Research Article

Received: 17 September 2021 / Accepted: 24 November 2021 / Published: 3 January 2022

The Effect of Test Length on the Accuracy of Estimating Ability Parameter in the Two- and Three-Parameter Logistic Models: Comparison by Using the Bayesian Method of Expected Prior Mode and Maximum Likelihood Estimation

Eisa Abdul-Wahhab Al-Tarawnah

Measurement and Statistics,
PO Box 2015, Amman 11181, Oman, Jordan

Mariam Al-Qahtani

Measurement and Statistics,
Ministry of Education,
Ahmadi City, Kuwait
Corresponding author

DOI: https://doi.org/10.36941/jesr-2022-0015

Abstract

This study aims to compare the effect of test length on the degree of ability parameter estimation in the two-parameter and three-parameter logistic models, using the Bayesian method of expected prior mode and maximum likelihood. The experimental approach is followed, using the Monte Carlo method of simulation. The study population consists of all subjects with the specified ability level. The study includes random samples of subjects and of items. Results reveal that estimation accuracy of the ability parameter in the two-parameter logistic model according to the maximum likelihood method and the Bayesian method increases with the increase in the number of test items. Results also show that with long and average length tests, the effectiveness is related to the maximum likelihood method and to all conditions of the sample size, whereas in short tests, the Bayesian method of prior mode outperformed in all conditions. Results indicate that the increase of the ability parameter in the three-parameter logistic model increases with the increase of test items number. The Bayesian method outperforms with respect to the accuracy of estimation at all conditions of the sample size, whereas in long tests the maximum likelihood method outperforms at all different conditions.

Keywords: *Maximum likelihood; Bayesian expected prior mode; the two-parameter logistic model; the three-parameter logistic model*

1. Introduction

Educational and psychological testing theories have played a significant role in developing education. These theories have established a framework, based on scientific and statistical hypotheses, to process different testing issues in a way that would help scholars and teachers handle numerous educational issues and ensure that reliable and precise statistics about the testing process are obtained. Consequently, sound educational decisions can duly be made about the students and the effectiveness of the program of the schools. Scientists and researchers wasted no time in improving the testing process by relating it to subjects' psychological features, such as personality, behavior, and academic achievement.

As in all of the sciences, educational and psychological measurement has gone through various phases of development to include psychological and educational tests. These tests have been based on different theories and principles; among those theories is the Classical Test Theory (CTT), which remained the norm in the psychological and educational field for a long time. It was used in a variety of testing situations, including psychological and educational tests. The theory has its defects, however, which results in narrowing its scope in important examinations and testing processes (Abu-Allam, 2005; Janssen et al., 2014). Specifically, the major deficiency in CTT is the dependence of the characteristics of items on those of the testees and vice versa. This means that the separate features (ability coefficients) depend on the testee sample used in its evaluation; the testees' features also depend on the items sampled in its evaluation. Because of these weaknesses in CTT, researchers developed a new theory called Item Response Theory (IRT), which had been called Latent Trait Theory (LTT). This theory comprises several models called item response models (Philip & Ojo, 2017). Lord (1953) introduced the principles of this theory, which amends the weaknesses in the classical theory by introducing a method of terminology that features fixed item parameters. These include the difficulty parameter, discrimination, and guessing, in addition to standardizing measurements of difficulty and testees' ability so that test developers can select the most suitable items to classify the testees on the basis of performance levels (Sulaiman, 2009); Costa and Ferrão (Costa & Ferrão, 2015). In the item response theory, the first step is specifying the item parameter and the ability parameter. Specifying these estimations defines the features of these models. The success of specifying this estimation depends on the availability of suitable procedures to estimate these parameters. In fact, there are many methods to specify item parameters and ability parameters. Such methods make use of numeric analysis, which can be implemented by using computer programs (e.g, BILOG-MG3, MULTILOG, and SYSTAT). Baker and Kim (2017) have stated that specifying the testee's ability is the principal goal of the test. The ability traits in this testing process can be personal or mental and are estimated in light of the IRT according to his/her response in the test. There are two methods for estimating the ability of the testee. The first is maximum likelihood estimation, which entails calculating the likelihood at more than one value at (θ) and (θ) of the maximum likelihood, which is the ability of the testee. The second method is Bayesian estimation, which is based on the supposition that there is a previous ability distribution of the testees' abilities in their society, and this distribution is mostly normal (Kim et al., 2015). Thissen and Wainer (1982) have pointed out that it is difficult to define the amount of error in estimating parameters with IRT. This error shows the precision in estimating the used model's parameters, because precision indicates the limit of the agreement between decisions based on the test results and decisions made when the results show no testing mistakes. Assuming that a test having no error is practically impossible, this would be how precision must be measured (Azizan et al., 2019). In the modern theory (IRT), scholars and researchers in the field of psychological and educational testing have tried to find factors that might affect the estimation precision (Finch & Edwards, 2016). This study is also an attempt to discover ability estimation precision in both methods: maximum likelihood estimation and Bayesian estimation in the three-parameter logistic model in the following variables: test length, test items difficulty coefficient, and sample size.

1.1 Problem and Questions of the Study

The success of ITR implementation basically depends on accurate selection of the suitable mode of ability estimation. Inaccurate selection affects three different aspects: selection of items used, finalizing the test, and the final estimation of the trait. Applications of IRT depend on the accuracy of item parameter estimates and on ability parameters. The decrease of ability parameter accuracy estimates renders the results of ability estimates doubtful. This has an automatically negative effect on decisions about the subjects made by the specialists using the tests in the educational and psychological field. There are several methods used to estimate the three-parameter logistic model. Which of those methods is most effective in estimating the subject's ability (within limited conditions) is the most important question, and, to the researchers' knowledge, no studies have focused on this topic. This study therefore aims to test the ability parameter accuracy estimates with regard to the test length variable to specify situations in which it is better to use Bayesian estimation and in which it is better to use the maximum likelihood estimation. More specifically, this study attempts to answer the following questions:

1. Comparison of the effect of test length on the estimation accuracy of the ability parameter in the two-parameter logistic model using the Bayesian method of expected prior mode and maximum likelihood.
2. Comparison of the effect of test length on the estimation effect accuracy of the ability parameter in the three-parameter logistic model using the Bayesian method of expected prior mode and maximum likelihood.

1.2 Purpose of the Study

The aim of this study can be further split into the following goals:

1. Comparison of the effect of test length on the estimation accuracy of the ability parameter in the two-parameter logistic model using the Bayesian method of expected prior mode and maximum likelihood.
2. Comparison of the effect of test length on the estimation effect accuracy of the ability parameter in the three-parameter logistic model using the Bayesian method of expected prior mode and maximum likelihood.

1.3 Justifications of the Study

- This study is of great importance owing to how significant accuracy is in the estimation of parameter in IRT models. This step is essential in applying this theory and maximizing its advantages, since these estimations determine the models' characteristics.
- Also critical is the difficulty in specifying the error that reflects the estimation accuracy of the two and three parameters according to each method used in estimating the ability parameter.
- Studies in Arabic that highlight the difference between the effectiveness of the maximum likelihood estimation and the Bayesian estimation are scarce, and this study will contribute to the literature in the IRT and the two- and three-parameter logistic models.
- The study is designed to focus researchers' attention on the accuracy of parameter estimation, because there are several statistical methods that require understanding of advanced mathematics and statistical issues, particularly Bayesian statistics.

