IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

Application of Item Response Theory as a Modern Statistical Tool to Test Item Development and Analysis

Solomon Chukwu Ohiri
Directorate of Academic Planning,
Alvan Ikoku Federal College of Education, Owerri Imo State, Nigeria
ohiri.chukwu2018@gmail.com

Abstract: *In the educational and psychological testing, there are two major theories through which tests can be developed, validated and ultimately used for assessing examinee's performance. These are classical test theory (CTT) and item response theory (IRT) and their corresponding models. Item Response Theory (IRT) as a test theory came into existence to provide probabilistic approach to surmount some of the inherent limitations of the classical test theory and maximize objectivity in educational assessment. Item response theory (IRT) is a quantitative approach to testing the reliability and validity of an instrument based on its items. It has statistics for evaluating individual items from a quantitative perspective. The purpose of this paper is to describe in details, the application of item response theory in test item development and analysis. The reason for the application of IRT is to have test items that will yield a reasonable degree of reliability. The statistics used in this respect are – item difficulty parameter, which is a measure of the proportion of testees who responded to an item correctly; the item discrimination parameter, which is the measure of how well the items discriminate between examinees with high and low levels of knowledge or ability and pseudo-guessing parameter, which expresses the probability that an examinee with low ability can be able to get an item correctly. The acceptable range of values of the aforementioned parameters were also discussed.*

Keywords: Item Response Theory, Statistical Tool, Test, Analysis.

I. INTRODUCTION

In the educational process, a variety of tools and techniques are used to measure the learning outcome. Test as one of the essential tools, is used to measure learners' academic outlook. Test is an instrumentfor 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 characteristic. 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.

Tests that are well-developed have essential place in educational and psychological programmes because they help to measure intended programme efficiency via educational outcomes. 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.

A test can be studied from different angles and the items in the test can be evaluated according to different theories. In the educational and psychological testing, there are two major theories through which tests can be developed, validated and ultimately used for assessing examinee's performance. These theories are classical test theory (CTT) and item response theory (IRT) and their corresponding models. The two frameworks are associated with the item development process in the field of educational and psychological testing. The CTT is a relatively simple theory for testing and has

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

been widely used for constructing and evaluating tests, particularly before the birth of IRT. Classical Test Theory (CTT) is a traditional quantitative approach in testing the reliability and validity of a scale-based items (Cappelleri, Lundy & Hays, 2014). This theory assumes that each observed score (X) of a person is the combination of an underlying true score (T) and unsystematic error (E). This can be written mathematically as:

$$X=T+E$$

The above equation is known as" classical test model". CTT assumes that observed score has a proportion of true score and random error due to errors of measuring instruments. In addition, variability of the test and examinee's conditions also contribute to these errors. If the same examinee takes the same exam different number of times (without the effects of any learning taking place) errors will approach zero, and the observed score will be equal to the true score (Oratokhai, 2021).

In classical test theory framework, using the selected sample, some indices like item difficulty, item discriminations are calculated for each item. The quality of item will be decided on the basis of these values. The quality of the test as a whole will be determined on the basis of some coefficients for reliability and validity.Mostly, item analysis is being carried out with the principles of classical test theory, but in recent times, the item response theory (IRT) is getting more popular among researchers, test experts and examination bodies (Abdullahi &Darazo, 2020).

II. ITEM RESPONSE THEORY

Item response theory (IRT) was first proposed in the field of psychometrics for the purpose of ability assessment and was developed to address the limitations in the classical test theory. It is widely used in education to calibrate and evaluate items in tests, questionnaires and other instruments, and to score subjects on their abilities, attitudes, or other latent traits (Xinming & Yiu-Fai, 2014). IRT is an area of test theory which provides probabilistic approach to overcome some of the limitations of classical methods (Ashraf & Jaseem, 2020). The item response theory (IRT), also known as the latent response theory refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations. It is a statistical technique involving models expressing the probability of a particular response to a scale item as a function of the ability of the subject. According toEleje, Onah and Abanobi (2018), item response theory (IRT) is a collection of measurement models that attempt to explain the connection between observed item responses on a scale and an underlying construct. IRT attempts to model the ability of an examinee and the probability of answering a test item correctly based on the pattern of responses to the items that constitute a test.

