independent of the item characteristics. The concept of test information function is one of the basic concepts in the Item Response Theory that makes measurement or required information as accurate as possible and through which the standard error of estimation can be determined.

When the maximum probability estimation of the ability parameter is extracted, the standard error in estimating ability is equal to the inverse of the test information function, which is given by the following mathematical relationship:

$SE(\theta) = 1/ I(\theta)$

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

Hambleton and Swaminathan (1985) indicated that the item information function is a relation that shows how the item contributes to the determination of ability. In general, high-discrimination items contribute more strongly to the accuracy of the value of their difficulty measurement (bi) on the ability continuum. Hambleton and Swaminathan (1985) also indicated that an easy test is expected to provide more accurate estimates at low ability levels, and a difficult test is expected to provide more accurate information at higher ability levels so, it is more useful for high-capacity examinees.

Item information function is used to select items when constructing tests, based on modern measurement theory, assuming that item information changes across different parameter levels; therefore, it is possible to select items that provide high measurement accuracy at a given point on the parameter continuum. Items with large distinction parameters provide greater information on the ability of the examinees; thus, providing greater accuracy. The test information function is given by the following equation:

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

Where:

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

θ: Examinee ability

$P_i(\theta)$:the probability that randomly selected examine with ability θ answers item i correctly.

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

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

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

There is a number of important steps that can be relied upon when constructing a test:

1. Describing the form of the target information function in the test.
2. Selecting items whose information functions cover levels of difficulty corresponding to the target information function.
3. After each item is added to the test, the test information function is calculated for the test items being tested.
4. Continue selecting test items until the test information function approaches the target information function.

The information function is influenced by the item parameters. In the one-parameter logistic model and the two-parameter logistic model, the highest amount of information is at the item difficulty parameter (when the ability is equal to the difficulty), because the shape of the item information function is generally close to the bell shape. Whereas in the case of the three-parameter logistic model the effect of item parameters on the test information function can be seen by simplifying the mathematical equation of the test information function as follows (Hambleton & Swaminathan, 1993):

$$I_i(\theta) = \frac{2.89 a_i^2 (1 - c_i)}{(c_i + e^{(1.7 a_i(\theta - b_i))})(1 + e^{(1.7 a_i(\theta - b_i)})^2}$$

Where :

b_i: item difficulty parameter , a_i: item discrimination parameter , c_i item guessing parameter.

Based on the above mathematical equation, the following can be observed:

1. The amount of information is generally larger when the discrimination parameter is large.
2. The amount of information increases as the guessing parameter approaches zero.
3. The amount of information when (θ) is close to (b_i) is greater than the amount of information when (θ) is far from (b_i).

Several studies have been conducted on the effect of the multiple-choice test length on estimating the item parameters and the test information function according to the three-parameter logistic model in the item response theory. One of those studies by Ababneh (2004) entitled "The effect of sample size, method of sampling, number of items and method of item selection on the accuracy of estimating item and ability parameters for a mental test utilizing item response theory". To achieve the objectives of the study, a mental ability test was prepared consisting of four sub-tests: vocabulary test, synonyms test, antonyms test, and arithmetic test. The number of test items was 71. The test items were applied to a sample consisting of 1000 male and female seventh grade students in public schools. The Biolog 3.11 program was used for estimating the examinee's abilities, test items parameters, standard errors in estimation, and contextual matching statistics for the three-parameter logistic model adopted for the purposes of the study. The study concluded that the accuracy of estimating the difficulty and ability parameters increases when the ability of the examinee is consistent with the difficulty of the items. The estimate of the discrimination parameter is also more accurate when using a sample of low-ability individuals to calibrate the items. The results also showed that estimates of the item parameters for a selected sample of items are relatively stable when calibrated within the scale as a whole or independently using the same sample of individuals. Moreover, the results showed that the accuracy in estimating the item parameters increases with increasing the sample size of the examinees.