1.4 Limitations of the Study

1. Results of this study are limited to the method of simulating data, more specifically using Monte Carlo (MC) of computer simulation by using the WinGen-3 program.
2. Results of this study are limited to the statistical methods used for estimating the ability parameter: maximum likelihood method and the Bayesian method (the expected prior mode).
3. Results of this study are limited to the logistic models used, specifically the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM).

1.5 *Terminology of the Study*

- Two-parameter Logistic Model: This IRT model depends on two parameters: difficulty and discrimination in addition to estimations of testees' ability.
- Three-parameter Logistic Model: This IRT model depends on three parameters: difficulty, discrimination, and guessing in addition to estimations of testees' ability.
- Estimation of Person Parameter: The estimation of the ability parameter by a value extracted from applying a mathematical function in item response models once the testee answers the items of a particular test.
- Accuracy of Estimation: An expression that refers to quality of estimation featured by maximum likelihood that the estimation is approximate to the actual value. That can be obtained by selecting the unbiased estimation whose variance is less than any other estimation by using error squares and standard error in estimation.
- Maximum Likelihood Estimation: A statistical method used to find out the estimation of the ability parameter in the IRT. This method is based on finding the ability parameter value (θ), which signifies that likelihood function.
- Bayesian Estimation Bayesian: Estimation is a statistical method used to determine the estimation of ability parameter in IRT. This method is based on the assumption that there is a prior probable distribution that can ensure more accuracy in the process of stimulating ability.

2. Review of Literature and Previous Studies

2.1 *Item Response Theory*

Research and educational centers have been using traditional instruments of measurement for a long time; however, those instruments do have aspects of weakness and lack of precision. Those weaknesses have been shown in the problems that appear while using these classically oriented methods. They fall short of attaining objectivity in the testing field in general and more specifically in behavioral testing (O'Connor et al., 2015; Reise et al., 2005; Vispoel & Kim, 2014). Several studies have been conducted to eliminate such issues, which have resulted in the emergence of modern testing theory. This has greatly helped in creating and ensuring the desired objectivity. Modern testing theories evolved, among which are the latent trait theory and the item response theory (IRT), also called the item characteristic curve (ICC) theory (Reise & Waller, 2003; Zanon et al., 2016). The measurement method that distinguishes IRT from classical theory is that it is based on probable mathematical modeling, which adjusts confounding factors that negatively affect the comparison between students' grades. This is caused by the mathematical modeling that the testing process undergoes. Moreover, the IRT is based on new rules that differ essentially from the rules used in the classical theory (Adedoyin & Mokobi, 2013; Downing, 2003; Embretson & Reise, 2000).

In this regard, Hambleton and Swaminathan (Hambleton & Swaminathan, 1985) listed several advantages that encourage specialists in educational and psychological testing to use the IRT. Those advantages are:

- The modern theory provides a statistical index that measures degree of accuracy in measuring each person's ability.
- When many test items measure the trait, the estimation of a person's ability occurs independently of the items sample applied to it, namely, the items used in estimating subjects' ability is free from characteristics of the items (item-free).
- In the case of many people involved, the items' psychometric characteristics are estimated freely from the subjects' sample (sample-free), which is used for estimating these parameters.

Recently the IRT has attracted the attention of researchers and test designers in many tests and orientations. This theory assumes that it is possible to predict a person's ability. Their performance in an educational or psychological test can be explained in light of the distinctive properties of this performance, called traits. This theory also estimates the individual's marks/grades in these traits. A group of models has developed from this theory, known as the latent traits model (Albano, 2017; Steel & Klingsieck, 2016; Zanon et al., 2016).

2.2 *Item Response Theory Assumptions*

The IRT model determines the relationship between the observed performance of the testee on the test and the latent abilities and traits that cannot be easily observed; however, these abilities can be deduced from the student's performance on the test. The relationship between observed performance and the latent trait can be shown in a mathematical model; therefore, the IRT can be said to be a mathematical model based on strong axioms. The importance of these assumptions comes from the necessity of proving them to ensure reliable and objective findings. There are other advantages to be obtained in applying this theory (IRT). They are as follows:

2.2.1 *Unidimensionality*

The IRT presupposes the availability of many dimensional latent traits, which makes it possible to locate every testee in the space of latent traits when identifying his or her position in each trait. The traits space is considered complete when it affects all the testee's marks in the exam. It is quite common for each testee to have one latent trait or ability that gives indications about his or her performance. This means that all the test items estimate the same ability or the same latent trait (Hamid, 2008; Jin & Wang, 2014).

Crocker and Algina (2009) state that unidimensionality can be defined by using statistical correlation language among items. For the test to be unidimensional is to attribute the statistical correlation among items to only one trait. The test is said to be unidimensional if its items are statistically correlated to the whole group of testees, and it can be statistically independent when each group of testees shares the same ability (Bazaldua et al., 2017).

Practically speaking, this assumption cannot be totally confirmed because of other factors that influence the testees' performance. Such factors can be natural, personal, or knowledge-based, or they might be related to answering skills such as motivation, speed of performance, test anxiety, and test instructions, all of which have an impact on an individual's performance to some extent (Allam, 2005; Kose & Demirtasli, 2012).

2.2.2 *Local Independence*

Local independence occurs when a testee's reactions on the test items are statistically independent when the general ability is taken into account. Thus, an individual's performance in one item does not affect his or her performance in the same test. Hambleton and Linn (1989) think that for an assumption to be correct, the individual's performance in a single item must be independent of any other item. In other words, performance in an item is affected by his/her abilities and features of this very item.

For an item to be hypothetically independent, the following equation has to be correct: $P_{ij}(\theta) = P_i(\theta) \times P_j(\theta)$ (1) where $P_{ij}(\theta)$ is the probable answer of the items (*i, j*). $P_i(\theta)$ is the probable correct answer of item *i*. $P_j(\theta)$ is the probable correct answer of item j. On reconsidering the previous equation, the local independence assumption confirms that a person with a particular ability of (θ) to answer two items (*i, j*) correctly equals the response probability multiplied for each item. This equation cannot be correct unless the testee's answers to items are statistically independent of each other. Local independence can be achieved on items whose numbers (*N*) with a particular ability if: $prob(U_1 = u_1, U_2 = u_2, U_3 = u_3, \dots\dots\dots, U_n = u_n|\theta) = P_1(\theta)^{u1}Q_1(\theta)^{1-u1}P_2(\theta)^{u2}Q_2(\theta)^{1-u2} \dots\dots. P_n(\theta)^{un}Q_n(\theta)^{1-un} = \prod_{i=1}^{n} P_i(\theta)^{ui}Q_i(\theta)^{1-ui}$ (2) where *Ui (i = 1, 2, ...n)* shows the testee's responses to the test items whose number is N and the response to it is (1, 0). This equation confirms that the assumption of an item's local independence is achieved when the probable responses of any testee equal the probabilities of the testee's answer to each item multiplied. Because this condition to fulfill local independence is strong, concerned researchers started looking for a weak form for this assumption, which examines the local independence of each pair each time instead of examining them throughout. This is called the weak independence local item (*WLI*), and it can be shown in the following equation: $COv(U_i, U_j|\theta) = 0, i \neq j$ (3)

This weak assumption states that if the shared variance of any two items equals zero, it means that the local independence is achieved (Dresher, 2003; Nofer, 2007). Having surveyed the unidimensionality and local independence, we can note that specialists and researchers do not agree on the fact of their equivalence; some say that they are equivalent and some say that they are not. Opinions supporting the equivalence of these two concepts are those of Hambleton and Swaminathan (1985) and Hulin et al. (1983), who stated that achieving local independence indicates that a group of items test only one trait, and that unidimensionality ensures local independence. Conversely, Crocker and Algina (1986) and Meara et al. (2000) state that they are not equivalent, justifying that by saying that test items can be statistically independent at a particular ability level. In addition, this test estimates more than one trait. They added that a unidimensional test in all cases will have its items statistically independent but not the opposite, which means that local independence is necessary but does not ensure unidimensionality.

