Ilkogretim Online - Elementary Education Online, 2021; 20 (1): pp. 887-898
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.82

The Effect of the Percentage of Missing Data on Estimating the Standard Error of the Items' Parameters and the Test Information Function According to the Three-Parameter Logistic Model in the Item Response Theory

***Dr. Habis Saad Al-zboon,** *Department of Curriculum and Instruction, College of Education -Al-Hussein Bin Talal University, Habis.s.alzboon@ahu.edu.jo*
Dr. Mo'en Salman Alnasraween, *Department Educational psychology Amman Arab University , mueen@aau.edu.jo*
Dr. Taha Oklah Alkursheh , *University of Tabuk, taha_oglah@yahoo.com*
*Corresponding Author

Abstract. This study aimed to investigate the effect of the percentage of missing data on estimating the standard error of the items' parameters and the test information function based on the three-parameter model using generated data. To achieve the objectives of the study, (1000) examinees' responses were generated on a test consisting of (50) dichotomous items according to the three-parameter logistic model(3PL) using (WINGEN3) software. By using (SPSS) and (EXCEL), data containing missing responses were obtained at the percentages of (0%, 5%, 10%, 15%). Four files were prepared that include the missing four percentages. The items' parameters for these files were identified, , the items and individuals were fitted for (3PL-IRT), number of persons were removed.
The results of the study showed that there are statistically significant differences between the means of the standard errors for estimating the difficulty and the discrimination parameter of the test based on the three-parameter model according to the difference in the missing data percentages of (0%, 5%, 10%, 15%) and in favor of the lesser missing percentages, which means that the fewer missing percentage**s** are, it is expected that the standard error value will decrease in estimating the difficulty and the discrimination parameter of the item, and that the best missing data percentage was (5%). The results also indicated that the information function increases as the missing data percentage decreases.

Keywords: Missing Data, Standard Error, Test Information Function

Received: 20.11.2020	Accepted: 12.12.2020	Published: 01.01.2021

INTRODUCTION

Tests are considered one of the most important, reliable, and common methods to measure student achievement. Test is a measurement tool that is prepared according to an organized method of several steps that include a set of procedures subject to specific conditions and rules (Odeh, 2010).

The use of school tests has spread widely in many areas, as these tests are designed for various purposes such as selecting a person for a job from among a group of applicants for this job, or for classification purposes such as determining the students' path in proportion to their abilities and skills, and in evaluating students' achievement through the achieved grades which they obtain in class tests, and many more. The process of evaluating individuals acquires great importance as the importance of the decisions based on which the process of evaluation is based, and as much as the seriousness of the wrong decisions that may result from this in various situations and fields, whether at the level of the individual or society, to an extent that is difficult to address, or may need treatment for a long time, which hinders the development process and coping with the development of other societies. To obtain more accurate decisions, valid and accurate information must be provided through good planning and preparation for the test (Al-Sharifain & Taamana, 2009)

The test makers face the problem of missing data when implementing tests, as most researchers face missing data problem. It is expected that the effect of the proportion of missing data on the test specifications

and the estimation of parameters will vary. In addition, ignoring the handling of values may negatively affect the specifications of the tests which affects the accuracy of decisions made based on the results of these tests. The missing data could be attributed to a wide range of reasons, some of which could be partially controlled by the researcher, and these reasons could be classified into three categories as follows: (1) Study participants: participants may not answer some because they may feel sick and tired. (2) Design specifications: the study may require a lot of time or lack of clarity (3) or the interaction between them, that is, between the participants' characteristics and the design specifications, as some participants do not answer the long or tiresome due to an illness. There are many problems associated with theoretically missing data and matching them with practical solutions, including that the samples are not representative, or that the available data reflect bias, which leads to biased estimates and misleading statistical conclusions (Mcknight, 2007).

The item response theory has contributed to solving most of the deficiencies in classical test theory. It has provided methods for selecting by presenting fixed parameters for (difficulty parameter, discrimination parameter, guessing parameter). In this theory, it is possible to link between the item characteristic and the measured ability on the one hand, and the probability of a correct answer, on the other hand through several mathematical models, including the three logistical models: the one-parameter logistic model, the two-parameter logistic model, and the three-parameter logistic model.

Little & Rubin (2002) showed that the data collected from the responses of individuals are significantly affected because a number of them did not respond to a number of the measurement tool, regardless of the reason for not responding, and this leads to the existence of missing values or data, and thus obtaining missing data that may affect the item parameters and thus the effectiveness of interpretations. Therefore, this study emanated to demonstrate the effect of the percentage of the missing data on estimating the standard error of the items' parameters and the test information function according to the three-parameter model.

