e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

APPLICATION OF CLASSICAL TEST THEORY AS LINEAR MODELING TO TEST ITEM DEVELOPMENT AND ANALYSIS

Solomon Chukwu Ohiri*1, Romy O. Okoye*2

*1Directorate Of Academic Planning Alvan Ikoku Federal College Of Education, Owerri Imo State, Nigeria.

*2Dept. Of Educational Foundations, Faculty Of Education Nnamdi Azikiwe University, Awka, Nigeria.

DOI : https://www.doi.org/10.56726/IRJMETS45379

ABSTRACT

In psychology and education, tests are made of items. The quality of each item contributes in no small measure to the quality of the test. In ensuring that the items are of good quality, they are subjected to what is known as item analysis. Item analysis is a process which evaluates testees' responses on individual test items in order to ascertain the characteristics of each item and the relationship between them. It provides constructive feedback about the goodness of items. Effective test item development requires an organized, detail-oriented approach based on solid theoretical education measurement procedures to ensure validity and reliability of the test items. Classical test theory (CTT) is a conventional quantitative approach to testing the reliability and validity of an instrument based on its items. As a theory of error measurement, it has statistics for evaluating individual items from a quantitative perspective. The purpose of this paper is to describe in details, the application of classical test theory in test item development and analysis. The reason for the application of CTT is to have test items that will yield a reasonable degree of reliability. The statistics used in this regard are – item difficulty, which is a measure of the proportion of testees who responded to an item correctly; the item discrimination, which is the measure of how well the items discriminate between examinees with high and low levels of knowledge or ability. Also of interest are reliability, which deals with the degree to which the same responses repeatedly given to the same questions attract the same scores, and standard error of measurement (SEM), which is an index that indicates the accuracy with which an individual's score approximates the true score for the same individual. At the end of analysis, items are selected if the difficulty indices fall between 0.3 and 0.7. On item discrimination, an item is acceptable or selected if the discrimination index falls between +0.3 and +1.0.

Keywords: Item Difficulty, Item Discrimination, Reliability, Standard Error Of Measurement.

I. INTRODUCTION

In teaching and learning processes, tests are administered with the expectation that the measurements derived from them will be helpful in making decisions with minimum of risk. Test is an instrument used for assessing or ranking students in terms of ability. According to Nkwocha (2019), a test is an instrument used to find out whether an object or person possesses a particular attribute or characteristics. A test is a device in which a sample of examinees' behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process (Sapmaz, 2019). Okoye (2015) saw test as a set of questions, tasks or statements that can be presented to an individual, responses to which would enable the tester establish how much of a desired characteristic is possessed by the testee.

Test is used to measure learning progress and achievement, and to evaluate the effectiveness of educational programme. Tests also measure students' progress towards stated important goals. Therefore, creating quality test is very important in assessing students' performance.

In education, psychometricians are concerned with the design and development of tests, the procedures of testing, instruments for measuring data, and the methodology to understand and evaluate the results (Erguven, 2014). Identifying cognitive abilities of a testee and representing them as a reliable numerical score is the main purpose of educational and psychometric measurement.

A test can be studied from different angles and the items in the test can be evaluated according to different theories. In educational and psychological testing, there are two main frameworks by which a test and the items

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

it contains can be studied. These are classical test theory (CTT) and item response theory (IRT). The two frameworks are associated with the item development process in the field of educational and psychological test. These frameworks have been widely used in test development, test score equating, and identification of biased test items.

Well-developed measurement instruments (tests) have an essential place in educational and psychological programmes because they help to measure intended programme efficiency via educational outcomes (Sapmaz, 2019). Test development is the process of producing a measure of some aspects of an individual's knowledge, skills, abilities, interest, attitudes, or other characteristics by developing questions or tasks and combining them to form a test, according to a specified plan (Standards cited in Sapmaz, 2019). There are many kinds of test, such as those measuring intelligence, attitude, or the ability of individuals and groups alike, that can be used for different purposes.