(Al-Thiabat, 2007) aimed at determining the effect of the test length on estimating the test reliability. The results of the study showed differences between the estimates of the test reliability attributed to the change in the test length. The results also showed that the estimates of reliability increase with the increase in the test length, where the overall test had the highest estimate of the reliability coefficient (0.96), and the least test length had the lowest estimate of the reliability coefficient (0.83).

In a study by Fitzpatrick (2009) entitled "the effect of placement tests on students' proficiency rate" which aimed to study the effect of reducing the placement tests length on students' proficiency rate by relying on (12) multiple choice tests. The tests consisted of (15, 10, and 5) items. After applying those tests, the analysis was based on the one-parameter model. The results of the study indicated that there was an increase in variations in the students' proficiency rates when using the short tests. The study recommended using tests containing more than (15) items to increase stability in estimating item parameters.

(Al-Zboon, 2013) aimed to investigate the effect of the sample size on estimating the test information function and the standard error in its estimation using the modern theory of measurement. For the purposes of the study, the responses of (7500) eighth-grade students on the National Test for Controlling the Quality of Education for mathematics were used. The test consisted of (40) multiple choice items. The students were randomly selected and distributed over five groups (500 students for the first sample, 1000 students for the second sample, 1500 students for the third sample, 2000 students for the fourth sample and 2500 students for the fifth sample). By using the Bilog-mg3 and the SPSS to analyze the responses of the individuals, the item parameters were found based on the modern theory of measurement according to the three-parameter model. The estimate of the information function and the standard error in estimating the information function for five test forms were also found. The results showed that the estimates of the information function vary with the variation of the sample size, as it increases with increasing the sample size. The results also showed that the standard error in estimating the information function varies by changing the sample size, as it decreases with increasing the sample size.

(Jubran, 2017) also aimed to detect the impact of the sample size, the test length, the items form and their interactions on the accuracy of estimating the item parameters (difficulty and discrimination), the individuals' ability, and the test information function. The study sample consisted of (1015) students. The researcher developed an achievement test consisting of (75) items divided into two tests: a long test consisting of (50) items and a short test consisting of (25) items of the types of true and false and multiple choice. The results of the study showed that the estimate of the difficulty coefficient is more accurate when using a long multiple-choice test on a large sample size. The results also showed that the estimate of the discrimination parameter is more accurate when using a multiple-choice test also but on a small sample size; and was found that the estimate of the information function is more accurate when using a long true and false test on a small sample.

(Al-Sarayrah, 2017) aimed at comparing the marginal method with the relational method in the calculation of item discrimination in the light of variation in test length and sample size. The study relied on a criterionreferenced test consisting of (35) multiple choice items. Three randomly selected sample sizes (150, 100, and 50) as well as three test lengths (35, 25, 15) were used. The results indicated that there was no significant effect in the calculation of discrimination in the light of the variation in test length and sample size according to the laws of classical theory of measurement for each method.

To answer the study questions, the researcher carried out the following procedures:
The WINGEN3 Data Generation Program designed and produced by Han & Hambleton (2007) was used to generate data. This program can generate responses to unidimensional tests, whether dichotomous or polytomous, in addition to a multidimensional test. It can also generate responses to more than one group of examinees with different types of distributions, as it uses three distributions to generate individuals' ability, namely: Normal Distribution, Uniform Distribution, and Beta Distribution through which skewed distributions - whether positively or negatively skewed - can be generated.

Stages of Data Generation:

First: Test Generation

1. Generating three test forms of different lengths (60, 40, 50) items according to the three-parameter model in the item response theory.
2. Generating the items discrimination parameter based on the log normal ~ (0,0.25) distribution according to the three-parameter model. After generating the data, it was found that the arithmetic mean of the discrimination parameter values were (1.09, 1.01, 1.04) and the standard deviation values were (0.22, 0.25, 0.27). These values were considered good compared with the standard set by (Hambleton & Swaminathan, 1985), which states that the values of the true discrimination parameter range around (0.2) logs.
3. Generating the items difficulty parameter based on to the normal distribution ~ (0,1) according to the item response theory models where the arithmetic mean of the difficulty parameter values was (-0.36, -0.12, 0.15) and the standard deviation was (0.74, 1.096, and 0.92) respectively.
4. Generating the items guessing parameter based on the Beta ~ (8.32) distribution according to the three-parameter model. This distribution produces values for the guessing parameter similar to the guessing values of the objective test (binary response) that has five alternatives. The arithmetic mean of the guessing parameter was (0.2, 0.22, 0.23) and the standard deviation was (0.07, 0.06, 0.06), respectively.

