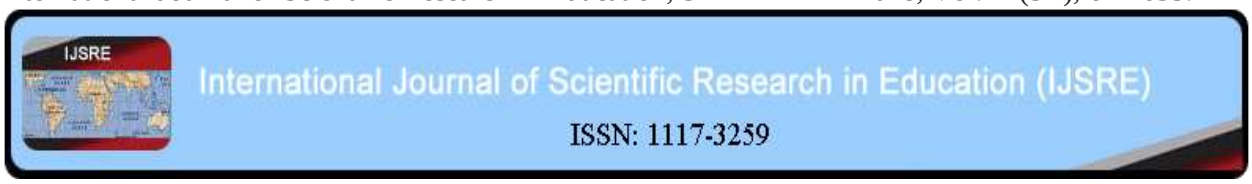
International Journal of Scientific Research in Education, SEPTEMBER 2018, Vol. 11(3B), 627-635.



Item Response Theory (IRT): A Modern Statistical Theory for Solving Measurement Problem in 21st Century

Isaac Bamikole Ogunsakin[i]
Department of Educational Foundations and Counseling, Faculty of Education,
Obafemi Awolowo University, Ile-Ife
sakinbamikole@yahoo.com

Yusuf. O. Shogbesan[ii]
Department of Educational Foundations and Counseling, Faculty of Education,
Obafemi Awolowo University, Ile-Ife
yusufolayinka813@yahoo.com

Abstract

Item Response Theory is a measurement framework used in the design and analysis of educational and psychological assessments (achievement tests, rating scales, inventories, or other instruments) that measure mental traits. This is becoming increasingly popular among educational assessment experts to analyse cognitive measurement but little is known about the theory to analyse non-cognitive measurement. (E.g. personality, attitude, and psychopathology) This paper provides information on models, assumptions and application of IRT in the analysis of cognitive and no-cognitive measurements. The paper concludes that, IRT is a better framework that can be exploited by researchers in analyzing cognitive data for assessment and evaluation research, and non-cognitive data, for sociological, psychological and psychopathological assessments. However, all the statistical assumptions must be met, and the test data must fit the IRT model for valid, reliable and credible results.

Keywords: Item Response Theory (IRT), Classical Test Theory (CTT), Cognitive, Non-Cognitive, Measurement.

Reference to this paper should be made as follows:

Ogunsakin, I. B., & Shogbesan, Y. O. (2018). Item Response Theory (IRT): A Modern Statistical Theory for Solving Measurement Problem in 21st Century. *International Journal of Scientific Research in Education, 11*(3B), 627-635. Retrieved [DATE] from http://www.ijsre.com.

INTRODUCTION

Identifying cognitive abilities of a test-taker and representing them as a reliable numerical score is the main purpose of educational and psychological measurements. However, the impact and outcome of education cannot be felt in our educational sector without measurement. This shows that measurements play clear roles in education which help in given feedback and help in decision making. The roles of measurement cannot be ruled out in education except there is no concern for feedback. The cardinal aim of the measurements is centered on diagnosing, examining or ranking individuals in order to acquire vital information that can be used for different types of activities or decisions. In order for these activities or decisions to be valid for its purpose, measurement has to give very reliable and secured information.

This can be archived by using a reliable instrument to carry out the act. It is not sufficient that the measuring instrument is of high quality in itself, it must also be used and interpreted on the basis of various quality requirements. The greater the consequences the measurement has for the individual of those who use the result, the greater are these quality requirements. The subject of educational measurement is about how relevant and reliable instruments are constructed and used in various different areas in society. Predominant issues concerned are and how a measurement is implemented. How the measurement functions, and what consequences it has for individuals, groups and society?

In educational measurement, certain theories help to attain reliable result among which are Classical Test Theory (CTT) and Item Response Theory (IRT). The CTT is a relatively simple theory for testing and has been widely used for constructing and evaluating tests, particularly before the birth of IRT. Hambleton and Jones (1993) state categorically that CTT models have served measurement specialists for a long time and have some positive points. The CTT models are straightforward, easy to understand and applicable in testing practice. Many useful models and methods have been formulated from the CTT framework and have addressed effectively important issues in measurement. Moreover, the mathematical analyses required for CTT models are usually simple compared to those required in IRT and do not require strict goodness–of–fit study to ensure the good fit of a model to actual test data. Unlike IRT models, CTT models do not require large sample sizes for analyses and basically depends on three basic mathematical concepts which are test score, true score and error score.

