ITEM RESPONSE THEORY: A TOOL FOR EDUCATION MEASUREMENT AND EVALUATION

Udoudoh, Juliana Francis and Umoobong, MfonObong PhD

Department of Educational Foundations, Guidance and Counselling,
University of Uyo, Uyo, Akwa Ibom State

ABSTRACT

Test theory comes in two principal ways: Classical Test Theory and Item Response Theory. While CTT makes no assumption about matters that are beyond the control of psychometrician, it test scoring procedures have the advantage of being simple to compute. Though CTT was initially confined to psychological tests, it has other shortcomings. Thus, the introduction of IRT, which is the branch of science, that comprises explanatory statements, acceptable principles and methods of analysis. IRT subsequently became the most important psychometric method of validating scales because it provides a method for resolving many of the measurement challenges that need to be addressed when constructing a test or scale, and widely used in the development and assessment in the field of education. This research identified IRT as the framework and popular development in psychometrics to overcome the shortcomings of CTT, and maximize objectivity in educational measurement and evaluation.

INTRODUCTION

In Educational Measurement and Evaluation, two basic test theories concerned with ensuring valid measurement of examinees' ability are used. These are; Classical Test Theory and Item Response Theory. A test is an assessment intended to measure a test taker's knowledge, skill, aptitude, physical fitness classification in many disciplines. A test may be administered orally, on paper, on a computer or in a confined area that requires a test taker to physically perform a set of skills. A test score is a number which purportedly reflects a candidate's proficiency in some clearly defined knowledge or skill domain. Test theory is essentially the collection of mathematical concepts that formalize and clarify certain questions about constructing and using tests, and then provide methods for answering them. A basic understanding of test theory is an important prerequisite before one may create or use a scale to measure behavior. Test theory is a principle that guides action or assists comprehension or judgement. It is the branch of science that comprises explanatory statements, acceptable principles and methods of analysis. Test theories as models are important to the practice of educational and psychological measurement because they provide a framework for conceding issues and addressing technical problems.

CLASSICAL TEST THEORY

The Classical Test Theory was originally the leading framework for analyzing and developing standardized tests. It has dominated the area of standardized testing based on the assumption that a test taker has an observed score and a true score. Classical Test Theory (CTT) started off as majority of practices developed during the 1920's. The conceptual foundation, assumptions and extension of the basic premises of Classical Test Theory (CTT) has allowed for the development of some excellent psychometrically sound scales. This theory has component theories like Theory of Validity, Theory of Reliability, Theory of Objectivity, Theory of Test Analysis, Theory of Item Analysis, among others. Most of the practices were initially confined to psychological tests and later on extended to educational testing. The analyses of Classical Test Theory(CTT) are the easiest and most widely used form of analyses. The statistics can be computed by readily available statistical packages.

The classical test theory explains what is mostly done today in educational measurement. It needs to be noted that, in Classical Test Theory, measurement of a person's ability and determination of item difficult are relative to the characteristics of the item and of the group of examinees used, respectively (Ojerinde, Popoola, Ojo and Onyeneho, 2014). This makes the resulting data largely wrongly interpreted. Moreover, Classical Test Theory(CTT) makes no assumption about matters that are beyond the control of psychometrician. It is difficult to determine what a particular examinee might do when confronted with a test item. It has also been noted that CTT has many shortcomings. The most prominent one being that each measurement depends for its meaning on its own family of test takers, and the ability of a test taker depends on the particular collection of items that measures this ability. To address the shortcomings of Classical Test Theory (CTT), more appropriate theories of psychological measurement have been proposed and investigated by experts and researchers in psychometrics. Items response theory is the most prominent of such modern measurement theories (Warm, 1978; Lord, 1980; Umobong, 2004; Nenty, 2004).