Second: Responses Generation

The responses of (4000) examinees were generated using the same values of the real items parameters that were previously generated based on the normal distribution of Normal ~ (0,1).

Data Analysis:

To achieve the objectives of the study, the researcher adopted (WINGEN3) program to generate data. The items parameters (difficulty, discrimination and guessing) were generated for three forms of tests and the resulting items parameters were used to generate abilities for (4000) examinees using the three-parameter logistic model. The information function of the three test forms was then calculated according to the three-parameter model in the item response theory.

Matching Data with the Model

The program (BILOG-MG3) was adopted in order to match individuals and items with the models of the item response theory. The data of (4000) examinees were analyzed. The results of the analysis indicated that all items were matching with the model, as the value of their chi-squared test was not statistically significant at the level of significance ($\alpha = 0.05$). The results of the analysis also showed that all the responses of the examinees were in line with the expectations of the models except eight individuals where their value of chi-squared test was statistically significant at the level of significance ($\alpha = 0.05$).

Statistical Processing

The following statistical processes were used to answer the study questions.

1. The (WINGEN 3) program was used to generate three test forms consisting of (60, 40, 20) items and to generate items parameters according to models of the item response theory.
2. The (WINGEN 3) program was also used to generate responses of (4000) examinees based on the parameters of the item generated.
3. The (BILOG-MG3) program was used to find the items parameters and information function of the items of the three test forms according to the three-parameter model in the item response theory.

Results and Discussion:

Question (1): Are there any statistically significant differences between the items difficulty coefficients of the test attributed to the test length according to the three-parameter model?
To answer this question, one-way ANOVA was used to analyze the items difficulty coefficients of the test

according to the test length (40,20,60). Table (1) shows these results:

Table 1: one-way ANOVA of items difficulty coefficients attributed to the test length

Source of variance	Sum of squares	DF	Means squares	F	Sig
Between groups	6.07	2	3.03	2.53	0.03
Within groups	140.54	117	1.20		
Total	146.6	119			

It is clear from table (1) that there are statistically significant differences between items difficulty coefficients of the test attributed to the test length according to the three-parameter model.

To determine the direction of the differences, and which of the test forms these differences belong to, Scheffe test for post-hoc comparisons was used. Table (2) shows the results of the post-hoc comparisons.

Table 2 : multiple comparisons between the mean of the items difficulty coefficients for three forms of the test

Test length		**Mean Difference**	**Std. Error**	**sig**
60 items	**20 items**	***0.32**	**0.30**	**0.00**
	40 items	***0.61**	**0.28**	**0.00**
20 items	**60 items**	***0.33**	**0.22**	**0.02**

It is clear from Table (2) that there are statistically significant differences in the arithmetic means of discrimination coefficient of the items attributed to the difference in the number of the test items between the two models. This difference is in favor of the model that has the highest number of items. Perhaps this is due to the fact that the more items the test has the more content is represented, which achieves content validity. Moreover, the test items in this case will gradually move from easy to more difficult, which minimizes guessing, since the items with high variation have a high discrimination and the difficulty coefficient which gives the highest variation is the best.

Question (2): Are there any statistically significant differences between the items discrimination coefficients of the test attributed to the test length according to the three-parameter model?

