The Effect of Methods of Estimating the Ability on The Accuracy and Items Parameters According to 3PL Model

Deyab A. Almaleki † and Ahoud Ghazi Alomrany †

†Department of Evaluation, Measurement and Research, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract

This study aimed to test method on the accuracy of estimating the items parameters and ability, using the Three Parameter Logistic. To achieve the objectives of the study, an achievement test in chemistry was constructed for third-year secondary school students in the course of "natural sciences". A descriptive approach was employed to conduct the study. The test was applied to a sample of (507) students of the third year of secondary school in the "Natural Sciences Course". The study's results revealed that the (EAP) method showed a higher degree of accuracy in the estimation of the difficulty parameter and the abilities of persons higher than the MML method. There were no statistically significant differences in the accuracy of the parameter estimation of discrimination and guessing regarding the difference of the two methods: (MML) and (EAP).

Key words:

Estimation-Accuracy; Three-Parameter Logistic Model (3PL); Items Parameters; Accuracy

1. Introduction

Measurement theories aim to provide a basis for making predictions about the traits or abilities that are measured by the test items. The classical measurement theory has been used for a long time to reach this objective. Within the framework of this theory, the concept of ability is expressed through the true score, which is defined as predicting the obtained score, which is obtained after re-applying the test a lot of times on the respondents [1–2]. Then came the Item Response Theory, which was mainly related to the objectivity of the measurement, as it assumes that there is a relationship between the amount of the respondent possessing of the trait "ability", expressed in the symbol (θ) and the difficulty parameter of an item, expressed in the symbol (bi), and the probability of the respondent obtaining the correct answer at the level of a specific ability, expressed in the symbol Pi (θ), and that this relationship takes the form of a mathematical curve whose shape is supposed to be like the letter S where the respondents' abilities and the difficulty of the items are calibrate on one curve [3–5]. This relationship has been formulated using mathematical or logistical equations called item response models, which can be represented graphically using the so-called item characteristics curve [6]. These models vary according to the relationship, is it logistical or mathematical, and according to the nature of the response; is it bilateral or multi-response, and according to the factor structure of the measured characteristic whether it is unidimensional or multi-dimensional, and also according to the parameters of the items that enter the relationship model [7]. Among the most famous of these models that are appropriate for the two-answer items are the One Parameter Logistic Model (1PLM), the Two Parameter Logistic Model (2PLM) and the Three Parameter Logistic Model (3PL)

Estimating the ability of respondents is a basic component of the Item Response Theory, and there are several methods of estimating ability, the most famous of which are:

- Maximum Likelihood (ML) method: Its most important methods are: Joint Maximum Likelihood Procedure (JML), Marginal Maximum Likelihood (MML), and Conditional Maximum Likelihood (ML)
- Maximum A Posteriori Method (MAP)
- Expected A Posteriori Method (EAP)

Given that being accurate in the measurement process leads to preventing wrong decisions at the personal and community level, the aim of this study was to reveal the effect of the method used to estimate the ability and to accurately estimate the characteristics of items and persons according to the 3PL Model and by using the data resulting from electronic test in chemistry level 3.

Estimating the parameters of IRT

The process of estimating the parameters of the items and the abilities of persons according to the item response theory is one of the basic steps in the application of this theory [8]. Views have varied on the best way to estimate the abilities of persons, as the results of studies conducted by Rajlic [9] and DeMars [10] show that there are clear differences between the ML method and the methods based on Bayes theory, among which is that the ML method gives high errors in estimating the ability parameter compared to Bayes' methods.

The study of Chen and Choi [11] showed that there are no statistically significant differences in the averages of standard errors for estimating each of the parameters of the items and the ability parameter due to the difference in the estimation method.

Li et al. [7] recommended studying the (EAP) method for estimating ability due to the low average standard errors in the estimation. They compared the accuracy of the estimates

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.12

for the ability parameter with the different nature of the data (real, generated) according to the difference in the length of the test.

On the other hand, there are studies that have confirmed the preference of the ML Method, including the study of Chen et al. [11] which used the data generated in their study.

It is noted from the previously presented studies that there is a difference in the accuracy of parameter estimation and ability due to the method used in the estimation process. Hence the problem of the study, which aimed to find out the effect of two methods of estimation: the MML method and the EAP method on the accuracy of estimating the statistics of items and persons using the three-parameter model, whose assumptions are more realistic than the one and the two-parameter model, as it allows for an opportunity for guesswork when estimating the abilities of persons in what is measured. Are the results of using of (MML) consistent with (EAP) when estimating the parameters of items and persons? This question can be answered by answering several sub-questions, namely:

Is there an effect of the ability estimation methods (MML method) and the (EAP) method on the accuracy of estimating the parameters of the items for the 3PL model?

Is there an effect of the ability estimation methods (MML method) and the (EAP) method on the accuracy of estimating the abilities of persons for the three-parameter model?

2. Theoretical Consideration

2.1 The Development of Measurement Theories Concept

Measurement processes and their tools are of interest to specialists in the natural, behavioural and human sciences alike [12]. All these sciences seek to develop accurate and objective methods for measuring phenomena related to them in order to understand, interpret and predict existing phenomena among their variables, in order to control them, leading to the accuracy of their results, as the progress of any science is measured by the degree of accuracy it attains in defining its concepts and in the accuracy of the tools used for measurement [3,13–17]. Psychological and educational measurement is more difficult than natural measurement, because the nature of psychological and educational phenomena is intertwined and affected by many variables, either directly or indirectly [18]. Therefore, the interest of psychometrists in the logic of measurement and quantitative methods was more than the scientists of natural sciences, due to the complexity of psychological phenomena and the multiplicity of their variables [10].

