International Journal of Evaluation and Research in Education (IJERE)
Vol.7, No.2, June 2018, pp. 142~151
ISSN: 2252-8822, DOI: 10.11591/ijere.v7.i2.pp142-151

Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development

Ado Abdu Bichi[1], Rohaya Talib[2]
Measurement & Evaluation, Faculty of Education, Universiti Teknologi Malaysia, Malaysia

Article Info

Article history:

Received Apr 23, 2018
Revised May 27, 2018
Accepted May 31, 2018

Keyword:

Classical Test Theory
Item Response Theory
Reliability
Validity

ABSTRACT

Testing in educational system perform a number of functions, the results from a test can be used to make a number of decisions in education. It is therefore well accepted in the education literature that, testing is an important element of education. To effectively utilize the tests in educational policies and quality assurance its validity and reliability estimates are necessary. There are two generally acceptable frameworks used in evaluating the quality of test in educational and psychological measurements, these are; Classical Test Theory (CTT) and Item Response Theory (IRT). The estimates of test items validity and reliability depend on a particular measurement model used. It is vital for a test developer to be familiar with the different test development and item analysis methods in order to facilitate the development of a new test. The CTT is a traditional approach which was widely criticise in the measurement community for its shortcomings such as sample dependency of coefficient measures and estimates of measurement error. However, the IRT is a modern approach which provides solutions to most of the CTT's identified shortcomings. This paper therefore, provides a comprehensive overview of the IRT and its procedures as applied to test item development and analysis. The paper concludes with some suggestions for test developers and test specialists at all levels to adopt IRT for its identified crucial theoretical and empirical gains over CTT. IRT based parameter estimates should be superior and reliable than CTT based parameter estimates. With these features, IRT can help resolve the problems associated with test design based on CTT

Corresponding Author:

Rohaya Talib,
Faculty of Education,
Universiti Teknologi Malaysia,
81310 Johor Bahru, Malaysia.
Email: rohayatalib@utm.my

1. INTRODUCTION

Assessment in education system serves a number of purposes, which includes improving instructional planning, acting as a mechanism to change instructional content, measuring learners' proficiency, comparison of student performances or achievement data, placement of students, determining a student fate (should he/she be retained or promoted) and holding schools and educators accountable. According to Gurski [1], tests might be helpful in evaluating the effectiveness of the instruction. The quality of such evaluations depends to a large extent on the nature and quality of the collected information during the assessment setting. Through the decades, the face of testing has undergone many changes. From oral to standardized testing, to authentic assessment, up to the present day it has continued to change with educational policy and practices.

High-stakes testing is used for the purposes of providing results that have important consequences such as, licensing, certifications or credentialing. Validity is the cornerstone upon which all measurement

systems are built. In educational measurements, the validity of inferences drawn from test results must be sound and well-grounded on principles and empirical evidence which should be able to withstand criticism [2]. It is clear that quality assessment is likely to lead to improvements in student learning [3].

Today testing is everywhere in our educational system; and with advancements in its design and technology, coupled with the advent of the "Age of Accountability", testing is an important element of education. However, for test quality to be established certain criteria on the areas of test design, test analysis techniques and test score interpretation must be met. In any official test, more especially in large scale assessment, the question of reliability and validity is of great concern. Educational assessment should follow the established criteria and guidelines of valid and reliable test development.

The process of test development includes five different steps, including test conceptualization, test construction, test try-out, analysis, and revision [4]. As can be seen, item analysis follows an initial try-out of the test. The goal of item analysis methods is to maximize the psychometric quality of scores from the test [5]. There are educational methods of assessing the items in a newly created test, which help in achieving accuracy and elaboration of the test results. It is important for a test developer to be familiar with the different test development and item analysis methods in order to facilitate the development of a new test. These frameworks are Classical Test Theory (CTT) and Item Response Theory (IRT). The estimates of validity and reliability of test items depends on a particular measurement model used.