Study problems

Test makers face a problem when they analyze the data, which is the incomplete answers of the respondents on some test , which may affect the characteristics of the parameters of the, the specifications of the test and the abilities of the respondents, and thus affect the decisions that will be taken based on the results of these tests. As an increase in the percentage of missing data may affect the results of these tests, and given the importance of decisions that are based on the results of the tests, if the purpose of the tests is to obtain important information, but it may be inaccurate due to the existence of missing data, therefore, the need for this study emanated to demonstrate the effect of the percentage of the missing data on estimating the standard error of the items' parameters and the test information function according to the three-parameter model.

Study questions

1. Are there statistically significant differences at the level of significance ($\alpha \leq 0.05$) between the means of the standard errors for estimating the item parameters (difficulty, discrimination, guessing) based on the three-parameter model with different percentages of missing data (0%, 5%, 10%, 15%)?
2. Does the test information function differ according to the different percentages of missing data (0%, 5%, 10%, 15%)?

Objectives of the study

This study aims to explore the effect of the percentages of the missing data on estimating the standard error of the items' parameters and the test information function according to the three-parameter model.

The importance of the study

The importance of the study stems from clarifying the concept of missing data and the effect of the percentage of the missing data on estimating the standard error of the items' parameters and the test information function according to the three-parameter model. This study could contribute to encourage researchers to conduct similar studies in the future and on new variables.

Definition of Terms: (Hambleton, Swaminthan, 1985)

Missing Data: Failure to respond to some of the test items by the respondent, leaving these empty without an answer.
Item parameters: item difficulty (threshold), item discrimination (slope), Item guessing Asymptote.
Item difficulty (Threshold): represents the point on the ability scale at which an examine has a 50 percent probability of answering items correctly when the guessing coefficient is equal to zero, the values of b very (typically) from about -2 to +2. Values of b_i near -2 correspond to items that are very easy, and values of b_i near +2 correspond to items that are very difficult for the group of examinees.
Item discrimination (Slope): is proportional to slope of $P_i(\theta)$ at the point θ = b_i the values of item discrimination Coefficient a_i very (typically) from about -2 to +2 , and the usual range for item discrimination parameters is (0,2), The discrimination factor values can be categorized as follows In (Baker, 2001) to the following:

Table 1. *Classification of the item discrimination coefficient*

Discrimination Coefficient a_i	Category
≤ 0	No discrimination
0.01-0.34	Very low
0.35-0.64	low
0.65-1.34	Medium
1.35-1.69	High
≥ 1.7	Very high

Item guessing (asymptote): The lower asymptote of the item characteristic curve which represents the probability that the examinee with low ability will answer the item correctly.
Test information function: The sum of the items' functions that make up the test.
Generated Data: The data used in this study generated by the data generation program (WINGEN) according to the three-parameter model.
Standard Error: It is the standard deviation of the error in the estimation of a parameter (Hambleton, Swaminathan & Rogers, 1991).

Theoretical framework and previous studies

Missing data is one of the most important problems that researchers face when collecting or analyzing data, and this problem arises from the moment of preparing and designing the test and during the application until the collection and correction of the response. The ideal case is the response of all members of the sample to all test items, and this ideal case does not appear in most research. Data collected are often incomplete and lead to biased and less efficient estimates (Litle & Rubin, 1987).

Huisman and Van Sonderem (1998) indicated that the missing data resulting from the respondent can be classified into two types. The first type is related to the complete failure to respond to all items, so that the examinee does not respond to any of the test items. While the second type relates to the non-response to the item, meaning that the respondent takes part in the test and answers some items and leaves some unanswered, which leads to incomplete partial data due to the missing of some of them, and this includes: The respondent skipping the item (i.e. leaving it without an answer) because he skipped it unintentionally or because the time is insufficient to answer, or because he does not know the answer for some reason, or because the item is difficult, or because the length of the test exhausts the respondents and affects their concentration. There are several ways to deal with missing values, as the researcher's knowledge of the pattern on which the missing values appear, as well as his knowledge of the mechanism of the missing of values, helps him choose the appropriate way to deal with the missing values. Little and Rubin (2002) distinguished between three types of the patterns of missing of values: arbitrary pattern, univariate pattern, and monotone pattern (Al-Zou'bi, 2013).

The use of Item Response Theory in analyzing psychological and educational test data is a solution to avoid most of the shortcomings of the Classical Test Theory that dominated the development of tests in the twentieth century.

Classical test theory (CTT) has been mainstay of psychological test development for must of 20th century. Guliksen's (1950) classic book, which remains in print , is often cited as the defining volume, however , CTT is much older. Many procedure were pioneered by spearman (1907,1913). CTT has defined the standard test development, beginning with the initial explosion of testing in 1930s.

