IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388,p-ISSN: 2320–737X Volume 3, Issue 5 (Nov. –Dec. 2013), PP 77-81
www.iosrjournals.org

A Comparative Study of Classical Theory (Ct) and Item Response Theory (Irt) In Relation To Various Approaches of Evaluating the Validity and Reliability of Research Tools

Mamun Ali Naji Qasem
Research Scholar, Department of Education Aligarh Muslim University, U.P and Member of Faculty of Education, University of Ibb, Yemen

Abstract: *Measurement theories are important to practice in educational measurement because they provide a background for addressing measurement problems. One of the most important problems is dealing with the Measurement Errors. A good theory can help in understanding the role of errors they play in measurement; (a) To evaluate the examinee's ability to minimize errors and (b) Correlations between variables.*

There are two theories addressing measurement problems such as test construction, and identification of biased test items: Classical Test Theory (CT) and Item Response Theory (IRT) (1950). As a result of a number of problems associated with the Classical Theory of Measurement, which cause inaccuracy in results i.e. methods and tools of measurement. There appeared a need to develop the methods of measuring behavior in a manner consistent with the Physical Measurement Methods. Based on the Philosophy of this measurement and assumption, which achieves the quality and safety of these methods, and acceptance of their results with a high Degree of Confidence. There were many research studies by professionals and those interested in behavioral measures, aimed and try to overcome some of the Behavioral Problems of Measurement. These studies have resulted in the emergence of Item Response Theory.

Item response theory is a Statistical Theory about Items, Test Performance and abilities that are measured by Items. Item responses can be discrete or continuous and can be dichotomous and the item score categories can be ranked or non ranked . There can be one ability underlying test, and there are many models in which the relationship between item responses and the underlying ability can be specified. Within the IRT there are many models that have been applied to test data really but most famous among them is Racsh model.

In this paper, both the theories i.e. Classical Test Theory and Item Response Theory (lRT) will be described in relation to approaches to measure the validity and reliability. The intent of this module is to provide a comparison of classical theory and item response theory.

Keywords *: Classical Test Theory (CT), Item Response Theory (lRT), Validity and Reliablity.*

I. Research Objectives

1- Study the concept of Classical Theory (CT) of Measurement and Item Response Theory (IRT). .
2- Knowing the Problems of Classical Theory (CT) of Measurement.
3- Knowing the assumptions of Item Response Theory (IRT).
4- Studying the methods of Validity and Reliability in Classical Theory (CT).
5- Studying the methods of Validity and Reliability in Item Response Theory (IRT).
6- To find out difference in calculating reliability and Validity through (CT) and IRT.

II. Importance of the Study

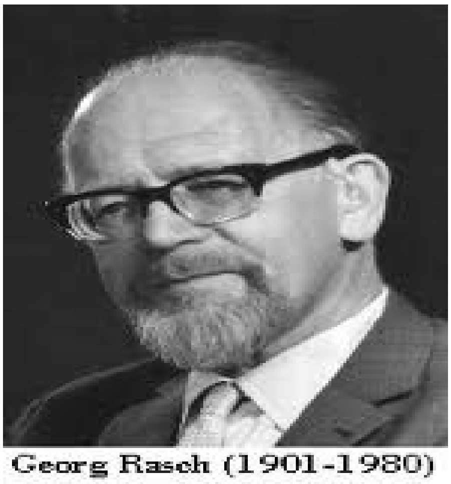
The current research is of a great importance because it deals with a theory advanced in Educational Measurement i.e. Item Response Theory, (IRT) which has become widely used and it has become popular among researchers in Educational and Psychological Measurement.

Index Terms: Classic Theory in Measurement , Item Response Theory**,** Validity and Reliability

III. Methodology of Research.

The Researcher has used a method called Content Analysis research, because it has seemed to be the most adequate method to fulfil the Aims of this Research.

Ronald Fisher -Classical theory- (CT)



Georg Rasch (IRT)

Classical Test Theory (CT)

Classical test theory introduces three concepts-test score, true score, and error score. Within that theoretical framework, models of various forms have been formulated. For example, we often referred the "classical test model," a simple linear model where the postulates linking the observable test score (X) to the sum of two unobservable (or often called latent) variables, true score (T) and error score (E), that is, X = T + E. (Hambleton and Jones, 1993)

Item Response Theory (IRT)

We call the new theory as item response theory because it focuses on the item, as opposed to the classical theory. IRT models the response of every examinee to every item in the test. The term item covering all types of items. Multiple choice questions that have incorrect and the correct answer .