2.2.3 Item Characteristic Curve (ICC)

The ICC indicates that answering a multiple-choice item correctly indicates a person's ability as well as the item's features that it aims to answer. Therefore, the aim of the IRT is to reach the individual's estimated values, and the others are item-related. These values are then used to estimate the probability of the correct answer for each item of the exam through the ICC. Allam (2006) defines this curve as a mathematical function that correlates with the possibility of a student's success in answering an item or the ability that a group of items estimates. Dresher (2003) defines it as an ascending cumulated function that describes the achievement of the testee in an item and his/her ability measured in the exam. This curve represents the probabilities of the testee's correct answer in the item's different ability levels. Because the curve is ascending and cumulative, it vividly indicates that the probability of an item being answered correctly increases as the testee's ability does. Kymlicka et al. (2002) think that the ICC is a basic concept in IRT. It is the probability of a correct answer to an item as an indicator of the measured latent trait of the test's items. Most IRT applications assume that the item's distinctive curve has an S shape (Edelen & Reeve, 2007). The researchers think that however the characteristics curve changes in form, the probability of a student's answering an item correctly depends on the form of the curve. This form is independent of the testee's abilities distribution. This characteristic indicates the stability of item characteristics curves that have been standardized to a community of testees and is considered the most important distinctive feature of all latent traits models.

2.2.4 Freedom from Speediness

Modern measurement theories agree that speediness does not play a critical role in the response to testees' answers; that is, bad performance in a given test is a definite result of low abilities. This is an implied assumption for all item response models as it is implied in unidimensionality, which stipulates that the test has been applied under the condition of speediness. Thus, a student's failure to answer a test's items correctly is due to their inability and not to lack of time. In this regard, Bolt et al. (2002) emphasized that the duration of the test has to be enough for all examinees. He thinks that leaving some items unanswered negatively affects the accuracy of ability estimation, which in turn deforms the person's ability estimation. We also have exaggerated estimations of item parameters that fall at the end of the exam. The influence of speediness can be determined by counting the number of students who could not answer all test items.

2.3 Factors on the accuracy of parameter estimation

Several studies investigated the effect of certain factors on the accuracy of parameter estimation in light of IRT. Some of those studies are presented here. Ban et al. (2001) compared and evaluated methods of computerized adaptive testing. Those methods are marginal maximum likelihood of the one-cycle estimation method, marginal maximum likelihood of the two-cycle estimation method, Stocking's method A, and Stocking's method B. Samples of the sizes 3000, 1000, and 300 were used, which were simulated. Results showed that the marginal maximum likelihood of the two-cycle estimation method provides better estimations with the standard error of estimation under the various conditions of sample size. Moreover, Stocking's method B performed better, but it requires anchor items, which leads to an increase in test length. Al-Darabee (2001) investigated the effectiveness of the one-parameter logistic model (Rasch model) for accurate estimation of a person's ability and the item difficulty parameter when using sample sizes of 500, 100, and 50 persons and the length of the test is 25, 50, and 300 items. To achieve the purpose of the study, two-item response data were estimated by simulation method using the computer program IRTDATA. For estimation of a person's ability and the item difficulty parameter, these data were entered into a specialized program of estimation (BILOG 3.0). RAMSE was determined for both a person's ability parameter and for item difficulty using a program prepared by the researchers and coded in Pascal. To answer the hypotheses of the study, a two-way analysis of variance 3*3 was used. Results then indicated that there were significant differences for the reaction of both the sample size variable and length of the test with the accuracy of estimating a person's ability. Results also showed that there were significant differences in the accuracy of the estimation of a person's ability, which was related to the variable length of the test only, whereas results did not show significant differences for sample size levels. Similarly, Pelton (2002) compared the accuracy and stability of the estimating difficulty parameter and the ability parameter using the classical theory of measurement and the logistic models in IRT by using the simulation method. Results of this study revealed that ability estimations could be compared through the classical theory and IRT. Ability estimation varied according to the available information in the simulated data group, which in turn was affected by unidimensionality, estimation degree, and variance range in item difficulty compared to the subject's ability. Results also showed that if average sample size is 999 testees and appropriate item number is 33 items, then the two-parameter logistic model will provide more accurate estimations for difficulty parameter than the Rasch model, the three-parameter logistic model, and the classical theory of measurement when there is less guess in the test items answer. Si and Schumacker (2004) compared accuracy estimation of testees' abilities when employing different permutations of the latent trait models, some of which are the partial estimation model and dichotomous gradation models, in addition to various distributions of testees' abilities, some of which are the positively skewed distribution and negatively skewed distribution, and normal distribution. The researchers used the simulation model by generating answers for 1000 testees for a test of 30 items; 84 experimental formations were used for the sake of comparing estimation accuracy of testees' abilities, which would result from each experimental formation relying on root mean squared error (RMSE) standard. Study results indicated that estimation accuracy of testees' abilities using a partial estimation model is much better than estimation accuracy of testees' abilities using dichotomous gradation models regardless the distribution of testees' abilities. Al-Ta'mari (2003) applied a three-parameter logistic model to estimate a person's ability and item features to test multiple choice items with different numbers of distractors. He also compared results in light of the classical theory of measurement. The sample of the study consisted of 1200 students of the 10th grade in the Educational Directorate of Amman in Jordan. The BILOG program was used to estimate subjects' ability and item parameters according to the modern theory of measurement. One of the significant results of the study was no statistically significant difference in estimation of person's ability to take the multiple-choice test with different numbers of distractors in light of IRT. There were statistically significant differences in a person's ability according to the classical theory of measurement, favoring the test with five choices. Moreover, there were statistically significant differences in item difficulty on multiple-choice tests with varying number of choices per the classical theory of measurement, favoring the test with three choices. There were no statistically significant differences in item discrimination in the multiple-choice test with a varying number of choices according to the IRT, favoring the test with five choices. Lastly, there were statistically significant differences per the classical theory of measurement, favoring the test with four choices. Ababnah (2004) investigated the effect of sample size and its selection on estimation accuracy of item parameter and ability to test mental ability. Therefore, the researcher prepared a mental ability test of four subtests (i.e., vocabulary test, synonyms test, antonyms test, and computer test). The test in its final draft consisted of 71 items. It was distributed to a sample of 1000 students of the 7th grade in Jordan. The researcher used the BILOG 3.11 program to estimate testees' abilities, test items parameters, standard errors of estimation, and statistics of data correspondence to the two-parameter model. Results of that study demonstrated that accuracy of estimation of item parameter increases by the increase of sample size of testees. Ability estimations of testees were shown to be stable when titrated samples with large sizes are used, and accuracy of estimation of item parameter increases by the increase in the number of test items or their percentage to the total test.