Psychometric and educational theories came to put forward assumptions through which psychological and educational measurements are reached to the maximum possible accuracy, and it was the classic theory that dominated for a short time in the study of the characteristics of the tests and the characteristics of their items, but it suffered from a major problem referred to by [19]–[23] that all psychometric characteristics such as difficulty and discrimination factors depend on the characteristics of the respondents, which is referred to as (Group-Dependent). The factors of difficulty and discrimination fluctuate with the change of the traits or abilities of the persons of the sample respondents, so the difficulty level of the items will increase if the ability of the sample members that were used in estimating the difficulty of the items was higher than the average ability of the sample population [10]. The discrimination indices tend to be higher when estimated by relying on a heterogeneous sample in ability than if it was estimated by relying on a homogeneous sample [24]. The average and extent of the ability level is affected by the values of the item's parameters, so persons who are exposed to items with high difficulty coefficients will be Low and vice versa, which is referred to as (Test-Dependent) [6]. The classical theory of the test did not provide mathematical models that contribute to estimating the probability of the respondent's correct answer on any item of the test [25].

In spite of these and other weaknesses, the classical theory of measurement still has the merit of establishing and developing the current concepts of measurement, but to avoid such weakness and try to reach an objective measurement of the trait through behavior, some scientists believe that it is necessary to search for a new measurement theory which is devoid of such weaknesses [26].

The Item Response Theory (IRT) came to avoid the weaknesses and shortcomings of the classical theory, which were mainly related to the objectivity of the measurement as it assumes that there is a relationship between the amount of the respondent's trait (ability), expressed in the symbol (θ) and the difficulty parameter of an item expressed in the symbol bi and the probability that the respondent will obtain the correct answer at a certain ability level and is expressed by the symbol $P_i(\theta)$and that this relationship takes the form of a mathematical curve whose shape is supposed to be like the letter "S" where the respondents' abilities are calibrated and the difficulty of the items on one curve. This relationship was formulated using mathematical or logistical equations called item response models, which can be represented graphically using the so-called item characteristics curve [16–17, 27–30].

This theory has a set of features pointed out by [3, 19, 21, 35].

- The existence of a large group of test items that measure the same trait, and the estimate of the person's ability is independent of the sample items applied to him (Item Free)
- The presence of a large population of persons, in which the psychometric properties of the items - parameters of difficulty and discrimination - are

Corresponding author: Almaleki, Deyab

independent of the sample of persons that were used to estimate these characteristics (Person Free)

2.2 Item Response Theory Assumptions

Like other theories, the item response theory is based on several assumptions that distinguish it from others:

Unidimensionality

Swaminathan et al. [37] and Hambleton and Regers [38] point out that Unidimensionality means that there is one underlying factor performance on a scale, and that factor is the measured ability or trait. Kishino et al. [39] also explained that Unidimensionality means homogeneity of the scale items among themselves, and its measurement is the same trait. One of the most popular methods used to verify Unidimensionality is to identify the factors whose Eigen value is greater than (1), then the graphic representation of the Eigen values of these factors. If there is a large regression between the first factor and the second factor in the Eigen value, then this is evidence that the variance in performance over the items refers to a large extent to the first factor, thus fulfilling the Unidimensionality condition [3-4, 36– [37].

Local Independence

Local independence refers to the fact that the answer to any of the scale items is not affected by the answer to any other term, either negatively or positively [3, 10, 38–39]. This means that the assumption of local independence is achieved if the probability of the correct answer on any of the scale terms is not related to the probability of the correct answer for any other item [43]. Kishino et al [39] also pointed out that local independence means that if the effect of the factor or the scale factor loading is removed; there will be no systematic variation between the items.

Local independence is expressed statistically by the absence of any statistical correlation between the scale item when the measured ability is fixed, that is, there is no statistical correlation between the scale items of respondents with the same measured ability Hambleton et al. [44] and Swaminathan et al. [37]. Among the most famous statistics used to verify local independence is the Fisher's Z index, in which the observed errors are converted into standard errors, and the logic behind this is that the distribution of the Fisher's Z index values for independent items should be distributed normally with an average of (zero), and then if the arithmetic mean of the Fisher indices of each of the items pairs, located between the lower limit and the upper limit of the Fisher index of the items pairs, this indicates that they are statistically independent items. The identification of the lower and upper limit of the confidence or the so-called confidence interval is determined by subtracting and adding two standard deviations of the Fisher's Z index values. The item pairs are statistically independent if their observed Fisher's Z index falls outside the lower and upper limits of confidence [3–4, 37, 42–48].

Item Characteristic Curve (ICC)

This curve is one of the central concepts in the Unidimensionality models, and this curve represents the mathematical relationship between the probability of a person answering a correct answer on the item and the ability or trait measured by the test items that contain items, which is a non-linear regression function. Knowing the test scores of the respondents of each specific ability level it is possible to draw a characteristic curve for any item of that test, which represents the regression line that passes the mean of the conditional distributions for each ability level [2, 24, 28, 34, 48].

Speediness

Most of the loading traits models assume that the speediness factor does not play a role in answering the item, and that the person’s failure to answer the test items is due to his reduced abilities and not to the effect of the speediness factor in answering or not reaching these items due to limited time [3–4, 36–37, 48–49].