However, since lord and Novick's (1968) classic book introduced model-based measurement, a quiet revolution has occurred in test theory. Item response theory (IRT) has rapidly become mainstream as a basis for psychological measurement. IRT, also known as latent trait theory, is model-based measurement in which trait level estimates depend on both person's response and on the properties of the items that were administered. Many new or revised tests, particularity ability tests, have been developed from IRT Principles. Yet, because most test users are unfamiliar with IRT, test manuals mention its application only in passing or in a technical appendix. Thus test users are largely unaware that psychometric basis of testing has changed.

Yet, in the new model-based version of test theory, IRT some well-known rules of measurement to longer apply. In fact, the new rules of measurement are fundamentally different from the old rules. Many old rules in fact, must be revised, generalized, or abandoned altogether.

One of these rule: the slandered error of measurement, in the old rule the standard error of measurement applies to all scores in particular population, but in the new rule the standard error of measurement differs across scores but generalizes across populations.

In CCT, the standard error of measurement is computed by the equation:

$$s_e = \sigma\sqrt{1 - r_{tt}}$$

Where : σ is slandered deviation , r_{tt} is reliability

Confidence intervals are constructed for individual scores under the assumption that measurement error is distributed normally and equally for all score levels.

In IRT models trait's scores are estimated separately for each score or response pattern, controlling for the characteristics of the items that were administered. Standard errors are smallest when the items are optimally appropriate for a particular trait score level and when item discrimination are high. (Embretson & Reise, 2000)

Item Response Theory is based on a set of assumptions that must be achieved in the data in order to obtain accurate results. The most important of these assumptions is the unidimensionality, which means that there is one characteristic that explains the examinee's performance on the test, (i.e. that the examinee's score on the test reflects the characteristic that the test measures only). There are different statistical methods used to examine the achievement of data for this assumption, the most important of which is the method of factor analysis. The second assumption is called the local independence. Achieving this assumption requires that the responses of the respondents to the test should be statistically independent at a certain ability level. In other words, that the examinee's response to one item does not positively or negatively affect his answer to the other items (Crocker & Algina, 1986). This assumption is only true in the case that the test is unidimensional, which means that the performance of individuals, of a certain ability level, on one item is not affected by their performance on another item (Hambleton & Swaminathan, 1985).

While the third assumption is called the Item Characteristics Curve (ICC). The concept of the curve of Item Response Theory is a mathematical association that relates the probability of the success of the examinee on the item with the ability measured by a group of items composing the test. The last assumption is called speediness, which assumes that the speed factor does not play a role in answering the test, i.e. the wrong answer of the item is caused by the ability and not by the time allotted for the test (Allam, 2005).

In this theory, it is possible to relate the characteristics of the items with the measured ability, on the one hand, and with the probability of the correct answer, on the other hand through several mathematical models, including the three logistic models: the one-parameter logistic model, the two-parameter logistic model, and the three-parameter logistic model. (Hambelton, Swaminathan & Rogers, 1991; Embretson & Reise, 2000)

One-Parameter Logistic Model (1PLM)

The One-Parameter Logistic Model, also known as the Rasch Model, is one of the broadest models used in the Item Response Theory, which assumes that all items do not differ from each other except with the difficulty parameter of the item (b_i). It also assumes that the discrimination parameter (a_i) is equal for all items, while the guessing parameter (c_i) for items approaches zero, and the model takes the following mathematical formula to express the probability of the correct answer to the item:

$$p_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}} \qquad i = 1, 2, 3....., n$$

Two-Parameter Logistic Model (2PL)

The one-parameter logistic model is a special case of the two-parameter logistic model, as this model assumes that the items differ in the difficulty and discrimination parameters, while the guessing parameter approaches zero, and the probability of a correct answer to the item is given by the following mathematical formula:

$$p_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

Three Parameter Logistic Model- 3PLM

The Three Parameter Logistic Model can be obtained from the two- parameter model by adding a third parameter, denoted ci, the mathematical form of the Three Parameter Logistic curve is written:

$$p_i(\theta) = c_i + (1-c_i)\frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

$P_i(\theta)$:the probability that an examinee with ability θ answers item i correctly .

b_i : item difficulty parameter , a_i : item discrimination parameter , c_i item guessing parameter, θ : trait level for person i.

e= The natural logist foundation that equals 2.718 D: The scaling factor equal to1.7

Information Function

The information function is considered one of the basic concepts that play an important role in the Item Response Theory. The information function determines the amount of information provided by the item or the test as a whole when estimating the ability of individuals or respondents, and through which the standard error in the estimation could be determined.