Minh (2004) makes a clear exposition on the CTT models that are fundamental: test score (sometimes called observed score), true score, and error score. Within that framework, many CTT models have been developed. In a simple CTT model, the observed test score X of an examinee for a certain test item can be expressed in terms of true score T and error score E in a simple equation:

X	**=**	**T**	**+**	**E**	1.0
Observed scored		**True score**		**Error**	

True score can be defined as expected observed test score over parallel tests or as the true but unknown value of the examinee on the ability being tested (Lord and Novick 1968). The error score is the difference between observed score and true score. The true score and error score are called latent (or unobservable) variables since they cannot be directly observed. Because there are two unknown variables (true score and error score) in the equation, we cannot solve it unless some assumptions are made. There are three fundamental assumptions in CTT models: true score and error score are not correlated, the average error score over population of examinee is zero, and error scores on parallel tests are not correlated (Hambleton & Jones, 1993).

Generally, parallel tests can be considered as tests that measure the same latent ability for which examinees have the same true score and the errors across tests are equal. CTT being the old form of measurement theory has done a lot in measurement before the advent of IRT. Ivailo (2004) asserts that the ultimate aim of both Classical Test Theory (CTT) and Item Response Theory (IRT) is to test latent ability of examinee. Hence, their primary interest is focused on establishing the position of the individual along some latent dimension. The latent trait is often called ability, but in other contexts it might be anxiety, neurosis or simply the authoritarian personality. However, the CTT and IRT approaches offer an interesting mix of both theoretical and practical benefits and shortcomings, with IRT emerging as a clear favorite in educational practices in 21st century.

Moreover, CTT is applicable in some areas but cannot be totally relied on though it has done excellently well. Dibu (2013) affirms that the Classical Test Theory (CTT) has served measurement practitioners for several decades as the foundation measurement theory. The conceptual groundwork, assumptions and extension of the basic principles of CTT have allowed for the development of some excellent psychometrically sound scales. Conversely, the shortcomings of CTT in measurements include the following: the true score is not an absolute characteristics of a test taker as it depends on the content of test, yet it is considered as true score. The two parameters item difficulty (p) and item discrimination (r) which form the cornerstones of CTT are group dependent; CTT does not take into consideration students responses to any specific item and score obtained are entirely test dependent. Consequently test difficulty directly affects the test scores (Xiato, 1988).

It was as a result of the weakness of CTT that IRT came into practice. Consequently, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. The IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991) has been widely accepted within the measurement community. It was further ascertained by Dibu (2013) that 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 is able to estimate the parameters of an item independent of the characteristics of both test takers to which it is exposed and other items that constitute the test. Four prominent equations termed 1PL, 2PL, 3PL and 4PL (parameter logistic) models are presently used to make predictions. These models are the cornerstone of IRT; they are the pivots upon which the theory depends and they reveal information about the latent behavior of the items and the examinee which make it easy for measurement community to make right predictions.

Models of IRT

Each IRT model's major cardinal aim is to predict the probability that a certain person will give a certain response to a certain item. The ability level can vary according to individual and items can differ in many respects; most importantly, some are easier and some are more difficult. In the IRT, the probability is denoted with Pij instead of simply P: the index i refer to the item and the index j refers to the person. When an item allows for more than two options, we shall also need an index for the options. Also, P (θ) is written to show that the probability of a correct response is a function of the ability θ. However, P also depends on the properties of the item as captured

by the item parameters. For dichotomous items, we shall examine IRT models having one, two, three or four parameters and the probabilities predicted by the models will be denoted as Pi j(θj , bi), Pi j(θj , bi , ai), or Pi j(θj , bi , ai , ci, di), where ai , bi , ci and di are all item parameters. However, 'θ' denote ability level, 'b' denotes difficulty parameter 'a' denotes discrimination parameter 'c' denote guessing parameter and 'd' denote carelessness parameter. The IRT models provide information about the individual items and the learner responses to those items. The collation of all person-item responses certainly provides richer data with which to assess the validity of the test construct and the reliability of the instrument. The Rasch model provides a "stringent modeling tool" that is useful when the data fit the model and that will provide information about test inequality or differential item functioning when misfit is identified (Ryan & Williams, 2005). This shows the extent to which IRT is very important in measurement. However its role cannot be over emphasized since it gives detail information about the examinee in relation to each of the item responded to. IRT has been a viable theory in measurement community in that it is helpful in analyzing cognitive and non-cognitive data. However, for valid, reliable and credible results to emanate, the models and the assumptions must play a crucial role which cannot be overemphasized. This shows that the assumptions and the model are the engine room of the theory.