Item Response Theory Models

The item response theory provides mathematical models that explain the relationship between a person's response observed on the test and the loading ability behind this response, and these models vary according to the diversity of the level of response on the test items, the parametric structure of the model used and the dimensions of the loading space behind the response of persons to the items [43]. The loading trait models have been classified into two types of models: the Unidimensionality models, which assume that there is one continuous trait that underlies the response of persons to the scale items, and the multidimensional models, which assume that there is more than one dimension that lies behind the responses of persons on the scale items [3, 36, 50–53]. Three of the Unidimensionality item response models are commonly used, and these models are appropriate for two-dimensional items, the most famous of which are:

- One-Parameter Logistic Model (Modal Rasch): It is one of the simplest models for responding to a two-stage item because this model contains only the parameter of item difficulty, and this model is a special case of the two- and three-parameter logistic model. This model assumes that all items distinguish equally between persons, and it is assumed that the answers are not influenced by the guessing parameter, but that they differ only in difficulty [37, 51, 54–55].
- Two Parameter Logistic Model (Lord Model): This model assumes that the items differ in difficulty and discrimination, and the answers are not

affected by the guessing parameter [53]. Item distinction can be defined as the slope of the item characteristic curve at the point of the curve inflection, which is the point where the probability of the person answering the item for a correct answer is equal to (0.5) since the greater the item distinction value, the greater the slope of the curve at the point of inflection, and thus the greater the distinction Item [37, 51, 54–55].

- Three-Parameter Logistic Model: the third parameter, which is guessing, because some test items sometimes allow some respondents with very low ability to arrive at the correct answer by guessing, so this is the most general form of the two-parameter model and the one-parameter. The assumptions of the three-parameter model are the most realistic in analysing and calibrating the items of recognition tests that guessing may affect the responses of persons, and the mathematical formula for this model is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{\left[1+\exp\left\{-Da_i(\theta-b_i)\right\}\right]} \quad i=1,2 \ldots n$$

Pi (θ): the probability that the respondent, chosen at random from the ability level (θ) for (i) item may answer a correct answer.

b_i: is the difficulty parameter for the item. (i)

θ: the ability parameter

D1: represents the scaling factor.

e: is the natural logarithmic base and equals (2,7183)

a_i: the discrimination parameter for item (i)

c_i: the guessing parameter for item (i)

It is clear from the previous equation that the three-parameter model assumes that there is an opportunity to guess when estimating the abilities of persons in what is measured, that is, that ($c \neq 0$) and this is in contrast to the two-parameter models which assume that ($c = 0$); Therefore, it is preferable to use the three-parameter model in the item's analysis of recognition questions. This parameter is also called the Pseudo-Chance level or the Lower Asymptote [53].

The items characteristic curve of the three-parameter model differs from the one and the two models in that the curve does not start from point zero on the y-axis which represents the correct probability of the answer; This is because the probability that low-ability persons in the test measures will reach the correct answer for the item is not equal to zero because of the probability of guessing [38, 40].

Estimation in Item Response Theory

One of the main issues when using one of the models of response theory to the item is in estimating the model parameters, which depend on the methods of numerical analysis through the use of different computer programs. Hambleton et al. [44] and Swaminathan et al. [37] counted the statistical estimate of the relationship between the probability of a correct response for an item of the test and the ability measured by the test, which is the main problem for the user of this theory. Hambleton et al. [44] showed that the estimation of the parameters of the item is one of the most important issues on which the success of the item response theory depends, especially in applications that depend a lot on those parameters, which made the psychometric search interested in searching for the best methods of statistical estimation of the parameters of the items and the persons abilities, in addition to that the development of probabilistic models to arrive at best estimates.

The standard error is a statistical indicator on which researchers rely to judge the accuracy of the sample's ability of the response to the population [22, 33, 36, 38, 53]. This is similar to the tests, as there is a contrast in the grades or in the estimation of the ability from one test position to another, and the researcher who uses IRT obtains the standard error of each ability. Hambleton et al. [44] and Swaminathan et al. [37] pointed out the importance of determining the amount of error in estimating the parameters, which expresses the accuracy in estimating the parameters of the model used, considering that if a small amount of standard error is obtained, it is an indication of the accuracy of the measurement function of the test, especially since it is an indication of the reliability of the test in IRT. It must be noted that the test information function has an important advantage, which is that it is independent of the respondent. Accuracy in estimation refers to the extent to which the decision based on test scores is consistent with the decision that can be made if the scores do not include any measurement errors[56–59].

From this standpoint, the psychometric research was concerned with the use of IRT in searching for the most accurate methods used in estimating the parameters of the items and persons. The results of studies have varied about the most accurate methods of estimating the parameters of the items and persons. The results of the study Swaminathan et al. [37] showed that there are clear differences between the ML method and the methods based on Bayes theory, and that the ML method gives higher errors in estimating the Ability parameter compared to the Bayes methods.

The results of the study DeMars [10] showed that there were no statistically significant differences in the mean of the standard errors for estimating each of the items parameters and the ability parameter due to the difference in the estimation method. As for the current study, it revolves around two methods of estimating the factor loading of the respondents' response to a group of items using actual data, and these methods are:

- Marginal Maximum Likelihood Estimation (MML) method: It is one of the most popular methods of ML, and this method is characterized by that it can be used to estimate the features of all Unidimensionality models, as well as multi-dimensional models, and this method is effective, whether the items of the test are few or many. The estimated values of the resulting standard errors are characterized by accuracy. By successive re-evaluation of the estimation processes, and we can obtain estimated values for the parameters of the persons who answered a correct or wrong answer on all items, and give estimates for the total score, and therefore there is no loss of information due to the deletion of the response of some of the study members, In addition to the resulting estimates being consistent, and approaching the real values by increasing the sample size, according to this method, a value (θ) is found that makes its value (θ) the largest through a mathematical equation, by finding the first derivative of that equation and equating it to zero , Whereby, according to this method, the marginal likelihood function of item parameters is found by integrating the density coupling over the ability parameters [22, 34, 40, 50, 53, 58].
- Estimation using Maximum A posteriori (MAP) method: It is one of the most popular methods that depend on Bayes Theorem, which is used to address some of the deficiencies in the high probability method related to the case of correct or wrong answer on all items. The ability parameter in infinity is positive in the case of all correct answer to items and minus infinity in the case of the wrong answer. To address this deficiency, the MAP method in estimating the ability depends on the use of the pre-distribution of ability, in addition to the procedures of the method of ML, and the habit of Pre-distribution of the ability is the normal distribution [9, 11, 59].