The information function, provided by both the item and the test, is of great importance in the Item Response Theory, the items can be chosen using the item information function, given that the item information changes across different levels of ability, and thus it is possible to choose items that provide high measurement accuracy at a certain point on the characteristic continuum, on the one hand. On the other hand, the items with high discrimination contribute to providing more information about the ability of the respondent and thus obtaining greater accuracy. Hence, the largest value of the information is fixed in the case of using the one-parameter model, and it is directly proportional to the item discrimination square for the two-parameter logistic model (2PL). The test items tend to contribute better to the accuracy of the measurement in the one-and-two-parameter models, about the difficulty parameter (b) on the ability scale, and the amount of information is largest when the ability level (Θ) is close to the difficulty parameter (b), because the information function takes a form that approximates the bell shape in general, while the highest value is obtained if using the three-parameter model at the level of (Θmax) ability (Hambelton & Swaminathan, 1985; Al-Zou'bi, 2013).

Baker (2001) mentioned different formulas of the item information function equation according to the three models of Item Response Theory. In the three-parameter model, the value of the test information function is calculated at each ability level (Θ) through the following equation:

$$I_i(\theta) = a^2 \left[\frac{Q_i(\theta)}{P_{ii}(\theta)} \right] \left[\frac{P_i(\theta) - c}{1 - c} \right]^2 \qquad Q_i(\theta) = 1 - P_i(\theta)$$

The data collected from the responses of the respondents is greatly affected by the failure of a number of them to respond to a number of items of the measurement tool; regardless of the reason of the non-response, and this leads to the existence of missing data, and thus obtaining incomplete data that affect the effectiveness of the interpretations (Little & Rubin, 2002) .

Missing values are a common problem in many measurement situations, and researchers have shown that these missing values create problems in estimating item parameters within the context of Item Response Theory, particularly if these missing values are ignored. Therefore, if there are missing data, it is important that we address them (Al-Zou'bi, 2013).

Previous Studies

Langkamp and Lemeshow (2010) conducted a study aimed at determining which method of dealing with missing values gives more accurate estimates of the parameters of the model used in child health surveys, in cases involving different proportions of missing values. To achieve the aim of the study, data were generated containing the missing values of four percentages (10%, 20%,30%, 40%) of the cases that were generated.

The study sample consisted of (18238) examinees, and the results of the study showed that when the missing values percentage exceeds (10%), the methods of reweighting and multiple imputation were much better than the methods of (Case-Deletion, and the Hot-Deck imputation).

Bani Awwad (2010) conducted a study aimed at examining the effect of a number of methods of processing missing data on the accuracy of estimating the item parameters and the ability of individuals. To achieve the objectives of the study, the test of Otis-Lennon mental ability test of (80) items was implemented on a sample formed of (1600) students randomly selected from eighth grade students in Irbid Governorate. The study found statistically significant differences between the means of the standard errors for estimating the ability parameter of individuals, as well as the existence of statistically significant differences between the means of the standard errors for estimating the parameter of the item (difficulty, discrimination, and guessing) in favor of the response function method.

In the study of Cokluk and Kayri (2011), which aimed to examine and compare the coefficients of reliability: Corrected Item-Total Correlation, Cronbach-Alpha Coefficient of Internal Consistency and Factor Structures, resulting from implementing five approaches to dealing with missing values, using different missing percentages. In order to achieve the aim of the study, the missing percentages were divided into two categories: the first and ranged between (15%) and (20%), and the second and ranged between (0%) to (50%), in order to examine the construct validity of the scale.

The study was considered as a comparative study to find the results of the factor analysis based on the Principal Component Analysis Method, which is used in determining the factor structures of the scale when using imputation methods that were dealt with in the study. The study sample consisted of (200) examinees. Regarding the validity of constructing the scale, the results of the study showed that the different methods of manipulating the missing values cause a decrease in the variance percentages explained for the methods used in the study. As for the eigenvalues and the coefficients of Cronbach-Alpha of the internal consistency of the scale, the results showed a similar decrease in the percentages of the explained variance.

Gemici, Bendnarz and Lim (2012) conducted a study aimed at investigating the effectiveness of a number of methods of processing missing data. To achieve its goals, real data related to individual responses on a scale related to education and vocational training were used. The study concluded with a number of results, the most important of which was that the multiple imputation method (MI) helps in reducing the problem of biasness and also reduces the standard error of the estimation compared to other methods.