Garre and Vermunt (2006) implemented a study that used the empirical curve and applied actual data to check results of the mathematical derivations related to accuracy of parameter estimations when using different prior distributions for ability parameters. Results indicated that parameter estimations that are obtained by the Bayesian method are more stable than the parameter estimations obtained by the maximum likelihood method, particularly at the estimation of abilities at the ends of the connector ability (Philipp et al., 2018; Wurpts & Geiser, 2014).

3. Item Response Theory Models

3.1 Two-Parameter Logistic Model (2TPLM)

This model was proposed by the statistician Birnbaum with a group of his colleagues at Columbia University in the United States. This model assumes that guessing has minimal influence, and that varies with regard to the degree of difficulty and degree of discrimination among examinees. What distinguishes this model from the previous one is that it allows variations of exam items regarding degree of difficulty and degree of discrimination. The equation runs as follows Allam (Abu-Allam, 2005): $P_i(\theta) = \frac{\pounds^{D_{ai(\theta-bi)}}}{1+\pounds^{D_{ai(\theta-bi)}}}$ (4)

Where: $P_i(\theta)$ is the probability of a correct answer by a randomly selected student from the ability level θ to item i; b_i is the difficulty parameter; a_i is the discrimination parameter; θ is the ability parameter; and D is the degree factor, which equals 1.7. Adding the discrimination parameter to this model overcomes defects in the one-parameter model, represented by the difficulty of finding a group of items with equal discrimination coefficients.

3.2 Three-Parameter Logistic Model (TPLM)

This model has been proposed to overcome the defects that can occur when applying one-parameter and two-parameter models, represented by the guessing factor as testees with low abilities whose percentage of correct answers is more than zero in the existence of guessing factor in objective tests that include yes/no and multiple-choice items. This model can be represented in the following mathematical equation (Al-Owidha, 2007): $P_i(\theta) = C_i + (1 - C_i)\frac{1}{1+\pounds^{-D_{ai(\theta-bi)}}}$ (5)

Where: $P_i(\theta)$ is the probability of a correct answer by a randomly selected student from the ability level θ to item i ; b_i is the difficulty parameter; a_i is the discrimination parameter; θ is the ability parameter; D is the gradation factor, which equals (1.7); and C_i is the guessing parameter. Thus, the 3PL model in the previous equation adds a third parameter, C_i which indicates that a person can answer correctly by guessing. This parameter is called lower asymptote (Allam, 2006).

3.3 Estimation of Ability

In tests, estimation of ability used to identify a subject's position in connected ability estimated in a test with a possibility of estimating that ability wrongly, which is estimated using standard error of estimation. The outcome value represents the variance between the testee's real mark and his/her observed mark. The symbol θ represents the testee's ability or the amount of the trait that he/she is in possession of. There is a certain probability of answering an item correctly at each ability level, represented by the symbol P (θ). Ability refers to the terms *achievement*, *attitude*, or *personality* (Hambleton et al., 1991). The first and most important step in using the IRT in designing tests and measurements and analyzing their items concerns the parameters that comprise a certain selected model. The effective use of any of these models depends on the availability of appropriate modes and suitable procedures to estimate these parameters, and on effective and suitable computer programs. The process of obtaining these estimations is called calibration. It is difficult to determine the estimated values manually because the available computer programs that have been updated in the last few decades implement various statistical methods to estimate the parameters of the different models (Allam, 2005). Embretson and Reise 2000) are of the opinion that all models based on the IRT in correcting the testees' responses are used to locate the testees' position in the ability connector, using testees' response pattern in addition to the estimated item parameters. Hambleton and Swaminathan (1985) describe four main steps to determine the testee's ability value measured in a particular test. They are as follows:

1. Compiling the testees' response data of one item including quite a large group.
2. Selecting one item response model after verifying that the data match the model.
3. Using a computer-based program in the estimation of ability and item.
4. Converting ability marks of appropriate measures.

It has been proved that the estimation of ability plays a vital role in logistic models and their applications in tests. Many indices to examine the precision of measurement of the ability estimation method (θ) also have been invented. These indices are indications to be referred to when preferring one ability estimation mode to another. The following are possible divisions of measurement precision indices:

A. General indices that give a brief image of the accuracy of estimation of a test as a whole, such as stability, approximation indices, accuracy, and efficiency.
B. Conditional indices that can be calculated at each point of ability calibrator, such as square root of the mean of error squared, bias, and standard error of estimating ability (Al-Zahrani, 2008).

3.4 Ability Estimation Methods

The successful application of IRT depends on precise selection of an appropriate ability estimation style, because an imprecise ability estimation style negatively affects many aspects, such as choice of items, stopping the test, and the final estimation of the trait. Generally speaking, there are two ability estimation styles: they are maximum likelihood estimation and Bayesian estimation. They will be illustrated here.

3.5 Maximum Likelihood Estimation

Frank (2009) states that this method tries to develop parameter estimation through parameter likelihood function, which is meant to be estimated as we gather enough information about the sample. This method aims to determine the ability parameter value (θ), which aims to indicate likelihood function, and more specifically if *ua=[u1a,u2a,....,una], a=1,2,...,N.* It represents the testee's response orientation (a) in a test with (n) items. In this case the maximum likelihood of the testee's marks (a) is the value (θ), which signifies the following function: $InL_a(\theta) = \sum_{j=1}^{n}\{u_{aj}In\,P_j\,(\theta) + (1-u_{aj})In(1-pj(\theta))\}$ (6) where $pj(\theta)$ is an item response function that conforms the item (j). Therefore, defining the value of (θ) necessitates finding the first derivative probability function of the parameter of the defined model parameters, then equalizing the derivative with zero, and after that answering the resulting nonlinear equation using numerical analysis.

3.6 Joint Maximum Likelihood Estimation

Using the maximum likelihood function, as shown previously, results in biased values in the three-parameters model. To avoid that, an ability values considerable scale is selected. This can be fulfilled by equalizing the mean and the standard deviation of a person or item parameters to (1, 0) consecutively (Allam, 2005). Estimation can be carried out in two steps, as Al-Zahrani (2008) and der Linden & Pashley (2009) point out:

1. Assuming ability parameter values by dividing the logarithm of the ratio among correct and incorrect responses of all testees. The resulting value is converted into standard values.
2. Item parameters are treated as pieces of information, whereas the item parameter is estimated.

These two steps must be repeated until we determine that the parameters' estimated value does not alter after two consecutive steps.