3. Methodology

3.1 Population

The study population consists of male and female students in the third year of secondary "level five of Curriculum system" of the "natural sciences course", who are attending the academic year 2020.

3.2 Sample

The study sample consisted of (507) male and female students in the third year of secondary "level five of Curriculum system" of the "natural sciences course", who are attending the academic year 2020.

3.3 Measure

An electronic test in Chemistry level 3 was prepared for students of the third year of secondary school, the fifth level of the curriculum system "Natural Sciences course" Edition 2020" according to the following steps:

- Determining the purpose of the test: The test aims to measure the achievement of male and female students in chemistry 3 in four cognitive levels: remembering, comprehension, application, and analysis.
- Determining the content: The test seeks to measure students' achievement in four subjects: states of matter, energy and chemical changes, speed of chemical reactions, and chemical equilibrium.
- Determining the objectives of the test: The test objectives were set to measure four cognitive levels, which are Knowledge, comprehension, application, and analysis. The number of objectives that the study seeks to measure reached 37 objectives. These objectives were distributed among the levels of objectives as follows: 4 objectives for memorization, 6 objectives for comprehension, 10 objectives for implementation, and 17 objectives for analysis.
- Determining the relative weights of the content subjects: This was done based on the number of classes allocated to teaching each topic, and these weights were as follows: states of matter 30%, energy and chemical changes 20%, speed of chemical reactions, and 25%, chemical equilibrium 25%.
- Determining the relative weights of the levels of objectives: By dividing the number of objectives in each of the levels by the total number of objectives, and these weights were as follows: Knowledge level 11%, comprehension level 16%, Application level 27%, and analysis level 46%.
- Determining the number of test items: The test objectives were measured through 25 items that were formulated in the form of multiple choice in which one correct alternative was chosen from among four alternatives.
- Establishing a table of test specifications: To ensure that the test items are distributed in proportion to the relative weights of target levels and content topics.

- Verification of the validity of the test: The validity of the test content was verified by presenting it to a number of arbiters, including supervisors and teachers of chemistry, and their number was 7; this is to determine the item's measurement of the goal to be measured, the correctness of its linguistic formulation, and its suitability for students. All arbiters have reported the appropriateness of the items to measure the goals that have been set to measure them as well as their suitability for students, with some suggesting to re-formulate items (2,9,13,21) and the required amendments have been made.
- The validity of the test was verified by fit, its items to the assumptions of the three-parameter model, and (12) items were deleted for non-fit.

3.4 procedures

The assumptions of the IRT were verified on the data obtained from applying the test items to the study sample according to the following:

First: Verifying the Unidimensionality assumption: By performing a factor analysis and calculating the ratio between the Factor Loading of the first factor and the factor loading of the second factor, and the results were as in Table 1.

Table 1. Factor loading of the first factor and the second factor and the ratio between them

Factor	Factor loading value	Explanatory Invariance ratio	Factor loading ratio of factor 1 to Factor Loading ratio factor 2
1	5.936	23.743	
2	1.500	6.000	3.95

Table 1 shows the ratio between the Factor loading of the first factor and the factor loading of the second factor exceeds the value (2), which is the value that Hambleton and Jones [62] specified as a condition for Unidimensionality verification in the scale, which confirms the fulfillment of Unidimensionality assumption in the study tool.

Second: Verification of Local Independence: This was done by using the method of average correlation coefficients of the intrapersonal items, where the average of the correlation coefficients of the intrapersonal items for the upper group was (0.011), while the average correlation coefficients of the test items for the lower group was (0.042), while the average Correlation coefficients between items for the sample as a whole (0.19). It is evident from the above that the value of the average correlation coefficients between the items of the upper group and the lower group is lower than the average value of the correlation coefficients between the items for the sample as a whole. We also note that all the correlation coefficients are close to zero, which indicates the realization of the assumption of local independence of the test items.

Third: Freedom from speediness: This was done by giving students sufficient time to answer as the test form was not linked to a specific time. By achieving the previous three conditions, the assumptions of the IRT are realized on the data derived from the application of the test items to the sample, which allows the use of the 3PL model.

Statistical treatment

The statistical methods and programs used in the study have been identified in light of the study’s problem and its objectives, in order to answer the study’s questions. The statistical methods and programs used in this study are as follows:

- Calculation of some descriptive statistics, a factor analysis of test items, and nonparametric statistics by SPSS (25)
- Estimating Parameters of the Item (Difficulty - Discrimination - Estimating), and estimating the abilities of persons according to the Item Response Theory (IRT) through PARSCALE program

4. Results

First research question: “Is there an effect of the ability estimation methods: It is the MML method, and the EAP method, on the accuracy of estimating the parameters of the items for the three-parameter model?” The following steps were followed:

Difficulty parameter

The means and standard deviations were calculated for the estimates of the mean values of the standard error parameters of the difficulty parameter for each one according to the estimation methods variable: (MML), and (EAP) method on the three-parameter model. Table 2 shows that there are differences in estimating the mean values of the standard error parameters of the difficulty parameter according to the estimation methods variable. The MML method and EAP method on the 3PL Model refer to Wilcoxon Test correlated Samples. To verify these differences, the Wilcoxon Test was performed for correlated samples to test the significance of the differences between the mean values of the standard error parameters of the difficulty parameter according to the three-parameter model as shown in Table 3.