EVOLUTION OF ITEM RESPONSE THEORY

ItemResponse Theory originated in the 19th century in Mathematics and Psychology. Item Response Theory (ITR) is a family of statistical procedures for analyzing and describing test performance (Ojerinde, Popoola, Ojo and Onyeneho, 2014). The concept of the item response function came to fore before 1950 when D. N. Lawley published a paper in 1943 that showed that many of the constructs of classical test theory could be expressed in terms of parameters of the items characteristic curve. This paper marked the beginning of item response theory as a measurement theory. The other pioneering researches included Federic M. Lord, an Educational Testing Service psychometrician, who in 1969 systematically and practically brought about an acceptable definition, and also expanded, explored the avenues and developed computer software to enhance easy application of the theory. His work became the driving force behind the development of the theory and its application in the last sixty (60) years. In 1980, Lord came up with a book titled Application of Item Response Theory to Practical Testing Problems. Other scholars were the Danish mathematician Georg Rasch and an Austrian sociologist Paul Lazarfeld who pursued parallel researches independently; Benjamin Wright and David Andrich also played prominent role in bringing to fore the importance of Item Response Theory in educational measurement. These authors had greatly contributed to the development of IRT to the level it is presently.

APPLICATION OF ITEM RESPONSE THEORY

Item response theory (IRT), which is also referred to as modern test theory, latent trait theory or item characteristic curve theory, actually gained attention in the 1970s when practitioners were made to know its wide applicability, usefulness and advantages; and the invention of personal computers gave many researches access to the computing power necessary for IRT. It was used in the development of standardized tests, such as Scholastic Aptitude Tests (SATs). It subsequently became the most important psychometric method of validating scales because it provides a method for resolving many of the measurement challenges that need to be addressed when constructing a test or scale.

IRT has been widely used in the development and assessment in the field of education and health. Yang and Kao (2010) applied the theory in evaluating health related scales and measures. In it, they adapted the IRT nomenclature from the field of education to measure mental health using a clinical example of a 65 year old woman who told her primary care clinician that she has been feeling depressed in recent time. During an interview session, the clinician asked her questions from the Centre for Epidemiologic Studies – Depression (CES-D) scale to assess her level of depression. These researchers employed IRT to assess her responses to determine whether or not the CES-D questions are valid.

IRT has been identify as the most significant and popular development in psychometrics to overcome the shortcomings of CTT, and maximize objectivity in measurement (Joshua, 2012). It explains what happens when an individual encounters a test item. When an examinee walks into an examination/testing room, he brings with him his ability or trait, and confronts, or is confronted by items of definite difficulty. The purpose of the test, therefore, is to increase the relative position of the examinees' ability and produces a measurement of ability often referred to as raw score. The probability of a correct response to test items therefore depends on the person's trait and the item's parameters.

The purpose of IRT is to provide a framework for evaluating how well assessment work and the individual items on assessment work. The most common application of IRT is in education, where psychometricians use it for developing and designing examinations, maintaining banks of items, and equating the difficulties of items for successive versions of examinations; for instance, it checks the differences between students' results over a period of time (Lord,1980).

BASIC ASSUMPTIONS IN IRT

IRT has several assumptions.

One of such assumptions is monotonicity, which is best displayed on a graph as a curve shaped like an **S** between the latent trait level on the X-axis and the probability of a more extreme response on the item on the Y-axis.

Another assumption is invariance in the item parameters and latent trait across different sample characteristics. In this case, the estimation of the item parameters and the latent trait are assumed to be independent of the sample characteristics within a given population.

There is an assumption of the local independence of items. This refers to the fact the chance of one item being used is not related to any other item(s) being used; and the response to an item is each and every test – taker's independent decision, that is, there is no cheating or pair or group work. It is assumed that the respondent's responses to questions are not statistically related to each other, even after the latent trait is taken into consideration or statistically held constant.

An important assumption that complements the local independence assumption is the unidimensional trait where only one latent trait is measured by the set of items in the scale or test. It is measured on a scale typically set to a standard scale with a mean of 0.0 and a standard deviation of 1.0. Unidimensionality should be interpreted as homogeneity, a quality that should be defined or empirically demonstrated in relation to a given purpose or use, but not a quantity that can be measured.