3.7 Conditional Maximum Likelihood Estimation (CML)

This pattern is used with the one-parameter model (Rasch model) and the models derived from it; however, it is not applicable with 2PL or 3PL, where item discrimination parameters are different. This tends to make the total grades inadequate to estimate the ability level of the testees because parameter estimations will depend on correctly answered items (Draxler & Alexandrowicz, 2015; Eggen, 2000). This method, the conditional maximum likelihood, is significant because of the conformity of parameter estimations; in addition, the likelihood function stipulates the number of the correct answer of testees to testing items, which means: $T = \sum_{g=1}^{n} X_g$ (7)

The grade T is statistically enough to estimate a person's ability, which means the conditioned probability of grades $X_1, X_2, X_3, \ldots\ldots X_n$ if we know that the total grade of an individual is independent of that person's ability θ. The likelihood function therefore can be taken as a conditioned probability for all testees' responses. Having elevated this function, we can obtain conditional maximum likelihood estimations of item difficulty parameters regardless of the ability parameter θ (Allam, 2005).

3.8 Marginal Maximum Likelihood Estimation (MML)

This method, developed by Bock and Aitkin (1981), handles indefinite ability level as far as the probabilities of answer patterns according to distribution of a particular community. The test data are viewed as a random sample taken from a particular community. This method is used to estimate all unidimensional models 1PL, 2PL, 3PL, as well as multidimensional models, and is effective whether the test has many or few items. Moreover, the resulting standard estimation value of mistakes is precise. BILOG-MG (designed by Mislevy and Bock in 1984; Rupp, 2003) is the program used to perform this estimation. This program is known for its strength, effectiveness, flexibility, and speed in estimating one-, two-, and three-parameter logistic models (Chaturvedi & Vyas, 2017). From the previous discussion of the maximum likelihood method and its three modes (JML, CML, and MML), the researchers implemented the MML, which the program BILOG-MG3 uses as one estimation method. They prefer using MML to JML and CML because parameter estimations resulted from quality stimulation of any specific deductive statistical mode; these include impartiality, consistency, and relative effectiveness, which are all achievable in MML. This method can estimate parameters of the model and individuals when the testee answers all the items either correctly or incorrectly. It is not a condition that the person who answers all the items either correctly or incorrectly be eliminated from the study.

3.9 Bayesian Estimation

In addition to maximum likelihood, which estimates personal ability, there are other means to estimate item response parameters, which depend on Bayes' theorem. This theory connects conditional likelihood with marginal likelihood. These methods require assuming prior probabilities of the parameters in the light of theoretical as well as empirical considerations. Regarding the availability of testees' ability, the methods based on Bayes' theorem can ensure obtaining reliable estimations of this ability (Allam, 2005). In fact, Bayesian estimations imply assumptions about parameter distributions; they are called prior distributions. Having included the prior distributions in the estimation process, it becomes improbable for the estimations to occur in less probable zones per the prior distribution. BILOG-MG3 ensures assumed prior distributions and also enables the user to start it if he wants to refer to it. The user can handle prior distributions in two ways: keeping the prior distributions unchanged in every repeated process, or updating them. Furthermore, prior distributions for definite parameters can be set to serve different purposes (Fox & Glas, 2003). Ababnah (2004) states that this method is recognized for using prior information in addition to the availability of information about the sample. No doubt, the Bayesian model is totally different from the classical method. In the classical method the parameter (θ) is indefinite, but it is a fixed value. The sample $X_1, X_2, X_3, \ldots\ldots X_n$ is usually drawn from a population identified by significance of (θ). The value of (θ) is determined on the basis of the values of the observed sample, but for the Bayesian approach (θ) is a quantity of a recognized variance through the description of a probable distribution of (θ), which is called prior distribution. Determining this probable distribution relies on the researcher's beliefs and previous experience about (θ) before collecting the data, which is why it is called prior distribution. A sample is then selected from the population expressed with the significance of (θ). Prior distribution is revised according to the information that the sample provides. The new type of distribution is called posterior distribution, and this newer probable distribution depends on Bayesian rule, which is the stem of Bayesian statistics. This method mostly assumes that the ability (θ) is distributed as per normal distribution, which means $\theta_a \sim N(0,1), a = 1,2, \ldots\ldots, N.$ Estimation of (θ) can be mathematically expressed by the Bayesian method as follows: $f(\theta_a|u) = \frac{f(u|\theta_a) f(\theta_a)}{f(u)}$ (8) θ estimation depends on a previously determined answer pattern, and $f(u)$ is a fixed value in this case, whereas f (u|θa) is a likelihood function, $f(\theta_a|u)$ is the posterior distribution of (θ_a), $f(\theta_a)$ is the prior distribution of θ_a. Here we assume that distribution is made according to the normal distribution. To simplify the definition of posterior distribution, the likelihood function $f(u|\theta_a)$ can be written as follows: $L(u|\theta_a) = L(u_1, u_2 \ldots un|\theta_{1,} \ldots, \theta_n) = \Pi_{a=1}^{N} L(u_a|\theta_a) = \Pi_{a=1}^{N} \Pi_{i=1}^{\wedge} P_{ia}^{u_i a} Q_{ia}^{1-u_i a}$ (9) Where: P_{ia} is the function of item response, n: Number of items, and N : Number of the testees.

3.10 Monte Carlo Methods of Simulating Data

Because it was difficult to implement the experimental approach in psychological and education studies, especially in the control field, technologic advancements in programs and computer devices have made experiments possible. In the field of psychology this has become a reality owing to various computer programs playing a prominent role in the advancement of research in educational psychology theories (Hadwin et al., 2005). Because of the potential and the effective solutions that Monte Carlo methods have provided, researchers can administer the highest degrees of experimental control in experiments and tests according to the specifications determined for samples, while ensuring the features that help to achieve research purposes. Monte Carlo methods therefore were a turning point and a contribution to the history of measurement development (Al-Zahrani, 2008). The term *Monte Carlo* usually is used when random numbers are simulated by data proposed by computer simulation methods used on sets of samples that simulate properties, features, and statistical distributions of the original population. Monte Carlo studies were used widely in the IRT field. The advantage of these modes was shown by using them to solve problems in educational and psychological measurement, and also some statistical issues, because they can describe and process parameter values properly. They can study the effects of many factors at once, and they can reduce time, effort, and cost, all of which have an impact on collecting actual data (Han & Hambleton, 2007; Harwell et al., 1996).

The authors of this study are of the opinion that there are many reasons to make the design of their experimental study depend on Monte Carlo data simulation. The most important reasons are:

Simulation programs using Monte Carlo methods provide large and random numbers for subjects' samples and tests with various lengths through different replication numbers. Such numbers are large enough that one researcher might be unable to obtain them from experimental studies with real data.

The present study compares maximum likelihood estimation and Bayes' technique of simulation of person parameter in the logistic model in various testing circumstances. This requires a high degree of experimental control of certain variables so that the comparison process is made properly, which cannot be achieved with actual data.

Many studies recommend that it is necessary to carry out experiments that depend on simulating data, such as the study of Al-Zahrani (2008), which revealed that most of the studies done by computerized adaptive testing were simulation studies. He stated that those studies obtained accurate results that conformed to the studies made by using actual data and added that at the beginning it is necessary to expand in carrying out experimentation of data simulation, and then apply that to actual data.

4. Methodology and Procedures

4.1 Methodology of the Study

This study followed the experimental approach because using the data based on Monte Carlo simulation methods achieves maximum degrees of experimental control because the data were selected in an absolutely random way through the simulated samples by the WinGen-3 program. All of the variables that could affect results of the present study, such as item features, were controlled. All of the items were unidimensional and dichotomous (0, 1), and had certain distributions in a certain range with a specific mean and standard deviation for each parameter of the logistic model used to compare the statistical methods used in this study (maximum likelihood, Bayesian prior mode). Therefore, the experimental variables represented in the present study by test length will affect the dependent variable represented by estimation accuracy of the ability parameter for both the maximum likelihood and Bayesian prior mode methods in isolation with any other variables.