Item Response Theory is a statistical theory about the item and test performance and the abilities that are measured by the items. Item responses can be discrete or continuous and can be dichotomous and the item score categories can be ranked or non ranked . There can be one ability underlying test, and there are many models in which the relationship between item responses and the underlying ability can be specified. Within IRT there are many models that have been applied to real test data but the most famous is **Racsh Model**.

IV. Problems of Classical Theory of Measurement

Despite Classic Theory is mostly used by researchers in educational research , but it is not devoid of shortcomings in the analysis of the results of the tests, The main shortcomings in this theory are as follows: -

1- That all Psychometric Properties that are built based on Classical Theory such as difficulty, discrimination, and stability depend on the characteristics of a sample of individuals to which the test is applied, and the level of difficulties in the items included in the test

2- We assumed that the scores of the individual test items will be on Linear Scale for all individuals. In other words the difference between the two scores is fixed. But in fact this Scale in Classic Theory Usually in the form of the Curve.

3- Assumes that the test scores that represent the feature or ability must be in a linear function steadily, if the scores of the individual increase in the test the amount of his ability must be in increase also . However, some individuals with high ability sometimes they get low scores on the tests, and maybe the opposite will happen for those with low ability .

4- The test construction is changing over time, that means that the construction and meaning of items test change from time to time. For the samples the environmental conditions are changing and test conditions are not standardized and delete or change any item of the test, may lead to a change in scores of individuals, and seriously affect the representation of items domain.

5- The results of individuals on the test depend on the characteristics of the sample items that include in the test, if we pulled samples of items differ in difficulty from a large group of items to measure the same ability, the expected degrees to individuals in the test will vary according to the difficulty of items.

V. The assumptions of Item Response Theory (IRT).

The mathematical models in item response theory determine the relationship between person performance on the test and the feature behind this performance, and this mathematical model is the equations which connect the person's ability and possibility to get the correct answer.

This theory is based on three basic assumptions:

1) **Uni-dimensionality**: means items of test measure only one ability (trait) .(Warm, 1978)
2) **Local Independence**: means answer of an item does not affect positively or negatively on the other item. (Crocker and Algina , 1986)
3) **Item Characteristic Curve**: represents the relationship between the probability of the correct answer to the item and the feature, where this relationship show through a mathematical function called the Item Characteristic Curve that linking the probability of the correct answer for the item and the ability .(Hulin ,and others, 1983)

VI. Methods of Validity in Classical Theory in Measurement

The validity means the extent to which an instrument measures what it purports to measure. **In Classical Theory of Measurement,** there are three methods to evaluate validity of research tool:

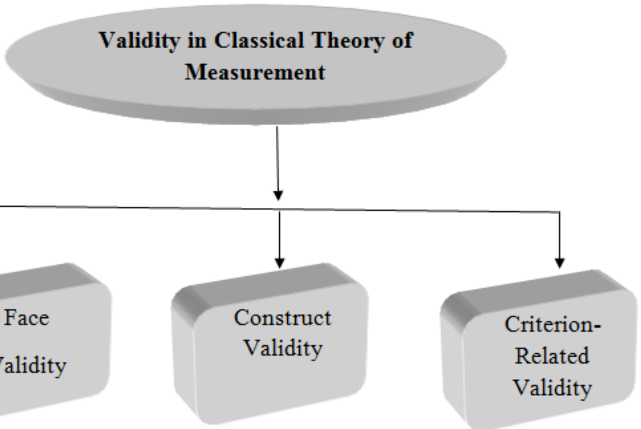

Validity in Classical Theory of Measurement
Face Validity
Construct Validity
Criterion-Related Validity

Figure 1 Methods to Evaluate Validity of Research Tool in Classical Theory

9.1 Face Validity (Content Validity).

Face validity refers to researchers’ subjective assessments of the presentation and relevance of the measuring instrument as to whether the items in the instrument appear to be relevant, reasonable, unambiguous and clear. Several authors have commented on the status of face validity in research. Most of these authors believe that face validity is not truly an indicator of validity and hence should not be considered as one. Practically, the quantitative assessment of face validity can be achieved by having experts in the field of study. (Oluwatayo, 2012)

9.2 Construct Validity.

Construct Validity. This type of validity is a judgment based on the accumulation of evidence from numerous studies using a specific measuring instrument. Evaluation of construct validity requires examining the relationship of the measure being evaluated with variables known to be related or theoretically related to the construct measured by the instrument. For example, a measure of quality of life would be expected to result in lower scores for chronically ill patients than for healthy college students. Correlations that fit the expected pattern contribute evidence of construct validity.