IRT allows the user to specify a mathematical function to model the relationship between a latent trait, Θ, and the probability that an examinee with a given Θ will correctly answer a test item (Erguven, 2014). Item response theory research focuses largely on the estimation of model parameters, the assessment of model-data fit, and the application of these models to a range of testing problems using dichotomously scored (Yes/No, 1 or 0) multiple-choice items. Item responses can be discrete or continuous and can be dichotomously or Polytomouslyscored; item score categories can be ordered or unordered; there can be one ability or many abilities underlying test performance; and there are many ways (that is, models) in which the relationship between item responses and the underlying ability or abilities can be specified. The most common models in IRT are used for dichotomous items, which are the one-/Rasch, two-, three-, four-logistic parameter models (Sapmaz, 2019).In IRT, the item responses are considered the outcome (dependent) variables, and the examinee's ability and the items' characteristics are the latent predictor (independent) variables (Le in Bichi, Embong, Talib, Salleh & Ibrahim, 2019).

The Item Response Theory is the study of test and item scores based on assumptions concerning the mathematical relationship between abilities and item responses. IRT is used in developing and refining tests and examinations, maintaining banks of items for examinations and comparisons between results over time. The IRT has the possibility of obtaining item characteristics which are not group dependent; ability scores, which are not test dependent; and a measure of precision for each ability level (Esomonu & Okeke, 2021). Under IRT, item difficulty describes where an item functions along the ability scale (Baker, 2001). It allows item difficulty to be estimated in an unbiased way. Item discrimination in IRT is the correlation between the item and test. IRT attempts to model the ability of an examiner and the probability of answering a test item correctly based on the pattern of responses to the items that constitute a test.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

The number of item parameters to be estimated determines which IRT statistical model will be used, and the test item analysis of any examination is based on item discrimination, item difficulty and the guessing parameters.

2.1 Assumptions of Item Response Theory

- Unidimensionality: This assumption requires that only one ability is measured by a set of items in a test. This means that the performance of each examinee is assumed to be governed by a single factor, known as ability. The assumption is that the probability that a test taker will answer a test item correctly depends on only one characteristic of the test taker.
- Local independence: This assumption requires that the responses of examinees to any pair of items are statistically independent when the ability influencing test performance is held constant. In the local independency assumption, responses for different items are not related. An item does not provide any clue to answer another item correctly. If local dependence does exist, a large correlation between two or more items can essentially affect the latent trait and can cause lack of validity (Erguven, 2014). Local independence does not mean that items do not correlate with each other, but that performance on different items is independent but conditional on the student's ability (Ojerinde in Eleje, Onah & Abanobi, 2018).
- Item invariance: This is another underlying assumption of item response theory. It can be described as the phenomenon in which estimated item parameters are constant across different populations. This allows for unbiased estimates of item parameters to be obtained from unrepresentative samples.
- Monotonicity: This refers to the phenomenon in which the probability of endorsing an item will continuously increase as an individual's trait level increases.

2.2 Models of Item Response Theory

A model is a representation of something that happens in real life. There exist numerous IRT models that differ on the type and number of item parameters estimated as well as in their suitability for different types of data.The models consist of a set of mathematical statements. Those statements are usually assumptions about the relationships between things that can be measured (Livingston, 2020).The first consideration when choosing the rightful model involves the number of item response categories. In the case of dichotomous items, the 1, 2, and 3 parameter logistic models(1PL, 2PL, 3PL,) are most common, and models including an upper asymptote parameter (e.g. 4PL) are also possible (Erguven, 2014). In the case of polytomous items, variations of the Partial Credit Model, Rating Scale Model, Generalized Partial Credit Model, and Graded Response Model are available for ordered responses, and the Nominal Model is appropriate for items with a non-specified response order (Erguven, 2014). The second important consideration is whether the item discrimination parameters, or slops, should be free to vary across items. The item response theory model (1PL, 2PL, 3PL) can be defined using the 3PL model formula:

$$P_i(\Theta) = c_i + (1-c_i)\ \frac{\exp\ a_i(\Theta-b_i)}{1+\exp\ a_i(\Theta-b_i)} \qquad i = 1, 2, \ldots, n.$$

where $P_i(\Theta)$ is the probability that a given testee with ability Θ answers a random item correctly, a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo guessing parameter (Hambletan, Swaminathan & Rogers in Erguven, 2014). The 2PL model is obtained when c is 0. The 1PL model is obtained if c is 0 and a is 1.