This paper therefore discusses the IRT framework, its assumptions its application in the development of test which includes validity and reliability in IRT, item analysis and Items selection. Using IRT to develop test items will overcome the limitations of the CTT. IRT can produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered [6].

2. ITEM RESPONSE THEORY

Item response theory or the latent trait models provide a rich statistical tool for analysis of educational test and psychological measurement scale. The IRT methods were largely developed in the 1960s through 1980s, though, as Bock [7] noted in his "Brief Historical Review of Item Response Theory". The foundation for these models began with Thurstone in the 1920s. In his paper titled "A Method of Scaling Psychological and Educational Tests." He provides a technique for placing the items of the [8] test of children's mental development on an age-graded scale. Another work that serves as a basis for later development of IRT was the [9] book entitled "Statistical Theories of Mental Test Scores". They provide a rigorous and unified statistical treatment of classical test theory, particularly the chapters written by Birnbaum in this book. Some of the recent collection and texts on the development and application of IRT include: [10]; [11]; [12]; [13]; [14]; and [15].

[6] Described Item Response Theory as a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test. Item responses can either be discrete or continuous and can be dichotomously or polychotomously scored; item score categories can be ordered or unordered; there can be one ability or many abilities underlying test performance; and there are many ways (i.e., models) in which the relationship between item responses and the underlying ability or abilities can be specified. Within the general IRT framework, many models shave been formulated and applied to real test data. According to [16] "Item Response Theory (IRT), based on latent trait theory, incorporates measurement assumptions about examinee item and test performance, and how performance relates to knowledge as measured by the items on a test. Within the general IRT framework, many models have been formulated. Famous names associated with these various scoring models are dichotomous, binomial, Poisson, rating scale, facets, multinomial logit, or polytomous. These scoring models handle item responses that are discrete or continuous and dichotomous or polytomous scored" (P. 1).

The characteristics of Item Response Models, as summarised by [17] are, first, an IRT model must specify the relationship between the observed response and underlying unobservable construct. Secondly, the model must provide a way to estimate scores on the ability. Third, the examinee's scores will be the basis for estimation of the underlying construct. Finally, an IRT model assumes that the performance of an examinee can be entirely predicted or explained by one or more abilities. In item response theory, it is often assumed that an examinee has some latent, unobservable trait (also called the ability), which cannot be studied directly. The purpose of IRT is to propose models that permit to link this latent trait to some observable characteristics of the examinee, especially his/her faculties to correctly answering to a set of questions that form a test [18].

Item Response Theory, item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Three most commonly used IRT models are; one-parameter logistic model (1PLM or Rasch model), two-parameter logistic model (2PLM) and three parameter logistics model

(3PLM). 1PM possesses only item difficulty parameter (b), the 2PLM in addition to (b) possess a second parameter known as discrimination parameter (a), which allows the items to differentiate or discriminate the examinees of different abilities. The 3PLM in addition to the (b) and (a) contains a third parameter, known as the pseudo-chance parameter (c). As noted by [10] the pseudo-chance or guessing parameter corresponds to the lower asymptote of the item characteristic curve which represents the probability that low-ability examinees will answer an item correctly in a test and provide an estimate of the guessing parameter

2.1. Assumptions of Item Response Theory

When identifying the major assumptions of the Item Response Theory stated that, the first assumption, [19] states that if the examinee knows the correct answer to the item, he/she will go directly to answer it correctly, this assumption relates to any test theory. Without this assumption, there may not be a good reason for testing. The two other strong assumptions of IRT are Unidimensionality and Local independence. These assumptions are paramount and should hold irrespective of the latent trait model used. This means test data can only be valid for latent trait model estimation only if these assumptions are met.

i. Unidimensionality

Unidimensionality state that there is only one ability being measured. Another researcher [19] further explain that, the theory of latent trait assumes that a set of traits underlies test performance. The examinee's ability in a set of unidimensional latent space can be represented by a vector of ability scores as (i.e., θ1, θ2, θ3….. θn). The Item response models that assume a single latent ability is referred to as unidimensional. This assumption means that the items measure only one area of ability or knowledge. The condition of unidimensionality does not portend that the items must correlate positively with each other. An item may negatively correlate with others item and can still be unidimensional.