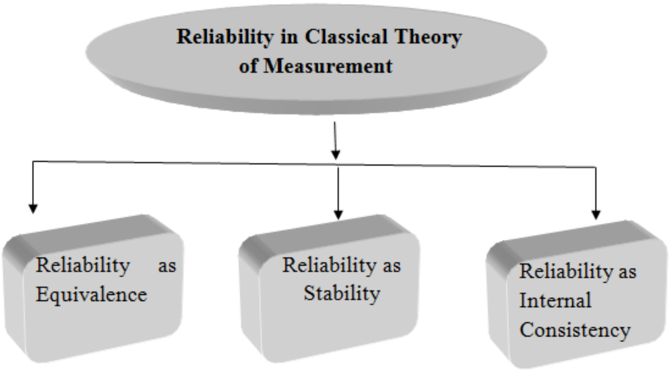
9.3 Criterion-Related Validity.

Criterion-Related Validity. This type of validity provides evidence about how well scores on the new measure correlate with other measures of the same construct or very similar underlying constructs that theoretically should be related. It is crucial that these criterion measures are valid in themselves. With one type of criterion-related is predictive validity and another type of Criterion-Related Validity is Concurrent Validity. (Carole and Winterstein, 2008)

VII. Methods of Reliability in Classical Theory of Measurement

Reliability. According to Classical Theory, any score (the observed score) is consisted of both the "True Score" which is unknown, and "Error Score" in the measurement process. There are different means of estimating the reliability of any measure. . (Carole and Winterstein, 2008)

Figure 2 Methods to Evaluate Reliability of Research Tool in Classical Theory


Reliability in Classical Theory of Measurement
Reliability as Equivalence
Reliability as Stability
Reliability as Internal Consistency

Reliability as Equivalence

Reliability as equivalence is of two sorts: alternate or parallel form and inter-rater form. Estimating reliability using alternate or parallel form requires developing two forms of an instrument using the same content domain, the same test specifications, the same number of items, the same item format and similar difficulty and discriminating indices. (Oluwatayo, 2012)

Reliability as Stability

Test-retest reliability is used to assess the consistency of a test across time. It is measured by the correlation between results from tests administered to the same group of people over two or more periods.

Reliability as Internal Consistency.

"Internal consistency gives an estimate of the equivalence of sets of items from the same test (e.g., a set of questions aimed at assessing quality of life or disease severity). The coefficient of internal consistency provides an estimate of the reliability of measurement and is based on the assumption items measuring the same construct should correlate. The most widely used method for estimating internal consistency reliability is **Cronbach's Alpha**. And there are others methods such as **Split Half and Kuder-Richardson**-20 & 21 (KR-20&21)." . (Carole and Winterstein, 2008)

VIII. The Methods of Validity and of Reliability in Item Response Theory (IRT).

In Item Response Theory (IRT) the meaning of validity and reliability differ in classic theory (CT) because the (IRT) theory focuses on the characters of the item.

Validity in Item Response Theory means to what extent individuals and items have a good ranking in the ability which the test measure, in other words the ability of any test to rank (order) the individuals according to their ability as well as rank the items according to their level of difficulty. (Hambleton, 1983)

The reliability in Item Response Theory (IRT) means to what exetent the measure is independent (free) from groups (samples) as well as from the test items, in other words the characteristics of items don't effected by the group which we apply to the test and if we apply many versions of test for the same group they must get the same score and same ranking.(Lord, 1968)

There are three models to evaluate the validity and reliability of items test according three parameters effect on the Psychometric characteristics for any test. The ability of the examinee, Level of difficulty of the item and Item ability to discriminate.