The One-Parameter Logistic (1PL) Model

Ogunsakin (2015) observes that the one-parameter model is most widely used and the simplest of the three IRT models. In the IRT-1PL model, the only parameter that is estimated by the model is the difficulty parameter bi. This is presented in the equation below.

$$P_i(\theta) = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{exp}(\boldsymbol{\theta}-\boldsymbol{b_i})} \qquad 1.1$$

From equation 1.1 the model presumes that the probability that a student will correctly answer a question is a logistic function of the difference between the student's ability 'θ' and the difficulty of the question 'b' $\{1 + exp(\theta - b_i)\}$. 1-PL is equivalent to the Rasch model which, although takes the student's ability and the difficulty of the question into account, is slightly different to the 1 PL model. In the Rasch model, each individual in the person sample has parameters defined for item estimation. On the other hand, when the person sample has the parameters defined by a mean and standard deviation for item estimation, it is called the 1PL model of IRT.

The Two-Parameter Logistic (2PL) Model

According to David (2011), the two-parameter model has the same function as presented for the one-parameter model. However, in the two parameter model, the item discrimination parameter will vary across items, as does the item difficulty parameter. The IRT-2PL model differs from the 3PL model only in that it assumes that pseudo-guessing is not a meaningful contributor to item fit, or more typically, that it is not applicable to the data (e.g., in the case of rater-scored constructed-response items. David further explains that the only difference is the absence of the guessing parameter, ci.

$$P_i(\theta) = \frac{1}{1+exp[-Da_i(\theta-b_i)]} \qquad 1.2$$

The Three-Parameter Logistic (3PL) Model

The three parameters being used for these models are the 'b' 'a' and 'c' parameters which are the difficulty, discrimination and guessing parameters. The three-parameter model includes a guessing parameter especially useful for multiple-choice and true-false testing. The mathematically representation of item characteristics curves (ICC) of 3-PL is given equation 1.3.

$$P_i(\theta)=c_i + \frac{1-c_i}{1+exp[-Da_i(\theta-b_i)]} \quad 1.3$$

The three parameters of the model are represented by ai, the discrimination power of item i, bi, the difficulty index that represents the value of the individual parameter (e.g., cognitive ability of a student). The function 1 [1+ exp−t] is a logistic function, with exp (–t) denoting e, the natural exponent. Within this logistic function, D is a multiplicative constant, typically set to 1.7 or 1.702, because this value helps the 2PL model approximate the normal ogive model (Yen & Fitzpatrick, 2006).

The Four-Parameter Logistic (4PL) Model

Four-parameter logistic (4PL) model as an extension of the usual three parameter logistic (3PL) model with an upper asymptote possibly different from1. The mathematical expression for 4PLM is given in equation (v).

$$P_i\ (Y_{is} = 1|\theta_s) = c_i + (d_i - c_i)\ \frac{exp(1.7a_i(\theta_s-b_i))}{1+exp[-Da_i(\theta_s-b_i)]} \quad 1.4$$

In equation 1.4 b*i* = item difficulty location, ai = item discrimination, slope ci = item lower asymptote "guessing" d*i* = item upper asymptote "carelessness" (so di < 1) Kalolina (2013). The four-parameter logistic model (4PLM) assumes that even high ability examinees can make mistakes (e.g. due to carelessness).

This phenomenon was reflected by the non-zero upper asymptote (d-parameter) of the IRT logistic curve. Karolina (2013) asserts further that 4PLM has been hampered, since the model has been considered conceptually and computationally complicate and its usefulness has been questioned. Although, the 4PLM has been abandoned for years due to its complexity in term of calculation which makes it unpopular in the measurement community. Evidence from the literature reveals that measurement experts have been making good use of it in recent years due to adequate information that it can reveal in the latent behavior of examinee and items. However, interest in 4PLM has been renewed in the field of clinical and personality psychology, as it is crucial to measure latent trait accurately at its extremes (Reise & Waller, 2003; Stark, Chernyshenko, Dras-gow & Williams, 2006; Waller & Reise, 2010). The model has been used in order to reduce the influence of examinees' early mistakes on estimation of their ability level and 4PLM has reduced such influence in a more effective way than 3PLM (Rulison & Loken, 2009; Loken & Rulison, 2010; Liao et al., 2012). Liao, Ho, Yen and Cheng (2012) also show that ability level was adequately estimated using 4PLM during Computerised Adaptive Testing (CAT) the model prevented the initial drop of ability estimates caused by the incorrect answers to the first two items.

Magis (2013) indicated that 4PL model allows more robust estimation of ability due to weighting the log-likelihood function (the aberrant item responses are down-weighted and have less impact on the estimation of ability).