The assumption of Unidimensionality requires that all items on a test measure a single latent trait and violation of this unidimensionality would lead to serious misleading result. The assumption can be satisfied if a single dominant factor underlie responses. Unidimensionality IRT analysis assumes the presence of a dominant ability or trait that influences test performance- which is called unidimensionality [13]. In other words, unidimensionality refers that there exist a single latent trait variable to explain the variability of observed score as well as assumption for the test development in classical test theory.

ii. Local Independence

The assumption of local independence means that, the probability of an examinee getting item correctly is not affected by the answer given to other items in the test. It necessitates that excluding the ability there is no relationship between the test item responses other than the relationship determined by the ability or other model parameters. For example, if the responses to one item structurally constrain the possible answers to other items, then the items are not locally independent. If these assumptions are met, an IRT model can be successfully employed [20]. Local independence means the performance on different items is independent but conditional on the student's ability and does not mean that items do not correlate with each other. suggests that, there is no correlation between test items when person's ability level is controlled [13].

Item Response Theory - the generalized model

$$P_g(\theta) = c_g + (1 - c_g)\frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}} \quad (1)$$

Where:

ag= gradient of the ICC at the point q(item discrimination)
bg = the ability level at which ag is maximized (item difficulty)
cg = probability of low persons correctly answering a question (or endorsing) g

Item Response Theory Models

Schumacker [16] summarised 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. 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 guessing, difficulty and*

discrimination. A popular one-parameter model, developed by George Rasch, is also commonly used where item difficulty provides an unbiased, efficient, sufficient, and consistent estimate of separate person and item calibrations (P. 1)

The One-Parameter Logistic Model (item difficulty)

The 1-parameter model explains the relationship between the ability and probability of a correct response on the item in terms of the item difficulty. An item's difficulty parameter (b) is the point on the ability scale corresponding to the location on the item characteristic curve (ICC) where the probability of a correct response is 0.5 [21].

$$P(\theta) = e^{a(\theta)}/1 - e^{a(\theta)} \quad (2)$$

Where:

P (θ) = ability of a student and
a(θ) = difficulty level of item and
e=2.73 = discrimination index
b(θ) = 1 in this model.

In 1PLM item discrimination is taken as 1 and this may not be of great utility where sharp measurement is required e.g. examinees with equal raw score in a test will have equal IRT score and thus may fail to produce ranking

The Two –Parameter Logistic Model (item difficulty and discrimination)

The 2-Parameter model makes use of the *b* parameter (item difficulty) just as in the 1PLM, and addition add an element that indicates how wills an item separates students into different ability levels this parameter is called item discrimination (a). The item discrimination (a) parameter used in the 2PLM is equal to the slope of the item characteristics curve when it is at its steepest [21].

$$P(\theta) = \frac{1}{1 + e^{-1.7a(\theta - b)}} \quad (3)$$

The Three-Parameter Logistic Model (Difficulty, discrimination, and guessing)

The 3PL model builds upon the two-parameter model by adding pseudo-chance-level parameter *c*. The *c* parameter is the value of the lower asymptote of the item characteristic curve and is indicative of the probability that an examinee with a very low ability score would answer an item correctly.

$$P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}} \quad (4)$$

All IRT models are derived to generate item characteristic curves. An item characteristic curve plots the probability that an examinee will respond correctly to an item solely as a function of the test's latent trait [21]. [17] noted "The main difference to be found in currently popular item response models is in the mathematical form of $P_i(\theta)$, the ICC. It is up to the test developer or IRT user to choose one of the many mathematical functions to serve as the form of the ICCs".

The values on the X-axis of an ICC represent the latent trait, usually ranging from -3 to +3. The Y-axis represents the probability of an examinee's success. As the latent trait increases, the probability of the examinee responding correctly will increase but with diminishing returns. In their discussion of item characteristic curves, [21] discussed two interpretations that they consider acceptable. The first interpretation of a correct response is "the probability that a randomly chosen member of a homogeneous subpopulation will respond correctly to an item" (p. 341). A second interpretation is that the probability represents the probability of a specific examinee responding correctly for a sub-population of items.