Classical Test Theory

According to Eleje, Onah and Abanobi (2018), classical test theory (CTT) has been the foundation for measurement theory for decades. The conceptual foundations, assumptions and extensions of the basic premises of CTT have allowed for the development of psychometrically sound scales in the assessment practices of educational bodies. This is due to the simplicity of interpretation which can usefully be applied to examine achievement and aptitude test performance. According to Bichi (2016) classical test theory was born only after the following three achievements or ideas were conceptualized:

- a recognition of the presence of errors in measurements
- a conception of error as a random variable
- a conception of correlation and how to index it.

Classical test theory is a psychometric theory of assessment/measurement that purports that every individual has some innate or "true" ability for any given attribute, and that the attribute can be measured, and the process of measurement inherently has error (Wang, 2018). Allen and Yen in Bichi (2016) stated that in 1904, Charles Spearman was responsible for figuring out how to correct a correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in making the correction. Spearman's findings are thought to be the beginning of classical test theory.

According to Bejar cited in Erguven (2014), random sampling theory and item response theory are two major psychometric theories in the study of measurement procedures. In random sampling theory, there are two approaches, the classical test theory approach and the generalizability theory approach. Marcoulides cited in Bichi (2016) maintained that, classical test theory (also known as classical true score theory) is a simple model that describes how measurement errors can influence observed scores. Classical test theory is the earliest theory of measurement. The major concern of this theory is estimating the reliability of the observed scores of a test. With the framework of classical test theory, each measurement (test score) is considered being a value of a random variable X consisting of two components: a true score and an error score. This relationship is represented below as the classical test model:

$$X = T + E \qquad (1)$$

This is a simple linear model that links the observable test score (X) to the sum of two unobservable variables - true score (T) and error score (E). It is so because the true score is not easily observed, instead, the true score must be estimated from the individual's responses on a set of test items.

Mathematically, classical test theory is based on the premise that the observed score from a test is composed of an immeasurable true score and error score. It is the error that is the most important aspect of the equation.

Score (X) = True score (T) + Measurement Error (E)
↓ ↓
Unmeasurable can be estimated

Error is inherent in almost all measurement devices one can think of. In the framework of classical test theory, the observed score (X) is assumed to be measured with error. On this premise, in developing measures, the aim of classical test theory is to minimize this error. In that case, importance of a reliability of a test and calculating the reliability coefficient increases (Erguven, 2014).

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Kaplan and Saccuzo cited in Bichi (2016) stated that, theoretically, the standard deviation of the distribution of random errors for each examinee tells about the magnitude of measurement error. This standard deviation of the distribution of random error around the true score is called the standard error of measurement. It is a number that indicates the accuracy with which an individual's score approximates the true score for the same individual. Mathematically, standard error of measurement can be computed using sample data as follows:

$$SE_x = SD_x \times \sqrt{1\text{-Reliability (X)}} \quad (2)$$

Where SE_x is the standard error of measurement (SEM)

SD_x is the standard deviation of the observed test scores

Reliability (X) = Estimated reliability coefficient of test score.

The smaller the standard error of measurement, the more reliable the test is.

Assumptions of Classical Test Theory

Classical test theory assumes linearity, that is, the regression of the observed score on the true score is linear. This linearity assumption underlies the practice of creating tests from the linear combination of items.

According to Nasir (2014), the following four assumptions are implicit with classical test theory:

- The observed score of a person is made up of the true score and random error.
- The expected value of any observed score is the person's true score.
- The covariance of error components from two tests is zero in the population. That is, error from two tests are uncorrelated.
- Errors in one test are uncorrelated with true scores in another. That is, measurement errors are not dependent on traits.

The assumptions can be readily derived from the definitions of true score and measurement error. Hence, they are commonly shared by all the models of classical test theory.