To answer this question, one-way ANOVA was used to analyze the items discrimination coefficients of the test according to the test length (40,20,60). Table (3) shows these results:

Table 3: one-way ANOVA of items discrimination coefficients attributed to the test length

Source of variance	Sum of squares	DF	Means squares	F	Sig
Between groups	0.006	2	0.003	0.110	0.00
Within groups	2.99	117	0.03		
Total	3.00	119			

Table (3) shows that there are statistically significant differences between the items discrimination coefficients of the test attributed to the test length according to the three-parameter model.

To determine the direction of the differences, and which of the test forms these differences belong to, Scheffe test for post-hoc comparisons was used. Table (4) shows the results of the post-hoc comparisons.

Table 4 : multiple comparisons between the mean of the items discrimination coefficients for three forms of the test

Test length		**Mean Difference**	**Std. Error**	**sig**
60 items	**20 items**	***0.17**	**0.44**	**0.01**
	40 items	***0.41**	**0.51**	**0.03**
20 items	**60 items**	***-0.39**	**0.32**	**0.00**

It is clear from Table (4) that there are statistically significant differences in the arithmetic means of the discrimination coefficients of the items attributed to the difference in the number of the test form items in favor of the test form that has the highest number of items. Perhaps the reason for this is that the model with the highest number of items is better as it will show the real level of the student. Moreover, in this case, the items will distinguish between students with low achievement and students with high achievement, and the time given for answering will be appropriate, which prevents students from guessing.

Question (3): Are there any statistically significant differences between the items guessing coefficients of the test attributed to the test length according to the three-parameter model?

To answer this question, one-way ANOVA was used to analyze the items guessing coefficients of the test according to the test length (40,20,60). Table (5) shows these results:

Table 5: one-way ANOVA of items guessing coefficients attributed to the test length

Source of variance	Sum of squares	DF	Means squares	F	Sig
Between groups	0.027	2	0.013	2.02	0.015
Within groups	0.72	117	0.006		
Total	0.74	119			

It is clear from table (5) that there are statistically significant differences between the items guessing coefficients of the test attributed to the test length according to the three-parameter model.

To determine the direction of the differences, and which of the test forms these differences belong to, Scheffe test for post-hoc comparisons was used. Table (6) shows the results of the post-hoc comparisons

Table 6 : multiple comparisons between the mean of the items guessing coefficients for three forms of the test

Test length		Mean Difference	Std. Error	sig
60 items	**20 items**	***0.35**	**0.21**	**0.01**
	40 items	***0.54**	**0.28**	**0.00**
20 items	**60 items**	***-0.61**	**0.12**	**0.02**

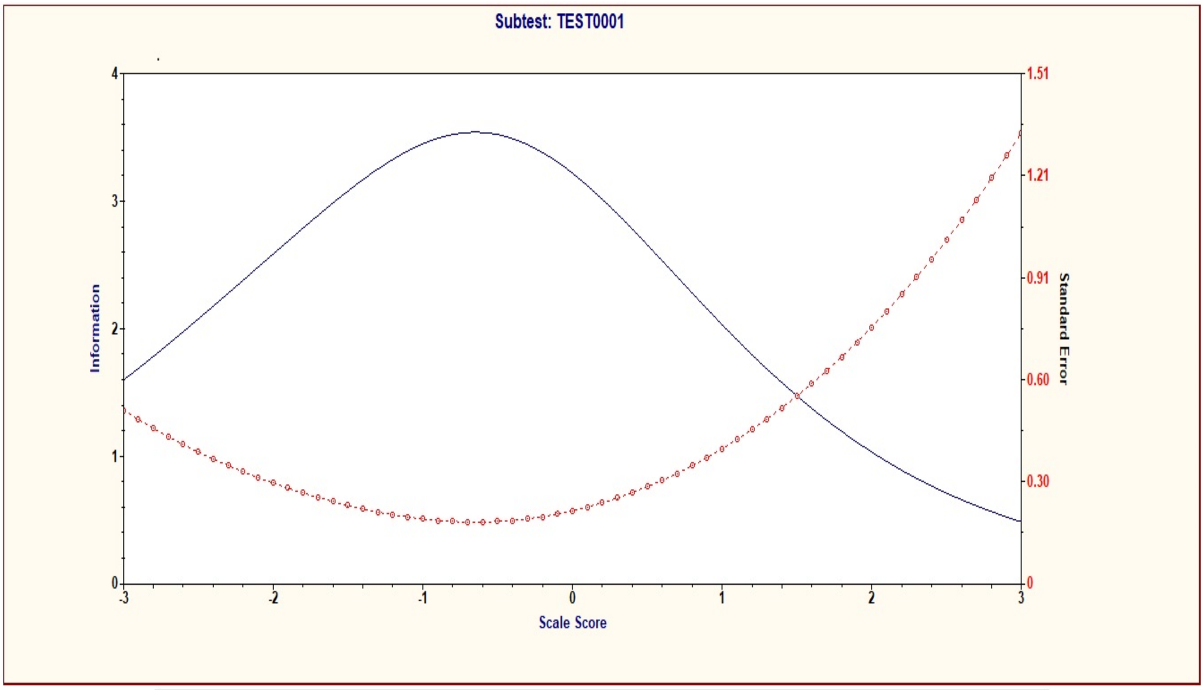
It is clear from Table (6) that there are statistically significant differences in the arithmetic means of the guessing coefficients of the items attributed to the difference in the number of the test form items. These differences are in favor of the test form that has the highest number of items; since the more items the test has, the more difficult the test will be. In this case, the test items will cover most of the subject and the time given for answering will be appropriate, which prevents students from guessing.