Al-Darabseh (2012) conducted a study aimed at demonstrating the effect of the method of dealing with missing values and the method of estimating the abilities of individuals on the accuracy of estimating the item parameters and individuals. In order to achieve the objectives of the study, a test consisting of (80) items was applied to (1500) examinees according to the three-parameter logistic model, where the missing percentage was (5%) of the volume of data, and the missing values were dealt with in three ways: the expectation maximization method, the multiple imputation method, and the item response function method. Two methods were used to estimate the ability, namely the maximum likelihood method and the expected a posteriori method of the estimation of the ability parameter. The results of the study concluded that there

was a difference in the accuracy of the estimation of the discrimination parameter in favor of the multiple imputation method, and there was no difference in the accuracy of the estimation of the difficulty parameter due to the method of manipulating the missing values. They also indicated that there was a difference in the accuracy of estimating the ability and in favor of expectation maximization method.

The study of (Alruhail and Aldrabsah,2014) amid to investigate the effect of ability estimation method and handling method with missing values on the accuracy of items and persons parameters. To achieve this aim, data were generated using (WINGEN) software. (1500) respondent on a test consisted of (80) dichotomous items fitting the three parameters logistic model were generated. Using (SPSS) and (EXCEL) that data with (5%) missing responses were generated, the data was processed through the three handling methods of missing values; Expectation maximizing (EM), Multiple Imputation (MI), and Response function(RF).

The finding showed that there were significant differences in the estimation accuracy of discrimination parameter attributed to estimation method is favor of (ML) and in the difficulty is favor of (EAP) method. Moreover, the study showed that there were no statistically significant differences in the estimation accuracy of item difficulty and guessing parameter attributed to the handling method or the interaction between the dealing method and the estimation method.

The study of Al-lasasmeh (2018) aimed to identify the effect of the sample size and the method of dealing with the missing values on the test reliability and the parameters of discrimination and difficulty of the items. To achieve the goal of the study, three versions of the test were generated with samples of (300, 500, 1000) examinees using the (WINGEN3) software. The responses were distributed on a test of (20) items according to the two-parameter model (2PLM), and the (BILOG-MG3) and (SPSS) programs were relied on to analyze the responses of individuals.

The results of the study showed that there were no differences between the parameters of reliability using different methods of missing data manipulating at all levels of the sample size (300, 500, 1000). They also pointed to the effect of the manipulating method on the discrimination parameter and in favor of the deletion method, and there was no effect for the manipulating method on the difficulty parameter. The results further showed that there was an effect of the sample size on the discrimination parameter, and in favor of the sample size of (500) examinees, and there was no effect on the difficulty parameter owing to the sample size.

Al-Zubi (2013) aimed to investigate the effect of the percentage of missing data and imputation method on the accuracy of estimating parameters of items and persons. To achieve this aim, data were generated using (WENGIN3) software. (1400) respondent on a test consisted of (100) dichotomous items fitting the one and two parameters logistic model were generated with the following ranges of discrimination (0.10-2.0), difficulty (-2.5-2.5), assuming that abilities are distributed normally.

Using (SPSS) and (EXCL) data with (5%,15%,20%,30%) missing responses were generated. The data was processed through the two handling imputation methods of missing values: expectation maximization (EM), and multiple imputation (MI). The data were tested for unidimensionality using factor analysis. The items and individuals were fitted to the used model. The number of items and persons were removed. The responses of 1254 persons for 67 items fitted for (1PL-IRT), the responses of 1365 persons for 77 items fitted for (2PL-IRT), the standard error were estimated through maximum likelihood (ML).

The findings showed that there was a significant difference in the accuracy of estimating of difficulty parameters attributed to the imputation method for (1PL-IRT) in favor of (MI), and the missing percentage for (1PL-IRT) and (2PL-IRT) was in favor of (5%) interaction between imputation method and the missing percentage for (1PL-IRT) and (2PL-IRT) in favor of (MI) with missing percentage (5%) in data.

Moreover, findings showed that there was a significant difference in the accuracy of estimating of discriminate parameters for (2PL-IRT) attributed to the imputation method in favor of (MI) and the missing percentage in favor of (5%) .

The methodology of the study

The experimental simulation approach was used on the examinees in order to identify the effect of the percentage of missing data in estimating the standard error of the item parameters and the test information function according to the three-parameter model.