Item Analysis

The qualities of items that make up a test determine the quality of the test as a whole and the assessment of these essential qualities of the items in a test constitutes item analysis. Item analysis is a technique that evaluates the effectiveness of items in tests. Test item analysis is broadly referred to as the specific methods used to evaluate items on a test, both qualitatively and quantitatively, for the purpose of evaluating the quality of individual items (Krishnan, 2013). The target is to help test developers to improve the instrument by revising or discarding items that do not meet a minimally acceptable standard. Item analysis is concerned with examining responses to individual test items to assess the item quality. Item analysis is important in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. Again, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity. The aim of achieving quality means minimizing the measurement error in scores. By using the internal criterion of test scores, item analysis presents such statistics as reliability coefficient to check for the internal consistency of items, which is also a first step in achieving the validity of test items.

Classical test theory has statistics for evaluating individual items from a quantitative perspective. The concern of item analysis is to use these detailed statistics to determine possible flaws in the item, and then decide whether to revise, replace, or retire the item (Thompson, 2016). This can be something as specific as identifying a bad distractor because it pulled a few high-ability examinees or something as general as: "this item is harder than the other".

Item analysis is typically carried out before the test goes live to ensure that only quality items are used. Many a time, it is done after pretesting the items on some small set of the sample. Item analysis is important because it is analogous to quality control of parts used in the assembly of a final manufacturing product (Thompson, 2016).

CTT-based Item Analysis

Since our concern here is on item analysis based on classical test theory, it is imperative to explore the basic ideas involved in order to fully understand the approach. Classical test theory as a body of theory and research,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

could be used to predict or explain the difficulty of questions, provide insight into the reliability of test scores, and help us towards coming up with an assessment of how to improve the test by maintaining and developing a pool of good items from which future assessments can be drawn (Krishnan, 2013). Therefore, special attention will be given to individual items, item characteristics, the probability of answering items correctly, the overall ability of the test taker, and the extent to which an item conforms with the rest of the items in a test.

Drawing on the important key concepts at a theoretical level, we explore the essential things to look for in a typical item analysis based on classical test theory. They include item difficulty, item discrimination, reliability and standard error of measurement (SEM).

Regarding the actual statistical analysis of the items during item analysis, the method is to compare the responses of testees in the upper one-third and the lower one-third continuum on the basis of total test scores. The responses of the testees in the middle one-third are not included in the analysis. The responses that have been made to each group are tabulated thus:

	A	B	C	D	E
Upper group (Upper $^1/_3$)					
Lower group (Lower $^1/_3$)					

a) Item difficulty

It is a simple concept in classical test theory. It simply refers to the proportion of examinees that correctly answered an item. This is called the p-value. This metric takes a value between 0 and 1. High values indicate that the item is easy, while low values indicate that the item is difficult.

This index denoted by 'P' is calculated thus:

$$P = \frac{U + L}{2N} \qquad (3)$$

Where U = the number of testees in the upper $^1/_3$ of the group who got the item right

L = the number of testees in the lower $^1/_3$ of the group who got the item right

N = the total number of testees in either of the upper or lower $^1/_3$ of the group

An ideal item is supposed to have a difficulty index of 0.5, but it may be difficult to have items with this index. Hence, an item is acceptable if the difficulty index falls between 0.3 and 0.7. If the difficulty index is less them 0.3, it shows that the item is difficult while any value greater than 0.7 indicates that the item is very easy.

b) Item discrimination

According to Okoye (2015), an item is considered good if it is got right by the bright students and failed by the dull ones. Item discrimination refers to the power of the item to differentiate between examinees with high and low levels of knowledge or ability (Thompson, 2016). It is the correlation between item scores and total test scores called the item-total correlation. A good item records more passes in the upper one-third than in the lower one-third.

The discrimination index of an item can be computed using any of the followings: the item discrimination index (d) and the item discrimination coefficient.

(i) **Item discrimination index (d):** The formula for calculating discrimination index is

$$d = \frac{U\text{-}L}{N} \qquad (4)$$

Where U, L and N are defined as in the case of item difficulty index.