IRT MODELS OR MEASUREMENT PARAMETERS

The type of IRT depends on the research question, field of study, and how item parameters are estimated and held constant. The models may come as unidimentional and multi-dimensional models. Uni-dimensional models require a single trait (ability) dimension, while multi-dimensional IRT often arise from multiple traits. Because of the complex nature of multi-dimensional models, most researchers prefer to apply a unidimentional model in carrying out measurement and evaluation in various fields.

There are several measurement parameters which IRT adopts. Dichotomous IRT models are described by the number of parameters they use. It may be 1-Parameter Logistic (1-PL), 2 Parameter Logistic (2-PL) or 3 Parameter Logistic (3-PL) or 4

Parameter Logistic(4-PL) ,and 5Parameter (5-PL).But 4-PL and 5-PL are not common .The one-parameter model (1PL) which is also called the Rasch Model holds the item discrimination constant so that only the item difficulty (location) is estimated. It assumes that guessing is a part of the ability and that all items that fit the model have equivalent discriminations, so that items are only described by a single parameter. This results in one-parameter models having the property of specific objectivity, meaning that the rank of the person ability is the same for items independently of difficulty. The 1PL does not only assume that guessing is not present or irrelevant, but that all items are equivalent in terms of discrimination, analogous to a common factor analysis with identical loading for all items. Individual items or individuals might have secondary factors but these are assumed to be mutually independent and collectively orthogonal (Wikipedia, 2014).

The two-parameter model (2PL) estimated both the item discrimination and the item difficulty. It assumes that the data have no guessing, but that items can vary in terms of location. For a 2-PL model, the item information is determined by the item information function for both the item discrimination and item difficulty (location) at each value of theta. In general, a higher item information curve is determined by higher item discrimination and greater item difficulty at a specific value of theta relative to other items in the scale. According to Yang and Kao (2010), the test information curve is the summation of the item information functions at each value of theta for all items in the scale.

The three-parameter model (3PL) is named so because it employs three item parameters. It estimates the item discrimination, item difficulty, and the guessing parameter. It is pertinent to point out that since the guessing parameter is not as relevant in the mental health field as in education field, a 3-PL is not commonly used in health related questionnaires. Analytically, 1 parameter models are sample independent, a property that does not hold two-parameter and three parameter models. Additionally, there is theoretically a four – parameter model (4PL) ,with an upper asymptote.

A COMPARISON OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORIES

These are some of the differences between classical test theory and item response theory.

i. Item Response Theory provides findings and assumptions than Classical Test Theory(CTT) primarily, characterizations of error. Although Classical test theory (CTT) results allow important practical results, the model – based nature of IRT affords many advantages over analogous CTT.

ii. Classical test theory (CTT) test scoring procedures have the advantage of being simple to compute and explain, whereas IRT scoring requires relatively complex estimation procedures.

iii. Item response theory (IRT) provides several improvements in scaling items and people. The specifics depend upon the IRT model, but most models scale the difficulty of items and the ability of people on the same metric. Thus, the difficulty of an item and the ability of a person can be meaningfully compared.

iv. The parameters of IRT models are generally not sampled or test dependent whereas true score is defined in CTT in the context of a specific test. Thus, IRT provides significantly greater flexibility in situations where different samples or test forms are used. These IRT findings are fundamental for computerized adaptive testing.

v. Developing tests based on IRT is more advantageous than developing them based on IRT, particularly in the aspects of item analysis, selection of items, test validity and reliability assessment.