Paradigm Shift from CTT to IRT

CTT has served measurement community for over 50 years, despite its simplicity in term of mathematical procedure, it has done more than imagined. IRT was introduced in place of CTT due to its inability to solve some measurement problems in research community. Minh (2004) affirms the limitations of CTT models and urges measurement specialist to develop new frame work to overcome those limitations. These limitations include the following among others: item difficulty and item discrimination are group dependent, the test score and the true scores are test dependent. This suggests that the testee's score depends on a particular set of test items being administered.

From this limitation, it was inferred by Minh that examinee usually has lower scores on difficulty test and higher scores on easier ones, although her ability remains constant over those tests. Another shortcoming of CTT is that CTT models lie on the assumption of equal errors of measurement across examinee. Concept of parallel test is another important limitation of CTT which IRT overcomes.

Literature studied shows that CTT is test oriented not item oriented. This shows that it is very difficult for researcher to predict and ascertain how an examinee respond to each of the test items. These leads to paradigm shift to IRT in order to overcome the limitation of CTT. Certainly, the framework for IRT that replaced CTT limitation include the following features: item statistics that are not group dependent , scores describing examinee's ability that are not test dependent, model that can solve the problem of parallel test and model that can interrelate items to examinee ability. Like CTT, IRT begins with the proposition that an examinee's response to a certain item is determined by an unobservable mental attribute of that examinee. That attribute is referred to as "trait" or "ability". Because traits are not directly observed, they are referred to as "latent traits" or "latent abilities".

IRT attempts to model the relationship between an examinee's latent ability and probability of the examinee correctly responding to a certain test item. (Minh 2004) points out that this relationship is modeled by a function called item characteristic function or item response function.

Assumptions of IRT

Literature reveals that there are several basic assumptions of IRT among others in measurement community that makes measurement of latent trait to be easier for Psychometricians in twenty first century; assumption about Unidimensionality, assumption about local independence, and assumption about mathematical form of the item characteristic curve

Assumption about Unidimensionality

Like classical test theory, IRT relies on a set of assumptions. First is the assumption of Unidimensionality. Central to all IRT models that is common to educational testing, uni-dimensionality is the assumption that all of the items on a test (or test section) measure only one

latent trait/ability/construct (e.g. reading comprehension, math proficiency, and verbal ability). This assumption is the basis of all measurement theory to the extent the sum of item scores is used to assign some overall value of ability to an examinee, as is the case on most tests. According to Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978), testing the assumption of uni-dimensionality takes precedence over all other goodness of fit tests because the results of all other tests will be difficult to interpret if the assumption of uni-dimensionality is untenable. The most common test of uni-dimensionality has been some variant of factor analysis. Stout (1984) was an early implementer of factor analysis in assessing uni-dimensionality, comparing classical and modern methods of factor analysis. However, Stout was not alone in suggesting a means of testing Unidimensionality.

Assumption about Local Independence

A second related but separate assumption central to IRT is Local Item Independence (LII), meaning the response to each item is not influenced by the response to any other item. In other words, LII is achieved if examinees' respective ability value (θ) explains fully their performance on all items. However, it assumes that item responses are independent given a subject's latent trait value, because the assumptions of Unidimensionality and local item independence are very strong assumptions.

Assumption about Monotonicity

All item characteristic functions are strictly monotonic in the latent trait. The item characteristic function describes the probability of a predefined response as a function of the latent trait.

Assumption about Item Characteristics Curve (ICC)

The key issue in IRT framework is the relationship between examinee's latent ability and probability of the examinee correctly responding to certain item. This relationship is modeled by a mathematical function called item characteristic function and the graph of this function is called Item Characteristic Curve (ICC). In simple terms, it is a linear or nonlinear function for the regression of item score on the trait or ability measured by the test (Hambleton, 1989). The assumption can be justified by how well the chosen IRT model accounts for the test data. Because the probability that an examinee correctly answers the item depends on the form of ICC. This probability is independent of distribution of examinee ability. Therefore, the probability of correct response by an examinee does not depend on number of examinees in the population who have the same ability.

An ICC presents the probability of responding correctly to an item as a function of the latent trait denoted by q underlying performance on the item (Crocker & Algina, 1986). The main difference found among IRT models is in the mathematical form of ICCs. Different IRT models employ different ICCs. The number of parameters required to describe an ICC depends on the chosen IRT model. Usually, an ICC has one, two, or three parameters that are called item parameters. Generally, an ICC can have linear or nonlinear form even though nonlinear forms are more useful. The most popular mathematical form of ICC is the logistic form whose graph is an S–shaped curve (Minh, 2004). Minh further explains that the S–shaped ICC (logistic form) just presented is widely used, we believe that it is not the only possible type of ICC. As a

mathematical function, an ICC can take any form. In fact, many people have proposed different forms of ICC such as step function or cubic function. We hope that those developments will bring significant insights on the relationship between ability and performance in the nearest future.