Stages of data generation

First: Generating the test

The data generated in this study was used using the (WINGEN3) software because of the standard conditions provided by this data that are difficult to obtain in the case of using real data. Lord (1980) recommended that the test should be of (50) items and the number of individuals should be (1000) to obtain the best estimates. Therefore, item parameters of the multiple-choice type suitable for the three-parameter logistic model (3PL) were simulated with a sample size of (1000) individuals. Parameters were simulated according to the following conditions:

1. Simulation of the discrimination parameter for the items according to the log normal distribution ~ (0,0.5) based on the three-parameter model. After generating the data, it was found that the mean of the discrimination parameter and the standard deviation were (mean a = 0.06) and (SD a = 0.48), respectively. This value was considered good by comparing it with the standard defined by Hambleton and Swaminthan (1985), which states that the true discriminant parameter values range from [2, -2] logit.
2. Simulation of the difficulty parameter of the items according to the normal distribution ~ (0, 1) based on the three-parameter model. After generating the data, it was found that the mean of the difficulty parameter and the standard deviation were (mean b = 0.17) and (SD b = 1.07), respectively.
3. Simulation of the guessing parameter of the items according to beta distribution ~ (10, 30) according to the three-parameter model. This distribution simulates the values of the guessing parameter, and the mean and standard deviation of the estimation parameter were (mean c = 0.24) and (SD c = 0.24), respectively.

Second: Generating the responses

Responses of (1000) examinees were generated using the same values for the real parameters of the previously generated items, depending on the normal distribution ~ (0, 1), and the mean and standard deviation of the simulated individuals' abilities were (0.049) and (0.973), respectively.

Third: Goodness of Fit data

First: The goodness of fit of items: The idea of proper matching of the items is based on comparing the expectation of the mathematical model of the respondent score with the apparent response at the level of the scale's items as a whole, or at the level of each item. Several methods work to examine the suitability of the items, including Chi-square (Mislevy & Bock, 1996), and the standardized residual indicator (Masters & Wright, 1996). In this study the value of the standardized residual indicator ranged between (1.09, 0.53) with a range of (0.56), which is not statistically significant, while the value of the Chi-square indicator ranged between (0.46, 0.31) with a range of 0.15. The general standardized residual indicator of the overall scale was (0.67), and the probability value was (0.27), which is a non-statistically significant value, as well as for all items of the scale. This indicates the ability of the model to explain the responses of individuals on the scale items.

Second: The goodness of fit of individuals: Several indicators are used to examine the suitability of individuals for the model, including: the standardized residuals (Masters & Wright, 1982), the probability indicator (Drasgow, et al, 1985), and the sum of squared residuals indicator (Almahrazi, 2003). In this study, the results of the standardized residual indicator (Masters & Wright, 1982) showed that the response patterns of 29 individuals were not appropriate, and they were, therefore, excluded from the study sample.

Fourth: The missing mechanism:

After generating the responses data of the respondents' to the test items, the data were transferred to the (SPSS) file and then to the (Excel) file, after which the missing process was carried out randomly at the percentages of (0%, 5%, 10%, and 15%) to prepare four files containing the missing values.

RESULTS AND DISCUSSION

To answer the first question that was “Are there statistically significant differences at the level of significance ($\alpha \leq 0.05$) between the means of the standard errors for estimating the item parameters (difficulty, discrimination, guessing) based on the three-parameter model with different percentages of missing data (0%, 5%, 10%, 15%)?” Means and standard deviations of the standard errors were calculated for the

estimates of the item parameters (difficulty, discrimination, guessing) based on the three-parameter model according to the difference in the percentages of missing data (0%, 5%, 10%, 15%) as shown in Table (2).

Table 2: *Means and standard deviation of standard error of items parameters*

Percentage of Missing Data	parameters	Mean of standard error of items parameters	Standard deviation of standard error of items parameters
0%	a	0.084	0.01
	b	0.5	0.33
	c	0.01	0.02
5%	a	0.092	0.008
	b	0.46	0.25
	c	0.009	0.002
10%	a	0.10	0.008
	b	0.39	0.21
	c	0.01	0.001
15%	a	0.21	0.05
	b	0.36	0.19
	c	0.009	0.002

It is clear from Table (2) that there are apparent differences between the means of the standard errors for the estimates of the item parameters (difficulty, discrimination, guessing). To verify the significance of the differences, the one-way ANOVA analysis was used to identify the significance of the differences between the means of the standard errors for the estimates of the item parameters (difficulty, discrimination, guessing) based on the three-parameter model according to the difference in the percentages of missing data (0%, 5%, 10%, 15%) as shown in Table (3):

Table 3: *One-way ANOVA of differences between means of standard error of items parameters*

ANOVA						
		Sum of Squares	df	Mean Square	F	Sig.
Standard error of item discrimination	Between Groups	.002	3	.001	.945	.04
	Within Groups	.151	196	.001		
	Total	.153	199			
Standard error of item difficulty	Between Groups	.973	3	.324	5.107	.002
	Within Groups	12.450	196	.064		
	Total	13.423	199			
Standard error of item guessing	Between Groups	.001	3	.000	2.545	.057
	Within Groups	.014	196	.000		
	Total	.015	199			

It is evident from Table (3) that there are statistically significant differences between the means of standard errors for the estimates of the difficulty and discrimination item parameters based on the three-parameter model according to the difference in the percentages of missing data (0%, 5%, 10%, 15%).