Unlike in CTT, where item statistics are sample dependent, test item in IRT can be calibrated without reference to the items and the quality of the sample-in-use, using the test-of-fit model. Once an item has shown to fit the model, such is selected for the

Table 1: Main Difference between Classical Test and Item Response Theory

S/N		Classical Test Theory (CTT)	Item Response Theory (IRT)
1	Model	Linear	Non-linear
2	Level	Test	Item
3	Assumption	Weak (ie easy to meet with test data)	Strong (more difficult to meet test data)
4	Item-Ability Relationship	Not specified	Item characteristic functions
5	Ability	Test scores or estimated test. Scores are reported on the test score scale or a transformed test-score scale	Ability scores are reported on the scale –a to +a (or a transformed scale).
6	Invariance of item and person statistics	No item and person sample dependent	Item and person parameters are sample independent if model fits the test data.
7	Item statistics	*p, r*	b, a and c (for the 3 parameter model) plus corresponding item functions.
8	Sample size (for Item Parameter estimation)	200 to 500 (in general)	Depends on the IRT model but model but larger samples, ie over 500 in general are needed.

Source: Ojerinde, Popoola, Ojo and Onyeneho (2012).

ADVANTAGES OF ITEM RESPONSE THEORY OVER CLASSICAL TEST THEORY

I. IRT provides several advantages over Classical Test Theory (CTT) methods for constructing tests and examining measurement equivalence. Unlike CTT which depends fundamentally on the subset of items and persons examined, IRT item and person parameters are invariant. This makes it possible to examine the contribution of items from a test.

ii. IRT allows researchers to calculate conditional standard errors of measurement based on test information function, rather than assuming average standard errors across all trait levels as in CTT. This allows researchers to select items that provide maximum measurement precision in a particular ability/trait range.

iii. IRT allows researchers to conduct rigorous tests of measurement equivalence across experimental groups. This is particularly important in cross-cultural research where groups are expected to show mean differences on the attribute being measured.

iv. IRT method can distinguish item bias from true differences on the attribute measured, whereas CTT method cannot.

v. IRT can be selected that provide the most information for each examinee. This can dramatically reduce the time and cost associated with test administration.

vi. IRT has extended the concept of reliability which refers to the degree to which measurement is free of error. Ordinarily, reliability is measured using a single index defined in various ways, such as the ratio of true and observed score variance. This index is helpful in characterizing a test's average reliability, for example in order to compare two tests. On the whole, IRT advances the concept of item and test information to replace reliability.

CONCLUSION AND RECOMMENDATIONS

The concept of Item Response Theory has been in use since 1943 when D. N. Lawley published a paper in 1943 that showed that many of the constructs of classical test theory could be expressed in terms of parameters of the items characteristic curve. This marked the beginning of item response theory as a measurement theory. The purpose of Item Response Theory is to provide a framework for evaluating how well assessment, and how well individual items on assessments work. The research has shown that the most common application of item response theory (IRT) is in education where psychometricians use it for developing and designing examinations, for instance, to allow comparisons between results over time.

Item response theory (IRT) is generally claimed as an improvement over Classical Test Theory (CTT). For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information.

REFERENCES

Akpan, S. M. (2002). *Test and Measurement: Concepts and Practice.* Executive Publishers, Owerri.

Joshua, M. T. (2012). *Fundamentals of Tests and Measurement in Education.* University of Calabar Press, Calabar

Lord,F. M.(1980). Applications of item response theory to practical testing problems. Mahwah,NJl: Lawrence Erlbaum Associates, inc

Ndem, U. D.; Udoh, A. O.; and Joseph, E. U. (2003). *Tests and Measurement fpr Teachers and Students.* Derand Publishers, Uyo, Nigeria.

Ojerinde, D.; Popoola, K; Ojo, F. and Onyeneho, P. (2014). *Introduction to Item Response Theory.* Marvelouse Mike Press Ltd, Abuja.

Ojerinde, D.; Popoola, K; Ojo, F. and Onyeneho, P. (2014). *Parctical Applications of Item Response Theory in Large Scale Assessmsnet*. Marvelouse Mike Press Ltd, Abuja.

Suen, H. K. (1990). *Principles of Test Theory.* Lawrence Erlbaum Associates Publishers,

Yang, F. M. and Kao, S. T. (2010). *Item Response Theory for Measurement Validity.* Printed Online.