Application of IRT in Measurement in 21st Century

Application of IRT is indispensable in measurement community; it is applicable in numbers of areas as it has overcome numerous insurmountable mountains which CTT cannot overcome. Evidence from literature reveals that it is applicable not only in educational testing which is more familiar by researchers; it is applicable in analyzing non-cognitive data. Templin (2012) affirms that Item Response Theory (IRT) is used in a number of disciplines including sociology, political science, psychology, human development, business, and communications, as well as in education where it began as method for the analysis of educational tests.

Egberink (2010) reveals that tests and questionnaires play a crucial role in psychological assessment. Both cognitive measures (e.g., intelligence tests) and non-cognitive measures (e.g., mood questionnaires, personality questionnaires) belong to the practitioner's toolkit in different fields of psychology. For example, in personnel selection procedures besides intelligence testing, often personality questionnaires are used to assess whether a candidate is suitable for a particular job.

Also, in the clinical field both cognitive and non-cognitive measures are used for diagnostic purposes and to select the most appropriate treatment for the diagnosed disorder, because psychological tests and questionnaires are used to make important decisions, high-quality standards for the construction and the evaluation of these instruments are necessary. One of these standards is Item Response Theory (Embretson & Reise, 2000; van der Linden & Hambleton, 1997). Although there has been no shortage of researchers demonstrating the potential of IRT in the cognitive domains; its use in the non-cognitive measurement area (e.g., personality, attitude, and psychopathology) has lagged behind as a results of little awareness of IRT applications in non-cognitive data, in the study carried out by Steven (2009) it is affirms that the applications of IRT in the non-cognitive domains of personality and psychopathology assessment, more importantly in patient reported outcome (PRO) measurement, constructs such as pain, depression, and anxiety may display qualitative differences in symptom expression across countries, cultures, ages, and gender. However, the ability to identify variation in the manifestation of trait and retain the possibility of scaling individual differences on common metric can be achieved with aid of IRT.

In educational testing IRT has been in use for many decades with reliable results. As a cornerstone in measurement it can be used in test construction, matrix sampling, it helps to improve the quality of the tests and scales produce, test equating and administration, it helps to develop an understanding about differential item functioning, computerized adaptive testing, health numeracy and test scoring and interpretation among others.

In conclusion, IRT is a better framework that can be exploited by researchers in analyzing cognitive data in assessment, evaluation research and non-cognitive data, in sociological, psychological, psychopathology assessments However, all the statistical assumptions must be met, and the test data must fit the IRT model for valid, reliable and credible results.

REFERENCES

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Harcourt Brace. Educational Research Institute.

David Magis (2013). A note on the item information function of the four-parameter logistic model. *Applied British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525.

Egberink, I. L. (2010). Applications of item response theory to non-cognitive data Groningen: s.n.

Embretson, S. E., & Reise, S. P. (2000). Multivariate Applications Books Series. *Item response theory for psychologists. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers*.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 147-200). New York, NY, England: Macmillan Publishing. American Council on Education.

Item analysis and evaluation using

Ivailo, P. (2004). A visual guide to item response theory.

Karolina Swist (2015). Item analysis and evaluation using a four parameter logistic model items? *Psychological Methods, 8*(2), 164–184.

Loken, E., & Rulison, K. L. (2010). Estimation of a four parameter item response theory model.

Magis, D. (2013). NJ. Erlbaum. *Psychological Measurement, 37*(4), 304–315.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184.

Rulison, K. L. and Loken, E. (2009). I've fallen and I can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101.

Ryan, J., & Williams, J. (2005). Rasch modeling in Test Development, Evaluation, Equating and Item Banking. Paper presented at the Third International SweMaS conference, Umea, October 14-15, 2003.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25–39.

Steven. P. R (2009) The Emergence of Item Response Theory Models and the Patient Reported Outcomes Measurement Information Systems. *Austrian journal of statistics*, *38*(4), 211–220.

Templin, J. (2012, July 9-13). Item Response Theory. Colorado, Georgia, USA

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). Handbook of modern item response theory. New York: Springer.

---

[i] **Isaac B. Ogunsakin** is of the Department of Educational Foundations and Counseling, Faculty of Education, Obafemi Awolowo University, Ile-Ife. Author can be reached via email at sakinbamikole@yahoo.com.

[ii] **Y. O. Shogbesan** is of the Department of Educational Foundations and Counseling, Faculty of Education, Obafemi Awolowo University, Ile-Ife Author can be reached via email at yusufolayinka813@yahoo.com.