4.2 Population and Sample of the Study

The population of this study comprises all of the people who have the specified level of ability of the design. All of the tests are dichotomous in gradation (0,1) and have the same specified distribution of parameters as in the design of the study. The study sample includes random samples of people, and random samples of items that are selected according to the varying test length (i.e., the number of test items), and they are chosen from the simulated data by the WinGen-3 program on the basis of steps in the computer program, and per the data that correspond with the design of the study.

4.3 Design and Variables of the Study

The Variable: Test Length
Three levels were determined for test length (10 items, 30 items, and 60 items). The variable shows in its three levels the effect of test length on the effectiveness of the maximum likelihood method and the Bayesian method in estimating the ability parameter to determine the appropriate statistical method of estimation at each level of test length.

4.4 Simulating the Data

This study used computer simulation methods to generate the required data for the study. The data for this study were simulated by the WinGen-3 program to comprise the variables in the study to compare the maximum likelihood method and the Bayesian prior mode method in estimating the ability parameter.

4.5 Estimating Subjects' Abilities with the Simulated Data

The BILOG-MG3 program was used to estimate the ability parameter of the models (two-parameter and three-parameter logistic mode) which were used in this study. This program was chosen because it provides estimations of high efficiency for two-parameter item response parameters and because it helps in balancing equivalent and inequivalent groups; in the vertical equation: tests of two stages, study of items bias, distributions of ability, statistics of suitability of item, and theoretical and experimental stability. In addition, this program contains three options for estimating item and person parameter methods (i.e., maximum likelihood, Bayesian expected posteriori, and Bayesian maximum a posteriori).

4.6 Simulation Steps and Procedures of Estimation

What follows is a description of the steps used by the researchers in simulation of the parameter item response and procedures of estimation, and then the measurement of accuracy of estimation of the ability parameter, which was obtained from the IRT models (the three-parameter logistic model in different conditions and circumstances).

4.6.1 Simulation Steps in Stimulating Data

The researchers investigated nine different conditions that were the focus of the comparison between the maximum likelihood method and the Bayesian prior mode method within each logistic model (one-parameter and two-parameter). To determine the different conditions, data were simulated by using the WinGen-3 program for 100 replications for each of these conditions according to the following standards:

1. Item difficulty and discrimination parameter, then simulating it from an ordinary distribution.
2. Item difficulty parameter set randomly from +3 to -3.
3. Item discrimination parameter was set randomly from 0.5 to 2, with a mean of 1.25 and a standard deviation of 0.25.
4. Subjects' ability belongs to the normal standard distribution from +3 to -3, with a mean of 0 and a standard deviation of 1.
5. Guessing coefficient was set from 0.10 to 0.30, with a mean of 0.20 and a standard deviation of 0.05.

6. Sample size was controlled according to each condition and within the three sets: 500, 1000, and 1500 testees.
7. The number of test items was controlled according to each condition and within the three sets: 10, 30, and 60 items.

Table 1: Description the settings from the WinGen-3 program

Determining the number of testees as 500	The logistic model used is the three-parameter version.
Distributing subjects' abilities normally with a mean of 0 and a standard deviation of 1	Items difficulty coefficients from +3 to -3
Setting the guessing coefficient from 0.10 to 0.30 with a mean of 0.20 and a standard deviation of 0.05	Distributing items discrimination factors normally with a mean of 1.25 and a standard deviation of 0.25
The response pattern on the test is binary.	Determining the gradation factor as D = 1.702.
The number of test items is 10.	Determining the number of replications as 100 times.

For example: To simulate the data comparing the maximum likelihood method with the Bayesian prior mode method in short tests applied to a sample of 500 testees within the three-parameter logistic model, we selected the following settings from the WinGen-3 program, as shown in Table 1. One hundred different copies of persons' responses were obtained from the program, for a test with the stated specifications.

4.6.2 Ability Parameter Estimation

To estimate the ability parameter and compare the effectiveness of the maximum likelihood method and the Bayesian method of estimation – the goal of the study.
Researchers performed the following steps:

- The data simulated for the two-item response models, and for each model separately (the two-parameter logistic model and the three-parameter logistic model) according to each of the mathematical design conditions (using the BILOG-MG3 program) were recalled.

Commands (syntax) for the BILOG-MG3 program to correspond with the conditions in which the comparison process were implemented, and the estimation process of subjects' abilities was done using the maximum likelihood method and the Bayesian prior distribution.

After obtaining the estimations of subjects' abilities in each of the study's conditions and with the two previously mentioned methods, we determined their effectiveness in estimating the ability parameter (θ) using the statistical index (relative efficiency criterion).

The previous steps were replicated 100 times for each of the test conditions, and the results of these steps were 100 ability estimations with the maximum likelihood method and 100 ability estimations with the Bayesian method, and 100 measurements for the relative efficiency index and for each model of the study.

The researchers then analyzed the results of each of the test conditions by using the descriptive and deductive approaches according to the questions and the experimental design of the study.

4.7 Statistical Treatments

To answer the questions posed by the study, data were simulated using the WinGen-3 program. These data were then recalled using BILOG-MG3 and the statistical package for the social sciences (SPSS), and the following statistical treatments were processed:

1. Estimating subjects' abilities by the maximum likelihood method according to each test

circumstance.

2. Estimating subjects' abilities by the Bayesian prior mode method according to each test circumstance.
3. Determining the standard error of estimation by these two methods.
4. Finding the mean of standard error of estimation and the outcome of the number of 100 replications for each test circumstance.
5. Calculating the relative efficiency index.

5. Results and Discussion

Does the accuracy of estimation of the ability parameter vary as the test length varies in the two-parameter logistic model when the maximum likelihood method and the Bayesian method of expected prior mode are used?

To answer this question, items were simulated by the process described earlier. Subjects' ability parameter and standard error of the degree of estimation were determined by both the maximum likelihood method and the Bayesian method of prior mode at each of the three conditions of test length in the control of sample size within its three levels (i.e., 500, 1000, and 1500). The average standard error of estimation of persons' abilities and the relativity efficiency factor were determined also. Table 2 shows the results of that process. The data in Table 2 related to standard error by ability parameter estimation methods SEML and SEB indicate that when the difficulty level and sample size are stable, the accuracy of the ability parameter estimation in the three-parameter model increases by the increase of the number of test items, assuming that the relationship between standard error and estimation accuracy is inversely proportional, and that this could reflect the data correspondence quality of long tests compared to short tests.

Table 2: Standard Error by Ability Parameter Estimation Methods and their Relative Efficiency Using the Two-parameter Logistic Model According to Difference of Test Length.