Table 2. Means and the standard deviations of standard error parameters of the difficulty parameter according to the two methods:(MML and EAP)

Item	Estimation method			
	EAP		MML	
	B	b_SE	B	b_SE
2	0.8542	0.0414	1.3153	0.0671
3	0.4187	0.0468	0.6621	0.0655
6	1.2192	0.0551	1.8626	0.0542
8	-1.1232	0.0654	-1.6325	0.0885
9	-0.1638	0.0566	-0.2113	0.0747
12	1.1221	0.0446	1.1019	0.0819
13	0.6663	0.0456	1.0335	0.0676
18	-1.0227	0.0197	-1.4994	0.0795
19	0.3219	0.0426	0.517	0.0583
21	0.1667	0.0502	0.2842	0.0672
22	0.3278	0.0377	1.4256	0.0639
23	0.0456	0.0559	0.1027	0.0747
25	0.2864	0.0463	1.0635	0.0271
M	0.2399	0.0467	0.4634	0.0669
SD	0.7073	0.0110	1.0648	0.0146

Table 3 shows that there are statistically significant differences between the mean of (MML) and (EAP), and in favors of (EAP), where the value of z was (-3.182) and a statistical significance of (0.01), Therefore, the (EAP) method gives the highest accuracy in estimating the standard error compared to the (MML) method as the value of the standard error of the item was low.

Table 3. Wilcoxon Test to detect differences between mean values of the standard error parameters of item difficulty according to estimation methods

Method (I)	Method (J)	M	Total	Z - value	Significance
MML	EAP	0.00 7.00	0.00 91.00	-3.182	.001

Discrimination parameter

The Means and standard deviations were calculated for the estimates of the mean values of the standard error parameters of the discrimination parameter for each one according to the estimation methods variable: (MML) and (EAP) on the 3PL model. Table 4 shows that there are differences in estimating the mean values of the standard error parameters of the discrimination parameter according to the variable of estimation methods: MML and EAP on 3PL model. To verify these differences, the Wilcoxon Test was performed for correlated samples to test the significance of the differences between the mean values of the standard error parameters of the discrimination parameter according to the 3PL model marked as shown in Table 5.

Table 4. Means and the standard deviations for the estimates of the mean values of the standard error parameters of the discrimination parameter according to the two methods: (MML) and (EAP)

Item	Estimation method			
	EAP		MML	
	A	a_SE	A	a_SE
2	1.3334	0.0885	0.8892	0.0823
3	1.0155	0.0811	0.6772	0.0821
6	1.0513	0.0891	0.701	0.0781
8	0.6129	0.0549	0.4601	0.0545
9	0.7535	0.0718	0.5025	0.0939
12	1.4283	0.0927	0.9524	0.0913
13	1.0718	0.0841	0.7147	0.0768
18	0.7753	0.0547	0.9388	0.0545
19	1.1245	0.0763	0.7499	0.0788
21	0.8938	0.0765	0.5988	0.0889
22	1.5305	0.0881	1.0206	0.0855
23	0.7726	0.0771	0.5152	0.0978
25	1.4078	0.0908	0.5170	0.0917
M	1.0655	0.0788	0.7105	0.0811
SD	0.2856	0.0124	0.19051	0.0134

Table 5 shows that there are no statistically significant differences between the mean values of the standard error of the estimation for each of the two methods (MML) and the (EAP) method, where the value of z is (-.385) and in statistical significance is (0.701), and therefore both methods give the same accuracy in estimating the standard error of the discrimination parameter.

Table 5. Wilcoxon Test to detect differences between the mean values of the standard error to distinguish the item according to estimation methods

Method (I)	Method (J)	M	Total	Z -value	Significance
MML	EAP	10.20	51	-.385	.701

Guessing parameters

The arithmetic means and standard deviations were calculated for the estimates of the mean values of the standard error parameters of the guessing parameter for each one according to the estimation methods variable: The (MML) and (EAP) on 3PL model.

Table 6 shows that there are differences in estimating the mean values of the standard error parameters of the guessing parameter according to the variable of estimation methods: MML and EAP on the 3PL model. To verify these differences, the Wilcoxon Test was performed for correlated samples to test the significance of the differences between the mean values of the standard error parameters of the Guessing parameter according to the 3PL model, using the SPSS program as shown in Table 7.

Table 6. Means and the standard deviations for the estimates of the mean values of the standard error parameters of the Guessing parameter according to the two methods: (MML) and (EAP)

Item	Estimation method			
	EAP		MML	
	C	C_SE	C	C_SE
2	0.2166	0.0521	0.2166	0.0508
3	0.2495	0.0642	0.2495	0.0676
6	0.2003	0.0489	0.2003	0.0478
8	0.2566	0.1096	0.2566	0.1519
9	0.2561	0.0807	0.2561	0.0943
12	0.1859	0.0453	0.1859	0.0432
13	0.2182	0.0574	0.2182	0.0586
18	0.2562	0.1073	0.2562	0.1514
19	0.2432	0.0647	0.2432	0.0689
21	0.2559	0.0711	0.2559	0.0787
22	0.1908	0.0481	0.1908	0.0462
23	0.2575	0.0755	0.2575	0.0852
25	0.1659	0.0429	0.1659	0.0409
M	0.2271	0.0667	0.2271	0.0758
SD	0.0325	0.0218	0.0325	0.0375

Table 7 show that there are no statistically significant differences between the mean values of the standard error of the estimation for each of the two methods (MML) and the (EAP), where the value of z is (-.843) and a statistical significance is (0.046), and therefore both methods give the same accuracy in estimating the standard error of the guessing parameter.