Question (4): Do the estimates of the test information function vary depending on the test length according to the three-parameter model in the Item Response Theory?

To answer this question, the information function was estimated for all test items and for the test as a whole based on the lengths of the test (40,20, 60). Table (7) shows the value of the information function for each of the three test forms.

Table 7 : information function for each of the three test forms

Test length	Information function
20	3.7
40	5.1
60	8.3



Test information curve: solid line
Standard error curve: dotted line

The total test information for a specific scale score is read from the left vertical axis.

The standard error for a specific scale score is read from the right vertical axis.

Figure 1 : information function of test length (20) items

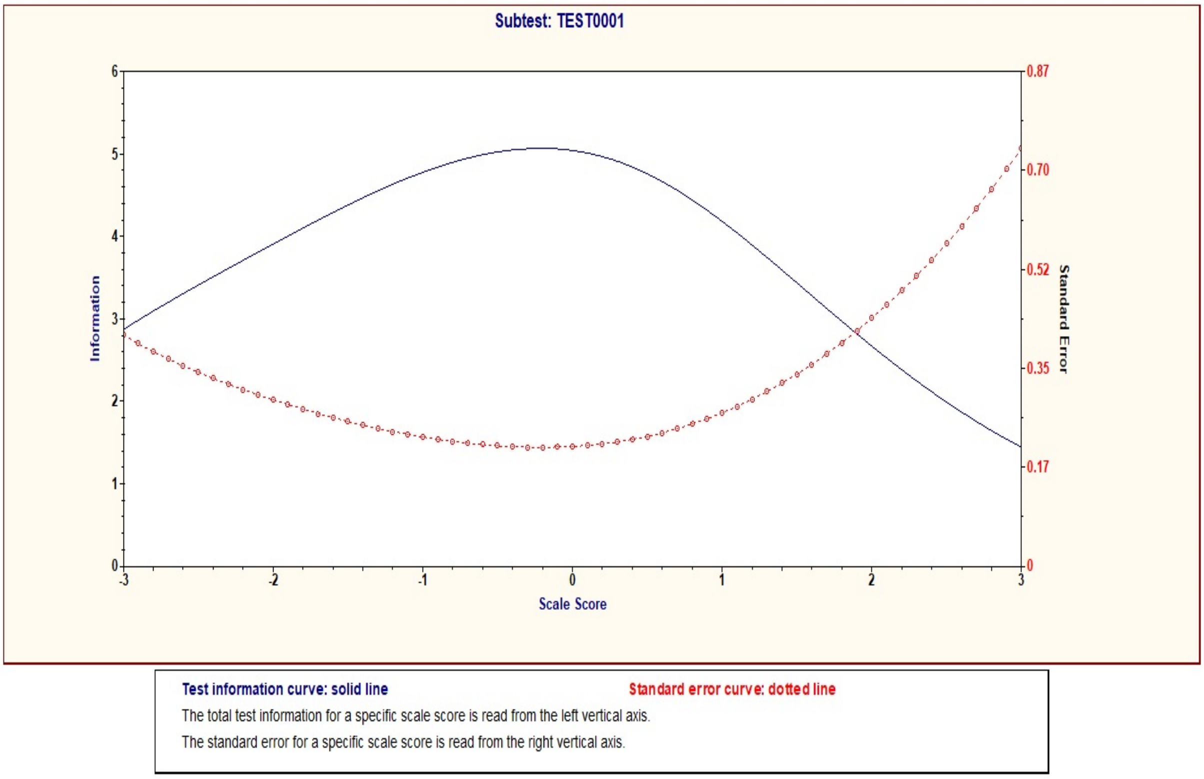


Figure 2 : information function of test length (40) items

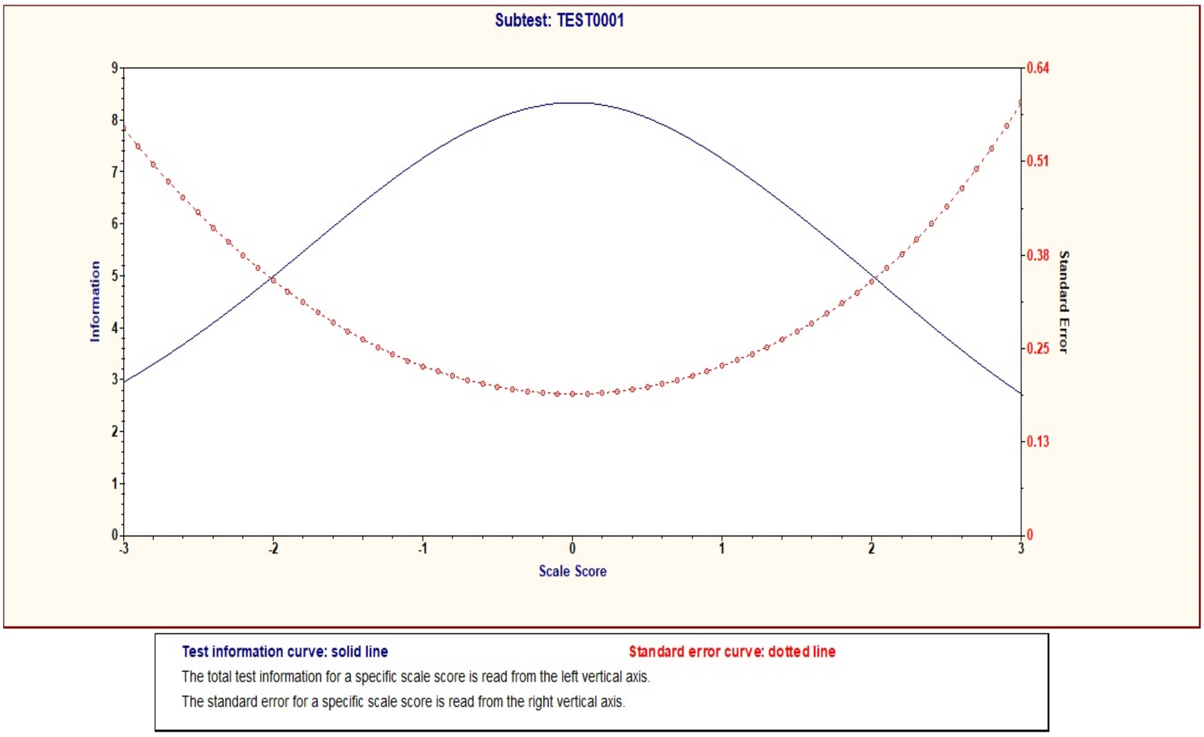


Figure 3 : information function of test length (60) items

Table (7) and Figure (1,2,3) show that the estimates of the information function vary depending on the length of the test. The amount of the information function increased from (3.7) when the test length was (20) items to (5.1) and (8.3) when the number of the test items was increased to (40) and (60) respectively. This increase in the information function by increasing the number of the test items may be due to the increase in content representation. Otherwise, because by increasing the number of the test items the guessing parameter average approaches zero, which results in an increase in the test information function as the amount of information increases when the value of guessing decreases. Moreover, this may be because when relying on the three-parameter