Sample Size	No. of Test Items	*Standard Error by Ability Parameter Estimation*		Test Relativity Efficiency
		**SEML*	** *SEB*	Support and RE(B/LM)
500	10	1.71	0.85	4.04
	30	0.51	0.53	0.93
	60	0.32	0.37	0.76
1000	10	0.69	0.63	1.21
	30	0.49	0.51	0.92
	60	0.40	0.43	0.86
1500	10	1.16	0.75	2.37
	30	0.47	0.50	0.86
	60	0.35	0.44	0.64

**SEML*: Standard error by Maximum Likelihood method, ***SEB* Standard error by Bayesian method.

This result corresponds with what Ababnah (2004), Al-Darabee (2001), and Hambleton et al. (1991) have noted, indicating that the increase of estimation accuracy of ability parameter increases with the increase in the number of test items. To compare the two methods with the accuracy of ability parameter estimation and under the same conditions, it is necessary to refer to relativity efficiency criterion RE(B/LM) at different test lengths. This showed that in long tests (60 items) and tests of average length (30 items), the effectiveness is related to the maximum likelihood method when compared with the Bayesian method for all conditions of the sample size. In short tests (10 items), the Bayesian method of prior mode for estimating accuracy outperformed when compared with the maximum likelihood method in all the sample size conditions.

Does the accuracy of estimation of the ability parameter vary as test length varies in the three-parameter logistic model when the maximum likelihood method and the Bayesian method of expected prior mode are used? To answer this question, items were simulated by the process described earlier and then the subjects' ability parameter and standard error of the accuracy of estimation were determined by both the maximum likelihood method and the Bayesian method of prior mode for each test length in the control sample size (i.e., 500, 1000, and 1500). Standard error of estimation of subjects' abilities and relativity efficiency factor were also taken. Table 3 shows the results of that process.

The data in Table 3 reflecting standard error by ability parameter estimation methods (SEML) and SEB indicate that when the difficulty level and sample size are stable, accuracy of ability parameter estimation in the three-parameter logistic model per the maximum likelihood method and Bayesian method increases by the increase in the number of test items, assuming that the relationship between standard error and estimation accuracy is inversely proportional. This result corresponds to the effect of test length on estimation accuracy of ability parameter for the two-parameter logistic model. This could be related to the same reasons stated by Hambleton et al. (1991), namely: the increase of estimation accuracy of ability parameter is linked to the increase of test items, and that the increase or

Table 3: Standard Error by Ability Parameter Estimation Methods and Their Relative Efficiency under the Assumption of the Three-Parameter Logistic Model According to Difference in Test Length.

Test Level	Sample Size	No. of Test Items	Standard Error by Ability Parameter Estimation		Test Relativity Efficiency
			*SEML	**SEB	RE(B/LM)
Difficulty Average	500	10	1.34	0.79	2.86
		30	0.72	0.61	1.39
		60	0.36	0.47	0.59
	1000	10	1.46	0.81	3.24
		30	0.66	0.56	1.39
		60	0.48	0.49	0.96
	1500	10	0.89	0.67	1.77
		30	0.57	0.46	1.54
		60	0.45	0.47	0.92

Decrease in test length to achieve the desired accuracy depends on the assumption that the added or deleted items are similar in their statistical features to the rest of the items of the test. Long tests therefore reflect a better-quality matching data than short tests. To compare the two methods with the extent of accuracy of ability parameter estimation and under the same conditions, it is necessary to refer to relativity efficiency criterion RE(B/LM) at the different levels of test length. This showed that in tests with short and average length (10 or 20 items), the Bayesian method outperformed estimation accuracy in all test conditions of sample size, whereas in long tests (60 items), maximum likelihood outperformed in all conditions of sample size.

In the light of the results obtained, the researchers recommend the following areas for further study:

Test length for the logistic method (two-parameter and three-parameter) for the increase of estimation accuracy of ability parameter in both methods, the maximum likelihood and Bayesian prior mode, should be increased.

In the two-parameter logistic model, and for long and average tests, the maximum likelihood method should be used for ability parameter estimation, and for short tests, the Bayesian prior mode is recommended.

In the three-parameter logistic model, the Bayesian method is recommended for the ability parameter estimation with short and average tests, whereas it is recommended that for long tests, the maximum likelihood method should be used for all conditions of sample size.

6. Conclusion

This study compared the effect of test length on the degree of ability parameter estimation in the two-parameter and three-parameter logistic models by using the Bayesian method of expected prior mode and maximum likelihood. The experimental approach was followed, using the Monte Carlo method of simulation. The population of the study consisted of all subjects who had the specified ability level in the design. All tests specifying parameter distribution in the study design were binary graded (0,1). In addition, the study included random samples of subjects and of items that were selected according to the variable of test length, which were selected from the simulated data by the program WinGen-3. The study revealed that estimation accuracy of the ability parameter in the two-parameter logistic model according to the maximum likelihood method and the Bayesian method increases with the increase of the number of test items. To compare the two methods by range of accuracy of estimation of the ability parameter, the results showed that with long tests (60 items) and those of average length (30 items), the effectiveness is related to the maximum likelihood method and to all conditions of the sample size, whereas in short tests (10 items), the Bayesian method of prior mode in estimating accuracy outperformed in all conditions of the sample size. Moreover, results showed that the increase of estimation accuracy of the ability parameter in the three-parameter logistic model, according to the maximum likelihood and the Bayesian method, increases with the increase in the number of test items. To compare the two methods with respect to the range of accuracy of estimation of the ability parameter, results showed that in the case of short tests (10 items) and those of average length (30 items), the Bayesian method outperformed with respect to the accuracy of estimation at all the conditions of the sample size. In long tests (60 items), however, the maximum likelihood method outperformed at all different conditions of sample size.

7. Declarations

Ethics approval: All procedures in this study were approved by the Research and Ethics Committee of the institution of the authors.

Availability of data and material: The data that support the findings of this study are available from the corresponding author, M. Al-Qahtani, upon reasonable request.

Authors' contributions: All authors contributed to the study conception and design. MAL-Q prepared the material. EAW collected the data. MAL-Q performed statistical analyses of the data and interpreted the results. The first draft of the manuscript was written by MAL-Q. All authors read and approved the final manuscript.

Consent to participate: Informed consent was obtained from all individual participants included in the study.

Code availability: The Code for Study can be obtained upon request to the corresponding author.

References

Ababnah, E. (2004). The Effect of Sample Size and Its Selection and the Number of Items and their Selection on the Accuracy of Estimation of Item Parameters and Ability of a Mental Ability Test Using Item Response Theory. Amman Arab University, Amman, Unpublished Doctoral Dissertation.

Abu-Allam, R. M. (2005). Evaluation of education.

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. International Journal of Asian Social Science, 3(4), 992–1011. http://www.aessweb.com/journal-detail.php?id=5007

Al-Darabee, M. (2001). The effectiveness of the one-parameter logistic model Rasch model in the accuracy of estimation of person's ability and the item difficulty coefficient with the variance of sample size and test length. Humanitarian Science Studies Journal, 1, 197–208.

Al-Owidha, A. (2007). A comparison of the Rasch model and the three-parameter logistic model applied to the quantitative subtest of the General Aptitude Test. University of Denver, USA., Unpublished Doctoral Dissertation.

Al-Ta'mari, B. (2003). Application of the three-parameter logistic model in estimation of person's ability and items parameters of testing a multiple choice test. Mutah University, Jordan., Unpublished Doctoral Dissertation.