Table 7. Wilcoxon Test to detect differences between the mean values of the standard error of the item Guessing parameter according to the estimation methods

Method (I)	Method (J)	M	Total	Z -value	Significance
MML	EAP	8.43	59	-.843	0.068

Second research question: "Is there an effect of the ability estimation methods: the MML method and the EAP method on the accuracy of estimating the abilities of persons for the 3PL model?" The following steps were taken:

The means and standard deviations were calculated for the estimates of the mean values of the standard error parameters of the ability of persons according to the variable of estimation methods: (MML) and (EAP) on the 3PL model.

Table 8 shows that there are differences in estimating the mean values of the standard error parameters of the abilities of persons according to the variable of estimation methods: The (MML) and (EAP) on 3PL model. Since the assumption of moderate distribution was not fulfilled in the study data, as the two statistical test values using the Kolmogorov-Smirnov test were (0.153-0.037) for the two methods MML and EAP respectively and at a significance level of (0.000), so the Wilcoxon Test was used to test the significance of the differences between the mean values of the standard error parameters of the abilities of persons according to the three-parameter model, as shown in Table 9.

Table 8. Mean and the standard deviation of the estimates of the mean values of the standard error parameters of the abilities of persons according to the two methods: (MML) and (EAP)

Estimation method	No.	M	SD
MML	507	0.5181	0.0277
EAP	507	0.4214	0.3219

Table 9 shows that there is a statistically significant difference between the averages of (MML) and (EAP) in favor of the method of (EAP), where the value of z was (-2.579) and a statistical significance of (.0000). The (EAP) method gives the highest accuracy in estimating the standard error compared to the (MML) method as the value of the standard error of the item is low.

Table 9. Wilcoxon Test to detect differences between mean standard error values of persons' abilities according to estimation methods

Method (I)	Method (J)	M	Total	Z - value	Significance
MML	EAP	309.17	41119.50	-2.579	0.000

5. Conclusion

It is noted from the previous results that there are statistically significant differences between (MML) and (EAP) in favour of (EAP) when estimating the parameter of difficulty and estimating the abilities of persons. These results can be interpreted in according to the mathematical structure of both methods, where we find that the methods that depend on the Bayes theory have made improvements in the mathematical equations used in estimating the ability compared to the methods that depend on ML, in order to process the problems related to estimation when using MML, as the method of (EAP) does not use (Iteration), as it depends on the use of the standard normal distribution, to divide the values of the Loading characteristic usually into (61) splits with periods of length (1.0), which gives the estimates of this method more accuracy which is reflected in the decrease in the value of the standard error of the estimation.

6. References

[1] S.-S. Lee and J. Kim, "An Exploratory study on Student-Intelligent Robot Teacher relationship recognized by Middle School Students," J. Digit. Converg., vol. 18, no. 4, pp. 37–44, 2020.

[2] S. Tibi, A. A. Edwards, C. Schatschneider, L. J. Lombardino, J. R. Kirby, and S. H. Salha, “IRT analyses of Arabic letter knowledge in Kindergarten,” Read. Writ., pp. 1–26, 2020.

[3] R. K. Hambleton and W. J. Van der Linden, Advances in item response theory and applications: An introduction. Sage Publications Sage CA: Thousand Oaks, CA, 1982.

[4] F. M. Lord, Applications of item response theory to practical testing problems. Routledge, 2012.

[5] W. Ma, N. Minchen, and J. de la Torre, “Choosing between CDM and unidimensional IRT: The proportional reasoning test case,” Meas. Interdiscip. Res. Perspect., vol. 18, no. 2, pp. 87–96, 2020.

[6] G. C. Foster, H. Min, and M. J. Zickar, “Review of item response theory practices in organizational research: Lessons learned and paths forward,” Organ. Res. Methods, vol. 20, no. 3, pp. 465–486, 2017.

[7] R. Liu, A. C. Huggins-Manley, and O. Bulut, “Retrofitting diagnostic classification models to responses from IRT-based assessment forms,” Educ. Psychol. Meas., vol. 78, no. 3, pp. 357–383, 2018.

[8] C. L. Azevedo, D. F. Andrade, and J.-P. Fox, “A Bayesian generalized multiple group IRT model with model-fit assessment tools,” Comput. Stat. Data Anal., vol. 56, no. 12, pp. 4399–4412, 2012.

[9] G. Rajlic, “Violations of unidimensionality and local independence in measures intended as unidimensional: assessing levels of violations and the accuracy in unidimensional IRT model estimates,” PhD Thesis, University of British Columbia, 2019.

[10] C. DeMars, “Group differences based on IRT scores: Does the model matter?,” Educ. Psychol. Meas., vol. 61, no. 1, pp. 60–70, 2001.

[11] J. Chen and J. Choi, “A comparison of maximum likelihood and expected a posteriori estimation for polychoric correlation using Monte Carlo simulation,” J. Mod. Appl. Stat. Methods, vol. 8, no. 1, p. 32, 2009.