The discrimination index ranges from -1.0 to + 1.0. An item is poor and unacceptable if the discrimination index is zero, because this implies that it has not been able to discriminate between the two groups. It is also unacceptable when the value is negative, because it implies that more of the lower group chose the correct answer than the upper group, which is an abnormal situation. On the other hand, when the discrimination index is positive, such an item is seen to be discriminating in the right direction. However, an item is considered acceptable only when the index falls between +0.3 and +1.0.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

(ii) **Item discrimination coefficient:** The method employed in (i) above gives a fairly stable index of discrimination. It is problematic in that the process of its computation ignores so much data (Adegoke, 2013). To correct the problem, we use the point-biserial correlation, though some researchers prefer to use its cousin called biserial correlation.

The point-biserial correlation coefficient, γ_{pb}, is a special case of Pearson's correlation coefficient. This provides an index of whether students who get the item correct are scoring high, which is the hallmark of a good item. It measures the relationship between two variables. The formula for the point-biserial coefficient is:

$$\gamma_{pb} = \frac{M_I - M_o}{S_n}\sqrt{Pq} \qquad (5)$$

Where M_1 is the mean (for the entire test) of the group that received the positive binary variable (i.e. the "1"),

M_o is the mean (for the entire test) of the group that received the negative binary variable (i.e. the "0"),

S_n is the standard deviation for the entire test,

P is the proportion of cases in the "0" group,

q is the proportion of cases in the "1" group.

Point-biserial coefficient values range from -1 to +1. A negative value of γ_{pb} indicates that the variables are inversely related. On the other hand, positive values indicate that the variables are directly related, while 0 indicates no association at all. Very low or negative point biserial coefficients help in identifying defective test items. An item is acceptable if the discrimination coefficient falls between +0.3 and +1.0.

(c) Reliability

Reliability is a classical test theory concept that seeks to quantify the consistency or repeatability of measurement (Thompson, 2016). A reliable test is one we can trust or we can use to measure a person's performance approximately the same way each time. The reliability of a test refers to the extent to which the test is likely to produce consistent scores. It can be described as the degree to which the same responses repeatedly given to the same questions attract the same scores (Nkwocha, 2019).

Technically, the reliability of a test deals with the proportion of the total variance of a test that is due to true variance. The degree of consistency of a test is expressed as a coefficient called the coefficient of reliability. It is usually estimated by correlating two sets of scores independently obtained with the test. This coefficient has been seen as a description of the loss in efficiency of estimation resulting from measurement error.

There are different means of estimating the reliability of any measure. However, in practice, we talk about those types of reliability we can estimate. Of the four general classes of reliability estimates researchers use, {test-retest (coefficient of stability), parallel forms (coefficient of equivalent forms), inter-rater and internal consistency}, the examination of reliability in this paper is focused on the internal consistency reliability. More specifically, we focus on Split-half reliability, Cronbach's alpha, Kuder Richardson (K-R) method and Rulon's method.

Internal consistency reliability estimation is based on a single test administered to a group of individuals in one occasion. It refers to the degree to which the items that make up the construct of interest are measuring the same underlying construct.

Reliability coefficient takes value between 0 and 1. The following guidelines can be used for the interpretation of the values of reliability coefficients according to Yolonda (2015):

- 0.9 and greater = excellent reliability
- 0.8 - 0.9 = good reliability
- 0.7 – 0.8 = acceptable reliability
- 0.6 – 0.7 = questionable reliability
- 0.5 – 0.6 = poor reliability
- 0.5 and less = unacceptable reliability

(i) **Split-half method:** In this case, the test is administered to the same group of testees once. The common procedure is to divide the test into two groups, with odd-numbered items usually placed in one group and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

even-numbered in the other. Each testee gets two scores, one from each half of the test. The scores on one half of the test are correlated with the scores on the second half. The computation is done with Pearson's product moment correlation method. Since the score of each testee has been divided into two, the correlation index estimated is the split-half reliability coefficient. To calculate the reliability index for the full test, Spearman-Brown computation formula is used based on the split-half.

Spearman-Brown formula is given as:

$$r_f = \frac{n\, r_s}{1+(n-1)\, r_s} \qquad (6)$$

Where r = Reliability of the full test

N = The number of times the test was shortened or elongated

r_s = The reliability of the shortened or elongated test

Split-half method is used when the items of the test are homogenous. That is, when the items measure one construct.