In this study, only a few of the models that (a) assume a single ability underlies test performance, (b) can be applied to dichotomously scored data, and (c) assume the relationship between item performance and ability is given by a one-, two-, or three-parameter logistic function will be considered. Typically, two assumptions are made in specifying IRT models: One relates to the dimensional structure of the test data, and the other relates to the mathematical form of the item characteristic function or curve (denoted ICC).

IRT - Item Characteristic Curves

An ICC is a plot of the respondents' ability (likeliness to endorse) over the probability of them correctly answering the question (endorsing). The higher the ability, the higher the chance that they will respond correctly. The probability of a correct response is determined by the item's difficulty and the examinee's ability. This probability can be seen as illustrated using item characteristic curve (ICC) in Figure 1.

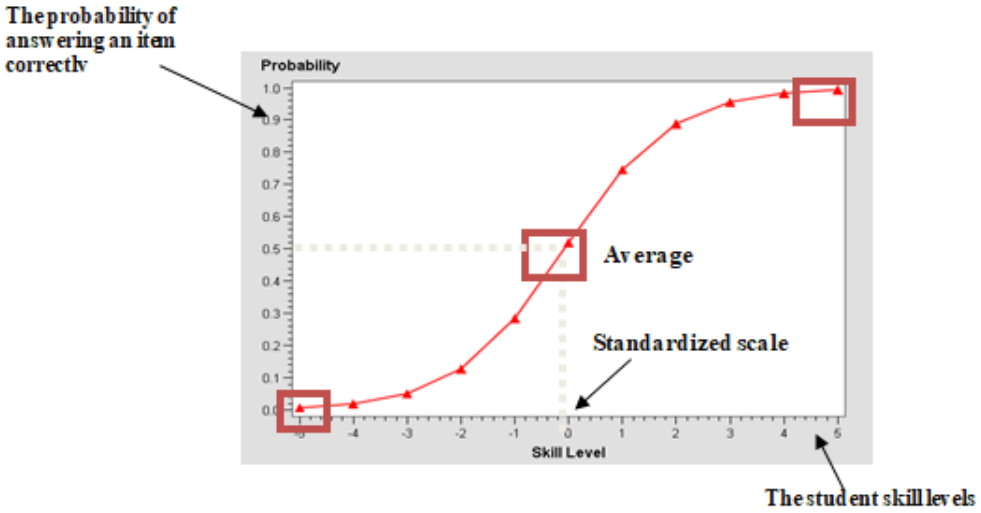

Figure 1. Item characteristic curve

From this ICC above, we can observe that as the examinee's ability increases, the probability of a correct response increases; this is what you would expect in practice.

As given earlier the item difficulty (a-value) measures the difficulty of answering an item correctly. The preceding discussion and equation suggests that the probability of endorsing an item correctly or a correct response is 0.5 for any examinee whose ability is equal to the value of the difficulty parameterpractice. Figure 2 and Figure 3 show the ICCs of two different items