model, the guessing parameter is taken into account, which reduces the effect of guessing and thus increases the test information function. It may also be the case that when the number of test items increase their difficulty level will gradually increase, which gives the student a greater chance to answer, and this in turn, increases the test information. Also, as the number of test items increases, it is expected that the discrimination parameter of the test items will increase, thus, increasing the value of the test information function.

Recommendations

1. Reliance on tests with number of items proportionate to the test period.
2. Conducting further studies based on the Rasch Model and the Two-parameter Model.

References

Ababneh, I. (2004). *The effect of sample size, method of sampling, number of items and method of item selection on the accuracy of estimating item and ability parameters for a mental test utilizing item response theory*. Unpublished doctoral thesis. Amman Arab University, Jordan.

Al-Zboon, H (2013). *The Effect of Sample in Estimating the Information Function of The Test and its Estimating Standard Error depending on The Modern Test Theory of Measurement.* An-Najah University Journal for Research, Vol. 27 (6), pp. 1323 - 1344.

Al-Sarayrah, I (2017). *Comparison of marginal method and relational methods in calculating the highlight the item in the light of the varying length of the test and sample size.* Unpublished Master Thesis. Mutah University, Jordan.

Allam, S (2001). *Diagnostic tests reference test in the educational, psychological and training fields*. 2nd ed., Cairo: Dar Al-Fikr Al-Arabi.

Allam, S (2002). *Educational and Psychological Measurement and Evaluation: Its Fundamentals, Applications and Contemporary Trends*, Cairo: Dar Al-Fikr Al-Arabi.