A "b" parameter defines how easy or how difficult an item is, and an "a" parameter determines how effectively this item can discriminate between highly proficient students and less-proficient students. The guessing parameter "c" determines how likely the examinees are to obtain the correct answer by guessing (Yu, 2013).

Schumacker in Bichi and Talib (2018)summarized the models when he said; IRT models differ depending on whether the relationship between item performance and knowledge is considered a one-, two- or three-parameter logistic function. Different IRT parameterization models adjust for different item properties leading to different ability estimation. In 1-parameter (1-PL), IRT adjusts for item difficulty; 2-parameter (2-PL), IRT accounts for difficulty and discrimination of an item; and 3-parameter (3-PL), IRT takes into account the effect of item difficulty, discrimination,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

and ease of guessing the correct answer. Using the appropriate IRT model, the ability level of an examinee is accurately estimated with any set of items that measure this ability (Esomonu & Okeke, 2021). The IRT is mostly used for modeling responses to items and scoring of educational tests. IRT is based on the idea that the probability of a correct response to an item is called latent trait or ability.

III. IRT-BASED 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. It is a process which examines students' responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. According to Urbina(2014), item analysis is a general term that refers to all the techniques used to assess the characteristics of test items and evaluate their quality during the process of test development and test construction. Within the item analysis, all the possible test items are subjected to stringent series of evaluation procedures, individually and within the context of the whole test. Item analysis is useful in both the development, evaluation of assessments tools and in computing standardized measures of student performance. Again, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which needs greater emphasis or clarity.

IRT positions all the test items on a common scale alongside the examinees and allows the measurement of any subset items to the person's ability on the latent trait (Shanmugam, 2020). Cohen, Bottge and Wells in Shanmugam (2020) clarified that a person's ability refers to the amount of latent trait and the test scores represent the amount of latent trait that the examinees have. The latent trait is assessed by the items composing the test. This is because the examinees' observed responses to the test items indicate their position in a scale of unobservable latent trait, which the test content assesses (Ellis, Becker & Kimmel in Shanmugam, 2020).

Since our concern here is on item analysis based on item response theory, it is imperative to explore the basic ideas involved in order to fully understand the approach. The focus will be on the a, b and c parameters

The "a" parameter: This determines how effectively an item can discriminate between highly proficient students and less-proficient students. According to Okoye (2015), an item is considered good if it is got right by the bright students and failed by the dull ones. Discrimination is the extent to which the item separates test takers above some point on the ability scale from test takers below that point. Item discrimination refers to the power of the item to differentiate between examinees with high and low levels of knowledge or ability (Thompson, 2016).The item discrimination "a" parameter expresses how well an item can differentiate among examinees with different ability levels. A test item has positive discrimination when lower ability students have a low probability of answering an item correctly, and higher ability students have a high probability of getting the item right. A test item has negative discrimination when high ability candidates have a low probability of answering an item correctly and low ability candidates have a higher probability of answering an item correctly. 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 choose 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. Theoretically, the scale for the IRT item discrimination ranges from $-\infty$ to $+\infty$ and its value does not exceed 2.0. Thus, the item discrimination parameter ranges between 0.0 and 2.0 in practical use.