Difficult Item vs. Easy Item

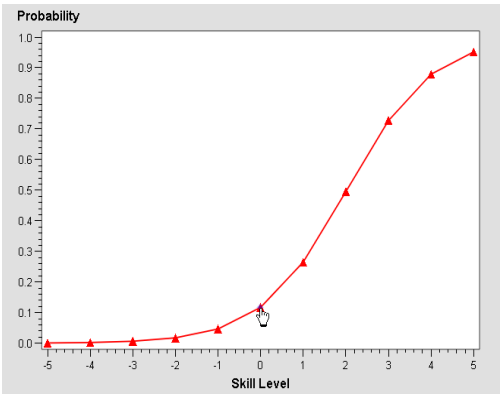

Figure 2. ICC in case of a difficult item

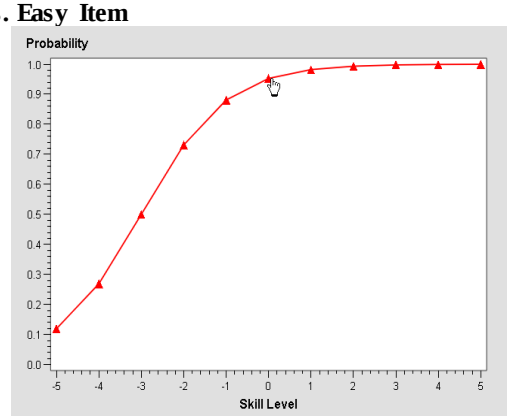

Figure 3. ICC in case of an easy item

Figure 2 and Figure 3 show the ICCs of two different items, with different item difficulty parameters and probability of endorsing an item correctly or a correct response. By comparing these two ICCs, we can see that the item difficulty parameter determines the location of the ICC. From Figure 2 the probability of them correctly answering the question (endorsing) is higher at 0.94, in order to get a 0.94 probability of a correct response for this item, the examinee must have higher ability skill level of about 5 to be able to get

the item, this signifies that, the item is difficult, because it can only be answered by a higher ability students. On the other hand, Figure 3 the probability of answering the question correctly (endorsing) is higher at about 0.94, in order to get a 0.94 probability of a correct response for this item, the examinee must have moderate ability skill level of about 0.7 to be able to get the item correctly, this signifies that, the item is easy, because it can only be answered by an examinee with moderate ability levels

2.2. Validity and Reliability in Item Response Theory (IRT)

a) Validity in IRT

The meaning of validity and reliability in IRT differ from that of CTT. The IRT focuses on the characters of the item. A validity in IRT means to what extent individual examinees and test items have a good ranking in the ability which the test items measure, this means the ability of any test to rank an individual according to their ability as well as rank the items according to their level of difficulty [22].

b) Reliability in IRT: Item and Test Information Functions

The reliability in IRT means to what extent the scores are independent of groups (samples) as well as from the items, in other words the characteristics of test items is not affected by the samples from which they were estimated, and even if same items were administered to different group it provide the same score and ranking.

Similarly, reliability according to IRT is an item and test information or, the degree to which an investigator or researcher can be certain of a person's location along θ. The amount of item information is proportionate to the standard error of estimate (SEE) for each possible θ [23]. A smaller SEE indicates a stronger certainty in the estimate of θ and therefore more information about individuals with that particular θ value. By rule, an item provides its highest amount of information near its difficulty value ("b") because there is the least amount of variability (error) near this value [24]. Figure 4 shows test information function

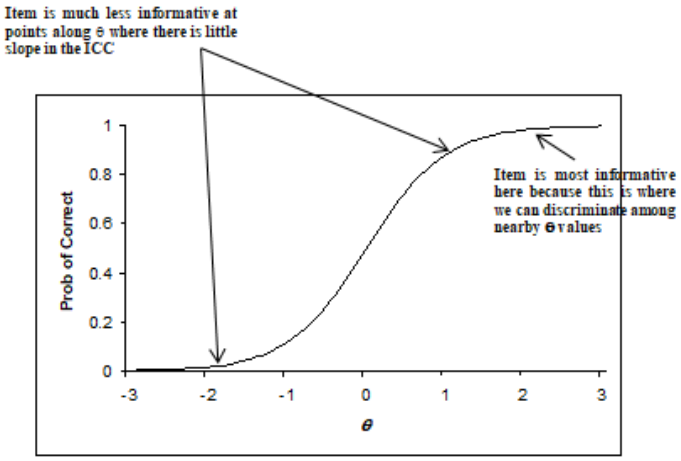

Figure 4. Test Information Function

The three IRT models are used to evaluate the validity and reliability of items test according to the three parameters. The ability of the examinee, level of item difficulty and ability of item to discriminate