It is also evident from the table that there are no statistically significant differences between the means of the standard errors for the estimates of the guessing parameter based on the three-parameter model according to the difference in the percentages of missing data (0%, 5%, 10%, 15%).

To identify the trend of the differences and to which missing percentage those differences could be attributed, Scheffe test for post-hoc comparisons was used as shown in Table (4).

Table 4. *Multiple comparisons between the means of standard error of items parameters*

Percentage of Missing Data	Mean Difference of standard error of item discrimination				Percentage of Missing Data	Mean Difference of standard error of item difficulty			
	0 %	5 %	10 %	15 %		0 %	5 %	10 %	15%
0 %		-0.08	-0.016*	-0.13*	0 %		-0.04*	-0.11*	-0.14*
5 %			-0.008*	-0.22*	5 %			-0.07	-0.10*
10 %				0.32*	10 %				0.03

It is evident from Table (4) that the differences in the estimates of the standard errors related to the estimate of the discrimination parameter of the test items were in favor of the missing percentages of (10%, 15%) compared with the missing percentage of (0%) and in favor of the missing percentages of (10%, 15%) compared to the missing percentage of (5%). The table also shows the preference of the missing percentage of (10%) compared to that of (15%), where the value of the mean for the estimates of the standard errors related to the estimate of the discrimination parameter of the test items is the lowest in the case of the lowest missing percentage , i.e. as the missing percentage decreases, it is expected that the value of the standard error in the estimates of the discrimination parameter for the item will decrease. The reason for this could be attributed to the fact that in the case of reducing the missing data percentage , the value of the discrimination parameter for the items will increase, which positively affects the specifications of the tests and items, which, in turn, affects the accuracy of the decisions taken based on the results of these tests.

Further, it is clear that the differences in the estimates of the standard errors related to the estimate of the difficulty parameter of the test items were in favor of the missing percentage of (5%, 10%, 15%) compared to the missing percentage of (0%) and in favor of the missing percentage of (15%) compared to the loss percentage (5%). It can be inferred that at the missing data level of (5%) and (0%), the mean standard error of the estimates of the difficulty parameter of the item was less compared to the missing percentages of (10%) and (15%). The reason for this may be that at a missing percentage of (5%), the value of the difficulty parameter approaches the value of the ability, which reduces the effect of guessing. Thus, the standard error of the estimates of the value of the difficulty parameter decreases, as the difficulty parameter approaches the missing percentage of (0%, 5%) of (0.5) where Frisbie (1973) (as cited in Fraihat, 2019) indicated that the extent of difficulty affects the discriminatory significance of the test , and that the items of medium difficulty are higher in the case of the very easy or very difficult items, where the amount of information is greater when the ability level (Θ) is close to the difficulty parameter (b). The results of the study concord with the results of Awad (2010) and Al-Darabaseh (2012).

To answer the second question that was "Does the test information function differ according to the different percentages of missing data (0%, 5%, 10%, 15%)?" The information function of the test was calculated according to the missing data percentages of (0%, 5%, 10%, 15%) based on the three-parameter model in the Item Response Theory.

The information function of the effectiveness of the item in measuring the ability containing all the specifications of the item (discrimination and thresholds) was estimated. Thus, it provides an opportunity for comparing different items and individuals, which is estimated through the formula ($I(\theta, u_i) = \frac{p`_i{}^2}{p_i q_i}$) in the logistic model, and is estimated according to the Biloge-mg software over a range of (+3, -3).

The test information function could be found among the sum of item functions, and it indicates the quality of the combined items in the estimate of the characteristic estimated by the scale. The Biloge-mg software also calculates the test information function at different ability levels. Table (5) shows the values of the test information function according to the different missing percentage.