"A reasonable assumption is that each examinee responding to a test item possesses some amount of the underlying ability. Thus, one can consider each examinee to have a numerical value, a score, that places him or her somewhere on the ability scale. This ability score will be denoted by **the Greek letter theta, è** . At each ability level, there will be a certain FIGURE 3 .A typical item characteristic curve probability that an examinee with that ability will give a correct answer to the item. In the case of a typical test item, this probability will be small for examinees of low ability and large for examinees of high ability. If one plotted P (θ) as a function of ability, the result would be a smooth S-shaped curve such as shown in Figure 1. The probability of a correct response is near zero at the lowest levels of ability. It increases until at the highest levels of ability, the probability of correct response approaches 1. This S-shaped curve describes the relationship between the probability of the correct response to an item and the ability scale. In item response theory, it is known as the item characteristic curve. Each item on a test will have its own item characteristic curve." (Bacer, 2001)

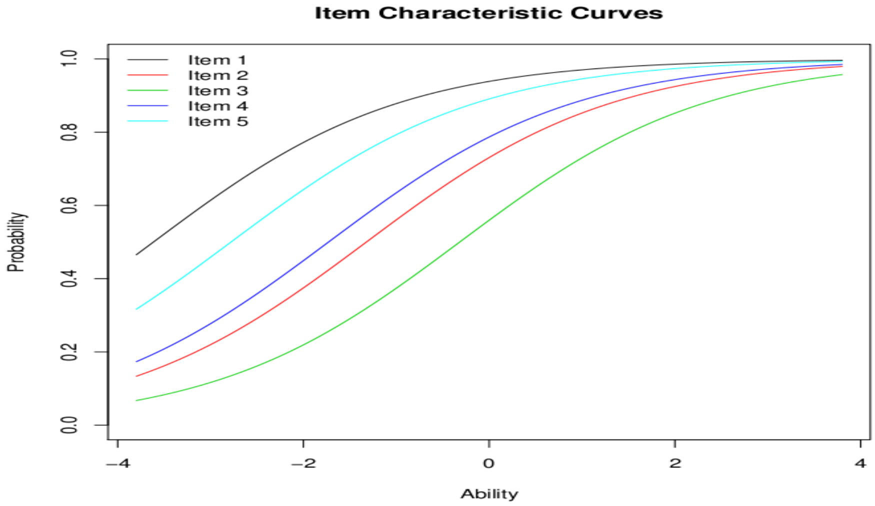


Figure 3 Item-Characteristic-Curve in (IRT)

IX. Conclusion

- Almost all the researchers depend on classic theory in their researches, because its concept is more clear to them. To evaluate the Validity and Reliability according to CT methods is easy.
- There are three methods widely used to evaluate validity of the research tool according to CT i.e. the Face Validity, Construct Validity and Criterion Validity .
- There are three methods widely used to evaluate reliability of the research tool according to CT the Face, Construct and Criterion Validity .
- As a result of a number of problems associated with the Classical Theory of Measurement the Item Response Theory (IRT) has been developed.
- The Item Response Theory (IRT) approved solutions for Classic Theory problems in Measurement.
- Although Item Response Theory (IRT) is more accurate in evaluating Validity and Reliability of Measurement Tools , but it is less used in Educational Research.

References:

[1]. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.
[2]. Carole, K. & Winterstein, A, (2008). Validity and reliability of measurement instruments used in research. Am J Health-Syst Pharm. Vol 65 Dec 1, 2008
[3]. Crocker , L. And Algina , J. (1986) Introduction to classical and modern test theory .New York: Holt , Pinehart and Winston .
[4]. Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, & Winston. Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.
[5]. Hambleton, R. K (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd Ed., pp. 147-200). New York: Macmillan.
[6]. Hambleton, R. K, & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.
[7]. Hambleton, R. K, & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.
[8]. Hambleton, R. K, Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
[9]. Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. Educational Measurement: Issues and Practice, 8, 35-41.
[10]. Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. 1. Thorndike (Ed.), Educational measurement (2nd Ed., pp. 130-159). Washington, DC: American Council on Education.
[11]. Hulin, L. , Drasgow, F. & Parsons , K. (1983) Item response theory application to psychological measurement, Ilinois: Dow Jones .
[12]. Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.
[13]. Loyd, B. H. (1988). Implications of item response theory for the measurement practitioner. Applied Measurement in Education, 1 (2), 135-143.
[14]. Oluwatayo, J. A. (2012). Validity and Reliability Issues in Educational Research, Journal of Educational and Social Research. Vol. 2 (2)
[15]. Warm , A. (1978) . A primer of Item Response Theory: U.S. Coast Guard Institute Oklahoma, 73/69 .