2.3. Application of IRT in Test Development Process

The pendulum swing in test development techniques is from Classical Test Theory (CTT) to Item Response Theory (IRT). According to [25], application of IRT in test development process is a recent trend which marks a departure from the traditional practice of basing test development on CTT. With IRT, items are calibrated without reference to the sample in terms of the trait level or ability level of an individual referred to as theta (ɵ) and item parameter estimates. The item parameter estimates are item discrimination power (parameter a), item difficulty (parameter b) and guessing (parameter c). Parameter 'a' indicates degree to which a person's response to an item relates or varies with his/her trait level or ability; parameter 'b' indicates amount of trait in the item; while parameter 'c' indicates probability that a person who does not possess the trait will respond to an item correctly. IRT is considered to be a powerful method of item

selection, which provides different estimates of error of measurement at each ability level. Its applicability in developing better tests, item bias, differential item functioning, item banking and tailored testing has been stressed [26; 27; 25].

2.4. Item Analysis In IRT

When employing item response theory, item analysis consists of (a) determining sample-invariant item parameters using relatively complex mathematical techniques and large sample sizes, and (b) utilizing goodness-of-fit criteria to detect items that do not fit the specified response model. The property of sample invariance inherent within IRT means that test developers do not need a representative sample of the examinee population to calibrate test items. They do, however, need a heterogeneous and large examinee sample to ensure proper item parameter estimation. However, the test developer using IRT is faced with a different problem. Because IRT requires larger sample sizes to obtain good item parameter estimates, the test developer must ensure that the examinee sample is of sufficient size to guarantee accurate item calibration. Poor items are usually identified through a consideration of their discrimination indices (the value of ai will be a low positive or even negative) and difficulty indices (items should neither be very difficult nor too easy for the group of students to be assessed) [28].

The *a, b* and *c* parameters

1. The *a* parameter

One feature of a good test item is that high-ability students will answer it correctly more frequently than lower-ability students. The item discrimination 'a' parameter expresses how well an item can differentiate among examinees with different ability. 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 (a-values) 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. The discrimination values of a good test item ranges between 0.5 to 2 and the steeper the slope of an item characteristic curve, the higher an item's discrimination values (a-values). High discrimination level indicates that the item discriminates well between low and high skilled individuals. A discrimination parameter is a measure that can be graphically expressed by the steepness of the item characteristics curve (ICC). The Item discrimination value (a-values) above 1 is normally desirable for a good test item and a-values above 0.75 can also be acceptable. Interpreting discrimination parameter values are presented in Table 1.

Table 1. Interpreting Discrimination Parameter Values [29]

Discrimination Value	Quality of an Item
$a \geq 1.70$	Item is functioning quite satisfactorily
$1.35 \leq a \leq 1.69$	Good item; little or no revision is required
$0.65 \leq a \leq 1.34$	Moderate: little or no revision is require
$0.35 \leq a \leq 0.64$	Item is marginal and need revision
$a \leq 0.34$	Poor item; should be eliminated or revised

2. The *b* parameter

Item difficulty refers to the b parameter, is the point where the S-shaped curve has the steepest slope. Examinee must have higher ability in order to answer a difficult item correctly. Item with high b values is a hard or difficult item, that is, value of b greater than 1 indicates a difficult item and low-ability examinees are more likely to fail because they will find it difficult to answer correctly. Similarly, an item with low b value below -1 indicate easy item, which most of the examinees with low ability level, will have at least a moderate chance of answering it correctly. When an item has a b-value of between -1.00 to 1.00, this value indicates an item with moderate difficulty [29]. Interpreting item difficulty values are presented in Table 2.

Table 2. Interpreting Item Difficulty Values [29]

Difficulty value *(b)*	Interpretation
$-3.00 \leq -2.00$	Very easy
$-2.00 \leq -1.00$	Easy
$-1.00 \leq 1.00$	Moderately difficult
$1.00 \leq 2.00$	Difficult
$b > 2.00$	Very difficult

3. The c parameter

The 3PLM include a pseudo-guessing parameter known as c-parameter, this parameter expresses the probability that an examinee with low ability can be able to get an item correctly 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. For example, a student, who randomly selects responses to items that have four response choices can answer these items correctly about 1 out of 4 times, meaning that the probability of guessing correctly is about 0.25.