Al-Zahrani. (2008). The Impact of the Difference in Sample Size and the Increase of Ability Range on Accuracy of Estimation of Actual Degree Estimated by the Traditional Models and One-Dimentional Models in the Modern Theory of Measurement. Um-Al-Qura University, Saudi Arabia. Unpublished Doctoral Dissertation.

Albano, A. D. (2017). Introduction to educational and psychological measurement using R.

Allam, S. (2005). Unidimensional and multidimensional testing item response models and their applications in the educational and psychological measurement.

Allam, S. (2006). Educational and psychological measurement and evaluation: Its fundamentals, applications, and contemporary orientations.

Azizan, N. H. B., Mahmud, Z. B., & Rambli, A. Bin. (2019). Rasch measurement model: a review of Bayesian estimation for estimating the person and item parameters. Journal of Physics: Conference Series, 1366(1), 12105.

Baker, F. B., & Kim, S.-H. (2017). The basics of item response theory using R. Springer.

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. Journal of Educational Measurement, 38(3), 191–212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x

Bazaldua, D. A. L., Lee, Y. S., Keller, B., & Fellers, L. (2017). Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. Asia Pacific Education Review, 18(4), 585–598. https://doi.org/10.1007/s12564-017-9507-4

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39(4), 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Chaturvedi, A., & Vyas, S. (2017). Estimation and testing procedures for the reliability functions of three parameter Burr distribution under censorings. Statistica, 77(3), 207–235.

Costa, P., & Ferrão, M. E. (2015). On the complementarity of classical test theory and item response models: Item difficulty estimates and computerized adaptive testing. Ensaio, 23(88), 593–610. https://doi.org/10.1590/S0104-40362015000300003

Crocker, L. & Algina, J. (2009). An introduction to the traditional and contemporary measurement theory.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. ERIC.

der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In Elements of adaptive testing (pp. 3–30). Springer.

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. Medical Education, 37(8), 739–745. https://doi.org/10.1046/j.1365-2923.2003.01587.x

Draxler, C., & Alexandrowicz, R. W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. Psychometrika, 80(4), 897–919. https://doi.org/10.1007/s11336-015-9472-y

Dresher, A. (2003). An empirical investigation of local item dependency in NAEP data. Annual Meeting of the American Educational Research. https://scholar.google.com/scholar?start=200&hl=en&as_sdt=0,5&sciodt=0,5&cites=1751750975108000882&scipsc=#14

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research, 16(SUPPL. 1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. Psychometrika, 65(3), 337–362. https://doi.org/10.1007/BF02296150

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Maheah. New Jersey: Lawrence Erlbaum Associates, Publishers.

Finch, H., & Edwards, J. M. (2016). Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach. Educational and Psychological Measurement, 76(4), 662–684. https://doi.org/10.1177/0013164415608418

Fox, J. P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. Psychometrika, 68(2), 169–191. https://doi.org/10.1007/BF02294796

Frank, R. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models. ETS Research Report Series, 2009(1), i--31.

Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. Behaviormetrika, 33(1), 43–59.

Hadwin, A. F., Winne, P. H., & Nesbit, J. C. (2005). Annual review: Roles for software technologies in advancing research and theory in educational psychology. British Journal of Educational Psychology, 75(1), 1–24. https://doi.org/10.1348/000709904x19263

Hambleton, R. K., & Swaminathan, H. (1985). 1985: Item response theory: principles and applications. Boston, MA: Kluwer-Nijhoff.

Hambleton, R., & Linn, R. (1989). Educational measurements. New York, NY: Macmillan Publishing Company.

Hambleton, R. K, Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Sage.

Hamid, S. (2008). The Effect of Item Response Model and Multiple Dimensions and Correspondence Method in Estimation of Items and Persons' Parameters. Amman Arab University, Amman.

Han, K. T., & Hambleton, R. K. (2007). User's manual for WinGen: Windows software that generates IRT model parameters and item responses. Center for Educational Assessment Research Report, 642, 516–524.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied Psychological Measurement, 20(2), 101–125. https://doi.org/10.1177/014662169602000201

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Dorsey Press.

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. Colombian Applied Linguistics Journal, 16(2), 167. https://doi.org/10.14483/udistrital.jour.calj.2014.2.a03

Jin, K. Y., & Wang, W. C. (2014). Item response theory models for performance decline during testing. Journal of Educational Measurement, 51(2), 178–200. https://doi.org/10.1111/jedm.12041

Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. ETS Research Report Series, 2015(1), 1–19.

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory interms of both variables of test length and sample size. Procedia - Social and Behavioral Sciences, 46, 135–140. https://doi.org/10.1016/j.sbspro.2012.05.082

Kymlicka, W., & others. (2002). Contemporary political philosophy: An introduction. Oxford: Oxford University Press.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13(4), 517–549. https://doi.org/10.1177/001316445301300401

Meara, K., Robin, F., & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. Multivariate Behavioral Research, 35(2), 229–259. https://doi.org/10.1207/S15327906MBR3502_4

Nofer, D. C. (2007). Conditional item dependence for testlet items.

O'Connor, B. P., Crawford, M. R., & Holder, M. D. (2015). An item response theory analysis of the subjective happiness scale. Social Indicators Research, 124(1), 249–258.

on Psychological Testing, I. V., & others. (2015). Psychological testing in the service of disability determination. Psychological Testing in the Service of Disability Determination, 1–246. https://doi.org/10.17226/21704

Pelton, T. W. (2002). the Accuracy of Unidimensional Measurement Models in the Presence of Deviations From the Underlying Assumptions. In Education (June issue). Brigham Young University, USA.

Philip, A., & Ojo, B. O. (2017). Application of item characteristic curve (ICC) in the selection of test items. British Journal of Education, 5(2), 21–41.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the Estimation of Standard Errors in Cognitive Diagnosis Models. Journal of Educational and Behavioral Statistics, 43(1), 88–115. https://doi.org/10.3102/1076998617719728

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science, 14(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Reise, S. P., & Waller, N. G. (2003). How Many IRT Parameters Does It Take to Model Psychopathology Items? Psychological Methods, 8(2), 164–184. https://doi.org/10.1037/1082-989X.8.2.164

Rupp, A. A. (2003). Item Response Modeling With BILOG-MG and MULTILOG for Windows. International Journal of Testing, 3(4), 365–384. https://doi.org/10.1207/s15327574ijt0304_5

Si, C.-F., & Schumacker, R. E. (2004). Ability Estimation Under Different Item Parameterization and Scoring Models. International Journal of Testing, 4(2), 137–181. https://doi.org/10.1207/s15327574ijt0402_3

Steel, P., & Klingsieck, K. B. (2016). Academic Procrastination: Psychological Antecedents Revisited. Australian Psychologist, 51(1), 36–46. https://doi.org/10.1111/ap.12173

Sulaiman, M. A. (2009). Measurement and Evaluation: Its fundamentals, tools and applications. In Cairo: Modern Book Publishing House.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47(4), 397–412. https://doi.org/10.1007/BF02293705

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the balanced inventory of desirable responding: Dichotomous versus polytomous conventional and IRT scoring. Psychological Assessment, 26(3), 878–891. https://doi.org/10.1037/a0036430

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. Frontiers in Psychology, 5(AUG), 920. https://doi.org/10.3389/fpsyg.2014.00920

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. Psicologia: Reflexao e Critica, 29(1). https://doi.org/10.1186/s41155-016-0040-x