Allam, S (2005). *Single-dimensional, multi-dimensional test response models and their application in psychometric and educational measurement*. 1st ed., Cairo: Dar Al-Fikr Al-Arabi.

Al-Kilani, A & Adas, A (1986). *Measurement and Evaluation in Education and Psychology.* Jordan Book Center, Amman, Jordan.

Al-Najjar, N (2010). *Measurement and Evaluation: An Applied Perspective with SPSS Software Applications*. Jordan, Amman: Dar Al-Hamed for Publishing and Distribution.

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

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

Fitzpatrick , Ann . R. (2009). The Impact of Anchor Test Configuration on Student Proficiency Rates. *Educational Measurement*: Issues and Practice, v27 n4 p34-40 Win 2008.

Gabrenya, W.K ,J R ,& Arkin , R, M.(1980). Self – monitoring scale : Factor structure and correlates, *Personality and social psychology* . 6 , 13-22.

Gibran, N (2017). *The impact of sample size, test length, and item format on the accuracy of estimating the item parameters, ability, and test information function*. Unpublished Doctoral Thesis. University of Jordan, Jordan.

Hambleton , R.H & Jones , R , W. (1994*). Item Parameter estimation errors and their Influence on test information function*. Applied Measurement in education. 7 (3), 171 - 186.

Hambleton,R.k., & Swaminathan,H, &Rogers. H.j.(1993). *Fundamentals of Item Response Theory :*International Educational and Professional. Publisher Newbury park

Hambleton,R.k., & Swaminathan, H.(1985). *Item Response Theory :Principles and applications*. Boston MA:Kluwer-Nyjhoff

Thiabat, L (2007). *The Effect of Test Length on the Characteristics of True Scores Distribution According to the Three - parameter Model.* Unpublished Master Thesis. Yarmouk University, Jordan.