2.5. Item Selection in IRT

As is the case with classical test theory, item response theory also bases its item selection on the purpose of the test. The final selection of test items will depend on the information each item contribute to the overall information supplied by the test. An especially useful feature 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 information function independent of other items in the test. As outlined by [15] a procedure, originally conceptualized by [30], for the use of item information functions in the test building process. This procedure entails that a test developer take the following four steps:

First, describe the shape of the desired test information function over the desired range of abilities. [15] calls this the target information function. Second, select items with item information functions that will fill up the hard-to-fill areas under the target information functions. Third, after each item is added, then calculate the test information function for the selected test items. Fourth, select items until the test information function approximate the target information function to a satisfactory level. Lastly, content validation considerations are monitored during the item selection process.

This procedure allows the test developer to build a test that will precisely fulfil any set of desired test specifications. Thus, it is possible to build a test that "discriminates" well at an; particular region on the ability continuum. That is to say, if we have a good idea of the ability of a group of examinees test items can be selected so as to maximise test information, the region of ability spanned by the examinees being tested, of course, this optimum selection of test items will contribute substantially to the precision with which ability scores are estimated. Furthermore, with criterion-referenced tests, it is common to observe lower test performance on a pre-test than on a post-test. Given this knowledge, a test instructor should select easier items for the pre-test and more difficult items for the post-test. Then, for both testing administrations, measurement precision will have been maximized in the ability region where the examinees would most likely be located. Moreover, because items on both tests measure the same ability, and ability estimates are independent of the particular choice of test items, the instructor can measure growth by subtracting the pre-test ability estimate from the post-test ability estimate

2.6. Benefits of Item Response Theory

Item Response Theory measurement models, when compared to classical models, offer several distinct advantages. These include the following: (a) Item statistics are independent of the sample from which they were estimated (b) Examinee scores are independent of test difficulty (c) Item analysis accommodates matching test items to examinee knowledge level (d) Test analysis doesn't require strict parallel tests for assessing reliability (E) Item statistics and examinee ability are both reported on the same scale [31].

2.7. Limitations of Item Response Theory

IRT models have several technical and practical shortcomings. Assumptions underlying the use of IRT models are more stringent than those required of classical test theory. IRT models also tend to be more complex and the model outputs harder to understand, particularly with non-technically oriented audiences. Additionally, IRT models require large samples to obtain accurate and stable parameter estimates, although Rasch measurement models are useful with small to moderate samples. Consequently, the choice of a model may depend on the sample available, particularly in the field-testing phase of a certification exam.

3. CONCLUSION

In conclusion, there are many limitations in the CTT that concerns with calibration of item difficulty, sample dependency of coefficient measures, and estimates of measurement error which in turn is addressed by the IRT. example IRT models represent the ability of the examinees and the difficulty of the items as independent parameters. It can separate these two parameters empirically in a way that no other psychometric models can do. Similarly, In contrast to CTT the estimation framework of IRT models make it straightforward to analyse items that have random missing data. IRT can still calibrate items and score subjects by using all the available information based on the likelihood; the likelihood-based methods are

implemented in the IRT procedure. IRT differs considerably from the linear approach to test and item analysis (CTT) and has some identified crucial theoretical and empirical gains over CTT due to this, it is expected that there would be appreciable differences between the two and IRT based parameter estimates should be superior and reliable than CTT based parameter estimates. With these features, IRT can help resolve the problems associated with test design based on CTT.

REFERENCES

[1] Gurski, L. F. "*Secondary Teachers' Assessment and Grading Practices in Inclusive Classrooms"*. A Thesis Submitted to the College of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of Master of Education, University of Saskatchewan, 2008.