(ii) **Cronbach's coefficient alpha:** This is a lower-bound estimate of reliability under the assumption that items are with uncorrelated errors. It can be used for any mixture of binary (true/false) and partial credit items (true/sometimes/false). This is computed by correlating the score for each scale item with the total score for each observation (usually –individual test takers), and then comparing that to the variance for all individual item scores.

The formula for coefficient alpha according to Nkwocha (2019) is

$$r = \left\{\frac{K}{K-1}\right\}\left\{\frac{V_t - \sum V_i}{V_t}\right\} \qquad (7)$$

Where r = coefficient alpha (the reliability index)

K = number of items that compose the test

Vt = variance of total scores of each respondent on the test

Vi = variance of scores obtained by all respondents on each item.

$\sum Vi$ = Sum of total variance of scores for all items.

(iii) **Kuder-Richardson (K-R) Method**: Kuder-Richardson's reliability is a method that makes use of the full test. It is used for frequency scores. Kuder-Richardson's approach avoids the problem of how to split the items, and it has two procedures-KR-20 and KR-21. The KR-20 is best used for a test that does not have many items because the formula requires computation of the proportion of those who passed each item and the proportion of those failed each item. KR-21 does not require such rigor and hence is used when the test items are many.

In the computation of K-R reliability coefficient, a single test is administered to a group of testees. It estimates the consistency of responses to all the items in a test.

$$\text{The formula for KR-20} = r = \left\{\frac{K}{K-1}\right\}\left\{1 - \frac{\sum pq}{SD_t^2}\right\} \qquad (8)$$

Where K = number of items the whole test is composed of

P = proportion of those who passed each item

q = proportion of those who failed each item

SD_t = square of standard deviation of testees scores of the whole test

$$\text{The formula for KR-21} = \left\{\frac{K}{K-1}\right\}\left\{1 - \frac{\bar{X}(K-\bar{X})}{K(SD_t)^2}\right\} \qquad (9)$$

Where K, SD_t are defined as in KR-20

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

X = mean of the summated scores

Kuder- Richardson method provides an estimate of the average reliability found by taking all possible splits without actually having to do so.

(iv) **Rulon's Method:** Rulon's Method is a simple method which does not require the computation of the reliability coefficient for the one half of the test.

The computational formula for Rulon's coefficient is given as follows:

$$r_{11} = 1 - \left\{ \frac{\sigma_a^2 + \sigma_b^2}{\sigma_t^2} \right\} \qquad (10)$$

Where σ_a and σ_b are standard deviations of the two halves of the test respectively. σ_t is the standard deviation of the whole test.

The common procedure is to divide the test items into two halves, with odd-numbered items usually in one group and even-numbered items in the other group.

(d)Standard Error of Measurement (SEM)

SEM is directly related to the reliability of a test. It is a number or an index that indicates the accuracy with which an individual's score approximates the true score for the same individual. If a test is administered an infinite number of times on an individual, we expect the scores to change from one another, i.e. the scores will vary. The mean of the scores obtained from the infinite tests, will be taken as the true score of that individual. Thus, some scores will be above the mean (the true score) while some will be below. The difference between each such score and the mean (the true score) is the error score. The standard deviation of these differences constitutes what is referred to as the standard error of measurement (SEM). The smaller the SEM, the more accurate the measurement. This is because, when the SEM is small, it implies that the scores obtained in the measurement are relatively close to the true score, thereby giving rise to minimal errors. It is this degree of accuracy that is sought for using reliability index. Standard error of measurement and reliability are therefore inversely related.

It is not practicable to administer a test an infinite number of times to an individual in order to ascertain SEM. However, reliability estimates can be obtained in various ways. Knowing therefore that SEM and reliability coefficient are related, the SEM can be computed when reliability coefficient has been determined with a set of obtained scores, using the formula

$$SEM_x = SD_x\sqrt{1 - \text{reliability (x)}} \qquad (11)$$

Where SEM_x is the standard error of measurement

SD_x is the standard deviation of the observed test score.