The "b" parameter: This is called item difficulty. It simply refers to the proportion of examinees that correctly answered an item. An ideal item is supposed to have a difficulty index of 0.5, but it may be difficult to have items with this index. The typical values of the item difficulty range from -3 to +3. Items whose difficulty indices are close to -3are termed very easy, while items with difficulty indices near +3 will be termed very difficult items. Items with difficulty parameter between 0.35 to 1.69 are considered appropriate (Esomonu& Okeke, 2021)

The "c" parameter: In the 3PLM, this is referred to as a pseudo-guessing parameter. This parameter expresses the probability that an examinee with low ability can be able to get an item correctly by guessing alone and, therefore, has a greater-than-zero probability of answering an item correctly in a test. The guessing parameter c is the lowest value that an ICC attains. Theoretically, the guessing parameter ranges between 0 and 1. Practically, values above 0.35 are not acceptable. Hence, the range $0 < c < 0.35$ is applied.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

3.1 Item Selection in Item Response Theory

Constructing test items calls for enough time and careful selection of the content that will produce the desired test results. In the item response theory, item analyses provide crucial information based on statistical criteria for the determination of sample specific parameters and elimination of bad items. The final selection of test items will be a function of the information each item contributes to the overall test. The beauty of the item information functions used in IRT test development is that, they permit the test developer to determine the contribution of each item to the test.

At the end of the item analysis, test items are listed according to their degrees of difficulty, discrimination and pseudo-guessing parameters. This arrangement provides a clear overview of the test items and can be used to identify items which are to be selected and those that will be discarded. Items are selected if their "a", "b" and "c" parameter indices fall within the acceptable range of values.

IV. CONCLUSION

The aim of item response theory (IRT) is to provide an improved framework for evaluating the validity of tests and their individual items. Item response theory (IRT) is seen as an improvement over classical test theory (CTT). In general, IRT offers greater flexibility and provides more sophisticated information about tests and their items.IRT is an indispensable modern statistical tool to test item development and analysis. Hence, in pursuance of quality psychological and educational tests, examination bodies should adopt IRT in test development, validation and standardization.

REFERENCES

[1]. Abdullahi, I. &Darazo, F.I. (2020). Analyses of psychometric properties of 2016 mathematics Basic Education Certificate Examination Questions (BECEQ) in Gombe State. Nigerian Journal of Educational Research and Evaluation, 19, 67-77.

[2]. Ashraf, Z.A. &Jaseem, K. (2020). Classical and modem methods in item analysis of test tools. International Journal of Research and Review, 7(5), 397-403.

[3]. Baker, F. (2001). The basics of item response theory. Washington D.C: ERIC Clearinghouse on Assessment and Evaluation.

[4]. Bichi, A. A. & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. International Journal of Evaluation and Research in Education (IJERE), 7(2), 142 – 151.

[5]. Bichi, A. A.; Embong, R.; Talib, R.; Salleh, S. & Ibrahim, A. B. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. International Journal of Engineering and Advanced Technology (IJEAT), 8 (5), 1260 – 1266.

[6]. Cappelleri, J. C., Lundy, J. J & Hays R. D. (2014). Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing Patient-Reported Outcome Measures. Clinical Therapeutics, 36 (5), 648 – 662.

[7]. 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.

[8]. EsomonuN. P. M. & Okeke, O. J. (2021). French language diagnostic writing skill test for junior secondary school students: Construction and validation using item response theory. International Journal of Education and Social Science Research, 4(2), 334 – 350.

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

[10]. Livingston, S. A. (2020). Basic concepts of item response theory: A nonmathematical introduction.New Jersey: Educational Testing Service.

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

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

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

[13]. Oratokhai, D. I. (2021). Investigating differential item functioning in National Business and Technical Examination English language multiple choice test items. Unpublished Ph.D research seminar, Department of Educational Evaluation and Counseling Psychology, University of Benin, Benin City.

[14]. Sapmaz, Z.M. (2019). Detection of gender-related differential item functioning (DIF) in the mathematics subjects in Turkey. Unpublished Master's Thesis, Department of Educational and Psychological Studies, University of South Florida.

[15]. Shanmugam, S.K.S. (2020). Gender-related differential item functioning of mathematics computation items among non-native speakers of English. The Mathematics Enthusiast, 17(1), 107-140

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

[17]. Urbina, S. (2014). Essentials of Psychological Testing (2nd ed.). Hoboken, NJ: John Wiley & Sons.

[18]. Xinming, A. &Yiu-Fai, Y. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. Controlled Clinical Trials, 24, 1-14.

[19]. Yu, C. (2013). A simple guide to the item response theory (IRT) and Rasch modeling. Journal of Education and practice,24 (11), 1-30.