[12] K. Matlock Cole and I. Paek, “PROC IRT: A SAS procedure for item response theory,” Appl. Psychol. Meas., vol. 41, no. 4, pp. 311–320, 2017.

[13] D. Almaleki, “Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Response Theory (IRT),” Eng. Technol. Appl. Sci. Res., vol. 11, no. 2, pp. 6889–6901, 2021.

[14] D. Almaleki, “Empirical Evaluation of Different Features of Design in Confirmatory Factor Analysis,” 2016.

[15] K. K. Tatsuoka, “Rule space: An approach for dealing with misconceptions based on item response theory,” J. Educ. Meas., pp. 345–354, 1983.

[16] M. Wu and R. Adams, Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions Melbourne, 2007.

[17] G. H. Fischer and I. W. Molenaar, Rasch models: Foundations, recent developments, and applications. Springer Science & Business Media, 2012.

[18] B. Zhuang, S. Wang, S. Zhao, and M. Lu, “Computed tomography angiography-derived fractional flow reserve (CT-FFR) for the detection of myocardial ischemia with invasive fractional flow reserve as reference: systematic review and meta-analysis,” Eur. Radiol., vol. 30, no. 2, pp. 712–725, 2020.

[19] D. R. Divgi, “A minimum chi-square method for developing a common metric in item response theory,” Appl. Psychol. Meas., vol. 9, no. 4, pp. 413–415, 1985.

[20] F. M. Lord, “Maximum likelihood and Bayesian parameter estimation in item response theory,” J. Educ. Meas., pp. 157–162, 1986.

[21] G. L. Candell and F. Drasgow, “An iterative procedure for linking metrics and assessing item bias in item response theory,” Appl. Psychol. Meas., vol. 12, no. 3, pp. 253–260, 1988.

[22] G. J. Mellenbergh, “Item bias and item response theory,” Int. J. Educ. Res., vol. 13, no. 2, pp. 127–143, 1989.

[23] C. J. Maas and J. J. Hox, “Sufficient sample sizes for multilevel modeling,” Methodology, vol. 1, no. 3, pp. 86–92, 2005.

[24] I. Paek, M. Cui, N. Öztürk Gübeş, and Y. Yang, “Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods,” Educ. Psychol. Meas., vol. 78, no. 4, pp. 569–588, 2018.

[25] L. R. Bonetto, J. S. Crespo, R. Guégan, V. I. Esteves, and M. Giovanela, “Removal of methylene blue from aqueous solutions using a solid residue of the apple juice industry: full factorial design, equilibrium, thermodynamics and kinetics aspects,” J. Mol. Struct., vol. 1224, p. 129296, 2021.

[26] K. M. Marcoulides, N. Foldnes, and S. Grønneberg, “Assessing model fit in structural equation modeling using appropriate test statistics,” Struct. Equ. Model. Multidiscip. J., vol. 27, no. 3, pp. 369–379, 2020.

[27] I. W. Molenaar, “Some background for item response theory and the Rasch model,” in Rasch models, Springer, 1995, pp. 3–14.

[28] C. E. Cantrell, “Item Response Theory: Understanding the One-Parameter Rasch Model.,” 1997.

[29] C. Magno, “Demonstrating the difference between classical test theory and item response theory using derived test data,” Int. J. Educ. Psychol. Assess., vol. 1, no. 1, pp. 1–11, 2009.

[30] S. E. Embretson and S. P. Reise, Item response theory. Psychology Press, 2013.

[31] T. Strachan et al., “Using a Projection IRT Method for Vertical Scaling When Construct Shift Is Present,” J. Educ. Meas., 2020.

[32] W.-C. Lee, S. Y. Kim, J. Choi, and Y. Kang, “IRT Approaches to Modeling Scores on Mixed-Format Tests,” J. Educ. Meas., vol. 57, no. 2, pp. 230–254, 2020.

[33] T. Strachan, E. Ip, Y. Fu, T. Ackerman, S.-H. Chen, and J. Willse, “Robustness of projective IRT to misspecification of the underlying multidimensional model,” Appl. Psychol. Meas., vol. 44, no. 5, pp. 362–375, 2020.

[34] D. R. Crişan, J. N. Tendeiro, and R. R. Meijer, “Investigating the practical consequences of model misfit in unidimensional IRT models,” Appl. Psychol. Meas., vol. 41, no. 6, pp. 439–455, 2017.

[35] M. N. Morshed, M. N. Pervez, N. Behary, N. Bouazizi, J. Guan, and V. A. Nierstrasz, “Statistical modeling and optimization of heterogeneous Fenton-like removal of organic pollutant using fibrous catalysts: a full factorial design,” Sci. Rep., vol. 10, no. 1, pp. 1–14, 2020.

[36] M. L. Stocking and F. M. Lord, “Developing a common metric in item response theory,” Appl. Psychol. Meas., vol. 7, no. 2, pp. 201–210, 1983.

[37] H. Swaminathan, R. K. Hambleton, and J. Algina, “Reliability of criterion-referenced tests: A decision-theoretic formulation,” J. Educ. Meas., vol. 11, no. 4, pp. 263–267, 1974.

[38] R. K. Hambleton and H. J. Rogers, “Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods,” Appl. Meas. Educ., vol. 2, no. 4, pp. 313–334, 1989.

[39] H. Kishino, T. Miyata, and M. Hasegawa, “Maximum likelihood inference of protein phylogeny and the origin of chloroplasts,” J. Mol. Evol., vol. 31, no. 2, pp. 151–160, 1990.

[40] R. G. Lim and F. Drasgow, “Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning.,” J. Appl. Psychol., vol. 75, no. 2, p. 164, 1990.