Table 5. *Information function value depending on the Percentage of Missing Data*

Percentage of Missing Data	Information function
%0	1.41
%5	1.56
%10	1.36
%15	1.34

It is evident from table (5) that the information function was the highest in the case of the missing data percentage of (5%), and that the order of values based on the missing data percentages was (5%, 0%, 10%, 15%), respectively. This confirms that the information function increases the more the missing data percentage decreases, and the reason for this may be due to the inverse relationship between the information function and the standard error in the estimates of the parameters of the items; as the missing data percentage decreases, it is expected that the percentage of standard error in the estimates of the item parameters will decrease, which leads to an increase in the value of the information function, as the relationship between the information function and the standard error in the estimate is shown in the formula ($SSE = \frac{1}{\sqrt{I(\theta)}}$). It is also clear that the information function in the case of the missing percentage of (5%) was higher than the information function if the missing percentage was (0%). The reason for that could be that the missing percentage is (5%) which does not force the student to guess, thus, leading to an increase in the discrimination parameters of the items, which reduces the percentage of standard error in the estimates of the parameters of the items. Furthermore, this increases the value of the information function for the test as guessing affects the estimate of the ability parameter and thus affects the amount of information. Therefore, the amount of information obtained when using the three-model is less than the amount of information that is obtained when using the two-model. However, if the missing percentage increases, this may negatively affect the specifications of the tests and their items, which affects the information function and, thus, the accuracy of the decisions made based on the results of these tests.

It could also be said that the decrease in the percentage of the missing data values increases the accuracy of estimating the maximum value of the informational function of the item, because the standard error associated with its estimation is related to an inverse relationship with the percentage of missing data values. Therefore, this result could be interpreted depending on the mechanism of estimating the item parameters and the ability in Item Response Theory; the greater the percentage of missing values, this is reflected in the accuracy of the estimates extracted for the statistics and characteristics of the items, including the maximum value of the item information function, which increases the standard error value of the estimate. In addition, the maximum value of the item information function is related to a certain ability level, and the item at a that level of ability is more effective in estimating the ability compared to other ability levels of the ability point of the ability continuum, which is reflected in the accuracy of decision-making related to choosing the item to measure that level of ability, which increases the accuracy of the ability parameter estimation.

RECOMMENDATIONS

1. Reducing the percentage of missing data of by no more than (5%).
2. Conducting further studies to find out the effect of the percentage of missing data on estimating the standard error of the item parameters and the test information function based on the one- and two-parameter models.

REFERENCES

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

Alruhail, R., & Aldarabsah, R. (2014). The effect of ability estimation method and handling method with missing values on the accuracy of items and persons parameters. The International Interdisciplinary Journal of Education,3(6),23-47.

Al-zubi, O. (2013). The effect of the percentage of missing data and imputation method in accuracy of estimating parameters of items and persons, Unpublished doctoral thesis. Yarmouk University, Jordan.

Allasasmah, S. (2018). The effect of the sample size and the method of dealing with the missing data on the test reliability and the item discriminate and difficulty Unpublished Master Thesis. Mutah University, Jordan.,

Baker, F. (2001). The basics of item response theory. ERIC. Clearinghouse on Assessment and Evaluation.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cokluk, O. & Kayri, M. (2011). The effects of methods of imputation for missing values on the validity and reliability of scales. Educational Sciences: Theory & Practice, 11 (1), 303 - 310.
Embretson, S. E. & Reise, S. P. (2000). Item response theory for psychologists. New Jersey: Lawrence Erlbaum Associates, Publishers.
Freihat, A, (2019). Investigating the effect of missing data imputation method and its percentage on the detection of differential item functioning for items of test using Raju's method. Basic Education College Magazine for Educational and Humanities Sciences,43(1), 801-827.
Hambleton, R. K. & Swaminathan, H. (1985). "Item response theory: principles and applications". Boston: Kluwer-Nijhoff publishing.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. London: Sage Publications, Inc.
Huisman, M., Krol, B. & Van Sonderen, F. L. P. (1998). Handling missing data by re-approaching nonrespondent quality & quantity. In Huisman, M. (edited). Item Nonresponse; Occurrence causes, and imputation of Missing Answers to test Item. DSWO Press, Lieden University, The Nethrlands, 1999.
Langkamp, D., Leman, A. & Lemeshow, S. (2010). Techniques for handling missing data in secondary analyses of large surveys. Academic Pediatrics, 10 (3), 205 – 211.
Little, R. J. A., Rubin, D. B. (1987). Statistical analysis with missing data.2nd edition, New York: john Wiley & sons.
Little,R.J.A;Rubin,D,B.(2002).Statistical Analysis with Missing Data. New York: Wiley.
Mcknight, P. E., Mcknight, K. M., Sidani, S. & Figueredo, A. J. (2007). Missing data: a gentle introduction. New York: Guilford press.
Sharifain, N. & Taamana, A. (2009). The effect of the number of alternatives in the multiple-choice test on estimating the individual's ability and psychometric characteristics of the items and the test according to the Rasch model in the item response theory. Jordan Journal of Educational Sciences,5(4),309-335.
Odeh, A. (2010). Measurement and evaluation in the teaching process. Dar Alamal, Irbed, Jordan.