Reliability (x) is the estimated reliability coefficient of test score

Based on the computational formula above, one can deduce that the smaller the standard error of measurement, the more reliable the test is. According to Verhelst cited in Krishnan (2013), standard error of measurement can help in the interpretation of scores and can be used to calculate confidence intervals.

SEM is expressed in the same scale as the test scores. The formula for SEM indicates that the standard error of measurement must be 0 when the reliability is +1; when reliability is 0, the SEM is equal to the standard deviation. This means that if measurements are entirely unreliable (0 reliability), the spread of obtained scores is due to chance conditions.

Item Selection in Classical Test Theory

Constructing test items calls for enough time and careful selection of the content that will produce the desired test results. In the classical test theory, item analyses provide crucial information based on statistical criteria for the determination of sample specific parameters and elimination of bad items.

At the end of the item analysis, test items are listed according to their degrees of difficulty and discrimination. This arrangement provides a clear overview of the test and can be used to identify items which are to be

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

selected and those that will be discarded. Items are selected if the difficulty indices fall between 0.3 and 0.7. If the difficulty index is less than 0.3, it shows that the item is difficult, while any value greater than 0.7 indicates that the item is very easy. For item discrimination, an item is acceptable or selected if the discrimination index falls between +0.3 and +1.0. In test development, items are selected on the basis of these two characteristics above: item difficulty and item discrimination. Hence, item analysis acts as quality control in test development.

II. CONCLUSION

Item analysis is a vital step in test development cycle, as all tests are composed of items and good items are necessary for a good test. Classical test theory provides some methods for evaluating items based on simple statistics like proportion, correlation, reliability of the measurement tools currently used in educational and psychological tests and research.

In classical test theory, the main concern of item analysis is to describe the statistical characteristics of each item. The total score of a test is considered the sum of scores on the individual items, and the individual item is of interest through its effect on the total test score. Thus, item analysis in classical test theory is focused on the degree to which each item influences the whole measurement.

Procedures commonly used in the development and analysis of test items under classical test theory include item difficulty, item discrimination, reliability, etc. These approaches to item analysis should be sustained in test development and test item analysis. This is based on its superiority and simplicity in the investigation of reliability and in minimizing measurement errors (Bichi, 2016).

III. RECOMMENDATION

In pursuance of quality psychological and educational tests, all tests should be made to pass through all the rigorous and meticulous processes of standardization and validation.

IV. REFERENCES

[1] Adegoke, B. A. (2013). Comparison of item statistics of physic achievement test using classical test theory and item response theory frameworks. Journal of education and Practice, 4(22), 87-96.

[2] Bichi, A. A. (2016). Classical test theory: An introduction to linear modeling approach to test and item analysis. International Journal for Social Studies, 2(9), 26-33.

[3] Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. European Journal of Educational & Social Sciences, 3(1), 57-75.

[4] Erguven, M. (2014). Two approaches to psychometric process: Classical test theory and item response theory. Journal of Education, 2(2), 23-30

[5] Krishnan, V. (2013). The early child development instrument: An item analysis using classical test theory on Alberta's data. Early Child Development Mapping Project (ECmap) Alberta, Community-University Partnership, Faculty of Extension, University of Alberta, Canada.

[6] Nasir, M. (2014). Application of classical test theory and item response theory to analyze multiple choice questions. Unpublished Ph.D. Dissertation, Department of Medical Science, University of Calgary, Alberta, Canada.

[7] Nkwocha, P. C. (2019). Basics of educational measurement and evaluation (Revised ed.). Owerri: Mercy Divine Publishers

[8] Okoye, R. O. (2015). Educational and psychological measurement and evaluation (2nd ed.). Awka: Erudition Publishers.

[9] Sapmaz, Z. M. (2019). Detection of gender-related differential item functioning (DIF) in the mathematics subjects in Turkey. Unpublished masters' thesis, Department of Educational and Psychological studies, University of South Florida.

[10] Thompson, N. A. (2016). Introduction to classical test theory with CITAS. Minnesota: Assessment System Corporation.

[11] Wang, V. (2018). Handbook of research on program development and assessment methodologies in K-20 education. Chicago: IGI global.