[41] H. Kishino and M. Hasegawa, “Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea,” J. Mol. Evol., vol. 29, no. 2, pp. 170–179, 1989.

[42] H. Swaminathan, R. K. Hambleton, and H. J. Rogers, “21 Assessing the Fit of Item Response Theory Models,” Handb. Stat., vol. 26, pp. 683–718, 2006.

[43] H. Swaminathan, R. K. Hambleton, S. G. Sireci, D. Xing, and S. M. Rizavi, “Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates,” Appl. Psychol. Meas., vol. 27, no. 1, pp. 27–51, 2003.

[44] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of item response theory, vol. 2. Sage, 1991.

[45] C. S. Wardley, E. B. Applegate, A. D. Almaleki, and J. A. Van Rhee, “A comparison of Students’ perceptions of stress in parallel problem-based and lecture-based curricula,” J. Physician Assist. Educ., vol. 27, no. 1, pp. 7–16, 2016.

[46] D. Almaleki, “Stability of the Data-Model Fit over Increasing Levels of Factorial Invariance for Different Features of Design in Factor Analysis,” Eng. Technol. Appl. Sci. Res., vol. 11, no. 2, pp. 6849–6856, 2021.

[47] D. Almaleki, “The Precision of the Overall Data-Model Fit for Different Design Features in Confirmatory Factor Analysis,” Eng. Technol. Appl. Sci. Res., vol. 11, no. 1, pp. 6766–6774, 2021.

[48] C. S. Wardley, E. B. Applegate, A. D. Almaleki, and J. A. Van Rhee, “Is Student Stress Related to Personality or Learning Environment in a Physician Assistant Program?,” J. Physician Assist. Educ., vol. 30, no. 1, pp. 9–19, 2019.

[49] J. Y. Park, F. Cornillie, H. L. van der Maas, and W. Van Den Noortgate, “A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments,” Front. Psychol., vol. 10, p. 620, 2019.

[50] G. J. Mellenbergh, “Item bias and item response theory,” Int. J. Educ. Res., vol. 13, no. 2, pp. 127–143, 1989.

[51] R. K. Hambleton and A. Kanjee, “Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations,” Eur. J. Psychol. Assess., vol. 11, no. 3, pp. 147–157, 1995.

[52] R. K. Hambleton, W. J. van der Linden, and C. S. Wells, “IRT models for the analysis of polytomously scored data,” Handb. Polytomous Item Response Theory Models, pp. 21–42, 2010.

[53] M. A. Tanner and W. H. Wong, “The calculation of posterior distributions by data augmentation,” J. Am. Stat. Assoc., vol. 82, no. 398, pp. 528–540, 1987.

[54] E.-Y. Mun, Y. Huo, H. R. White, S. Suzuki, and J. de la Torre, “Multivariate higher-order IRT model and MCMC algorithm for linking individual participant data from multiple studies,” Front. Psychol., vol. 10, p. 1328, 2019.

[55] S.-K. Chen, L. Hou, and B. G. Dodd, “A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model,” Educ. Psychol. Meas., vol. 58, no. 4, pp. 569–595, 1998.

[56] D. A. Almaleki, W. W. Khayat, T. F. Yally, and A. A. Al-hajjaji, “The Effectiveness of the Use of Distance-Evaluation Tools and Methods among Students with Learning-Difficulties from the Teachers’ Point of View,” Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 5, pp. 243–255, May 2021, doi: 10.22937/IJCSNS.2021.21.5.34.

[57] D. A. Almaleki, R. A. Alhajaji, and M. A. Alharbi, “Measuring Students’ Interaction in Distance Learning Through the Electronic Platform and its Impact on their Motivation to Learn During Covid-19 Crisis,” Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 5, pp. 98–112, May 2021, doi: 10.22937/IJCSNS.2021.21.5.16.

[58] D. A. Almaleki, “The Psychometric Properties of Distance-Digital Subjective Happiness Scale,” Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 5, pp. 211–216, May 2021, doi: 10.22937/IJCSNS.2021.21.5.29.

[59] D. A. Almaleki, “Challenges Experienced Use of Distance-Learning by High School Teachers Responses to Students with Depression,” Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 5, pp. 192–198, May 2021, doi: 10.22937/IJCSNS.2021.21.5.27.

[60] A. Preti, M. Vellante, and D. R. Petretto, “The psychometric properties of the ‘Reading the Mind in the Eyes’ Test: an item response theory (IRT) analysis,” Cognit. Neuropsychiatry, vol. 22, no. 3, pp. 233–253, 2017.

[61] J. J. Hox, C. J. Maas, and M. J. Brinkhuis, “The effect of estimation method and sample size in multilevel structural equation modeling,” Stat. Neerlandica, vol. 64, no. 2, pp. 157–170, 2010.

[62] R. K. Hambleton and R. W. Jones, “Comparison of classical test theory and item response theory and their applications to test development,” Educ. Meas. Issues Pract., vol. 12, no. 3, pp. 38–47, 1993.

Deyab A. Almaleki is an Associate Professor in the Department of Evaluation, Measurement and Research. Dr. Almaleki received his Ph.D. from Western Michigan University (USA) in 2016 in Evaluation, Measurement and Research. Since 2011, Dr. Almaleki has authored and co-authored in more than 20 peer-reviewed journal articles, and over 30 peer-reviewed presentations at professional conferences. Dr. Almaleki has extensive experience in educational and psychological research, research design, applied statistics, structural equation modeling, design and analysis of psychological measurement.

Ahoud Ghazi Alomrany is a researcher in Evaluation, Measurement and Research - Umm Al-Qura University.