[2] Chester, M. D. "*Ensuring Technical Quality: Policies and Procedures Guiding the Development of the MCAS Tests"*. Technical Report of the Massachusetts Department of Elementary and Secondary Education, 2008 http://www.doe.mass.edu/mcasappeals

[3] Hamilton, L. S., Stecher, B. M., & Klein, S. P. "*Making Sense of Test-Based Accountability in Education"*. Santa Monica, *CA: RAND, 2000*

[4] Cohen, R. J., Swerdlick, M. E. & Phillips, S. M. Test development. In Psychological testing and assessment: An introduction to test and measurement (3rd Ed). Mountain View, CA: Mayfield, 1996.

[5] Ebel, R. L. and Frisbie, D. A. "*Essentials of Educational Measurement (5th Ed)"*. Prentice Hall, Engelwood Cliffs, New Jersey, 1986, ISBN: 13-9780132846134, Pages: 370.

[6] Hambleton, R. K., & Jones, R. W. "Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development". *Educational Measurement: Issues and Practice, 12*(3), 38-47, 1993

[7] Bock, R. D. "A Brief History of Item Response Theory." *Educational Measurement: Issues and Practice, 16*(4), 21–33, 1997.

[8] Binet, A., & Simon, T. "Methods Nouvelles Pour Le Diagnostic Du Nieveau Intellectual Anormoux [New methods for the diagnosis of levels of intellectual abnormality]". *AnneePsychologique*, 1905

[9] Lord, F. M. & Novick, M. R. "*Statistical Theories of Mental Health Scores"*. Addison Wesley, Reading, MA, 1968

[10] Embretson, S. E. & Reise, S. P. "*Item response theory for psychologists"*. Erlbaum,Mahwah, NJ, 2000

[11] Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag

[12] Heinen, T. "*Latent Class and Discrete Latent Trait Models"*. Sage, Thousand Oaks, CA, 1996

[13] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. "*Fundamentals of Item Response Theory"*. Newbury Park, CA: Sage Publications, 1991

[14] Hulin, C. L., Drasgow, F. S., & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL: Irwin

[15] Lord, F. M. "*Applications of Item Response Theory to Practical Testing Problems"*. Lawrence Erlbaum Associates, Inc. New Jersey, 1980

[16] Schumacker, R. E. "Item Response Theory". 2010[b] http://appliedmeasurementassociates.com/ama/assets/File/ITEM_RESPONSE_THEORY.pd. Retrieved on 13 August, 2017.

[17] Hambleton, R. K., & Swaminathan, H. "*Item response theory: Principles and applications* (Vol. 7)": Springer, 1985

[18] Magis, D. "Influence, Information and Item Response Theory in Discrete Data Analysis". 2007 http://bictel.ulg.ac.be/ETD-db/collection/available/ULgetd 06122007-100147/. Accessed on 20 June, 2014.

[19] Ojerinde, D. "Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of the Comparability of Item Analysis Results". A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May, 2013

[20] Courville, T. G. "*An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics"*. Unpublished Ph.D Dissertation, Texas A & M University, 2004.

[21] Crocker, L. & Algina, J. "*Introduction to Classical and Modern Test Theory"*. Holt, Rinehart and Winston, New York, USA., ISBN-13: 9780030616341, Pages: 527, 1986

[22] Hambleton, R. K. "The Rise and fall of criterion referenced measurement". *Educational Measurement: Issues and Practice, 13(4). 21-26, 1994*

[23] De Ayala, R. J. "*The theory and practice of item response theory"*. New York: Guilford Press, 2009

[24] DeMars, C. (2010). Item Response Theory (Understanding Statistics). Oxford, Oxford University Press. ISBN-13: 978-0195377033

[25] Nworgu, B. G. "Challenges of Quality of Assessment in a Changing Global Economy". *Journal of Educational Assessment in Africa*, 5*:* 13-35, 2010

[26] Nenty, H. J. "From Classical Test Theory (CTT) to Item Response Theory (IRT): An Introduction to Desirable Transition". In O. A. Afemikhe & J. B. Adewale (Eds.), *Issues in Educational Measurement and Evaluation in Nigeria*. Ibadan: Educational Research and Study Group, 2004