International Journal for Social Studies
Available at https://edupediapublications.org/journals
ISSN: 2455-3220
Volume 02 Issue 09
September 2016

CLASSICAL TEST THEORY: An Introduction to Linear Modeling Approach to Test and Item Analysis

Ado Abdu Bichi
Department of Arts and Social Sciences Education,
Northwest University, Kano-Nigeria
adospecial@gmail.com +2348032928301

Abstract:
The practice of testing has become increasingly common and the reliance on information gained from test scores to make decision has made an indelible mark on our culture. The entire educational system is today highly concerned with the design and development of the tests, the procedures of testing, instruments for measuring data, and the methodology to understand and evaluate the results. In theory of measurement in education and psychology, Classical Test Theory (CTT) is a popular framework. The techniques of CTT are applied in assessment situations to improve test analysis and test refinement procedures. The main purpose of this paper is to provide a comprehensive overview of the CTT and its procedures as applied to test item development and analysis. The usage of CTT in measurement is to determine maximum information about an individual. It is a scientific framework which has a pioneer role in educational measurement and psychometric process. CTT has served the measurement community for decades, besides depicting the simplicity of the CTT model from multiple points of view; various limitations of the model were highlighted. These limitations are detailed in item, person and ability level. Despite the shortcomings attributed to CTT it is recommended that, Classical test theory approach of item analysis should be maintained in test development and evaluation, because of its superiority and simplicity in the investigation of reliability and in minimizing measurement errors.

Keywords
Classical Test Theory, Test Development, Educational Measurement

1. Introduction

Assessment of students learning is very important in education. The assessment of students' cognitive abilities, academic skills and intellectual development involves certain techniques employed to sample students' performance on a particular learning outcome targeted by the instructional objectives one of that techniques is test, the test is expected to sample students' behaviours. Thus creating quality tests is very important in assessing the students' performance; many indices have been developed in order to construct valid and reliable items during test development. These indices developed mostly rely on the two popular statistical frameworks Classical Test Theory and Item Response Theory. The two frameworks are associated with the item development process in the field of educational and psychological test. These frameworks are widely been used in test development to ensure quality of measuring instruments and discuss in various literatures in the field of psychological and educational measurements on their suitability and effectiveness in test development process. In the theories the models associated with each have been described and compared, and the ways in which test development generally proceeds within each frameworks have demonstrated [1] the existence of the theoretical as well as empirical differences and similarities of the two frameworks were extensively described in many studies. This paper provides a critical review of the existing empirical studies conducted to describe and compare the two popular frameworks.

2. Classical Test Theory: Overview

The Random sampling theory and item response theory are two major psychometric approaches used in a measurement. Classical test theory approach and the generalisability theory are the two approaches in random sampling theory [2]. [3] maintained that, Classical test theory is a simple model that describes how measurement errors can influence observed scores. According to [4] Classical Test Theory (CTT) is an emancipation of the early 20th century approaches to measuring individual differences. CTT was born after the following three achievements or ideas were conceptualized: 1. recognition of the presence of errors in measurements, 2. a conception of that error as a random variable, and 3. a conception of correlation and how to index it. [5] stated that in 1904, Charles Spearman figured out how to correct a correlation coefficient for attenuation due to error measurement and how to obtain the reliability index needed in making the correction. His finding is considered to be the beginning of Classical Test Theory. Some other scholars who played a significant role in the Classical Test Theory's approach include: Truman Lee Kelley, George Udny Yule, Louis Guttman,

those involved in making Kuder-Richardson Formulas [6].

2.1. What is Classical Test Theory?

Classical test theory has been used for decades to determine reliability and other characteristics of measurement instruments. According to [1] Classical test theory is a theory about test scores that introduces three concepts (1) test score (often called the observed score), (2) true score, and (3) error score. Within this framework, various models have been formulated. Example, in what is often referred to as the "classical test model,"

$$X = T + E \qquad (1)$$

This is a simple linear model that links the observable test score(X) to the sum of two unobservable variables, true score (T) and error score (E). Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items. Therefore the equation is not solvable unless some simplifying assumptions are made.

The major assumptions underlines the CTT are: true scores and error scores are uncorrelated, the average error score of the examinees is zero, and error scores on the parallel tests are uncorrelated. According to [7] the assumption of classical test theory is that, each individual examinee has a true score (unobservable) which would be obtained if there were no errors in measurement. However, because the instruments used are imperfect, the score observed for each individual may differ from an individual's true ability. This difference between the observed score and the true score results from measurement error. Error is often assumed to be a random variable having a normal distribution. The Classical test theory's implication for examinees is that tests are fallible imprecise tools. The score obtained by an individual is called the individual's true score. This means that even with the repeated application of the same test, the true score for an individual will not change. This CTT's observed score is always the true score influenced by some degree of error, the influence of this error on the observe score can be positive or negative.

[8] stated that, theoretically, the standard deviation of the distribution of random errors for each examinee tells about the magnitude of measurement error. Usually, it is assumed that the distribution of random errors will be the same for all test takers. The standard deviation of errors is uses as the basic measure of error in Classical test theory. In practice, the reliability of the test and standard deviation of the observed score are used to estimate the standard error of measurement. The smaller the standard error of measurement the more certain is the accuracy with which the attribute measured which also tell us the individual score is close to the true score. Conversely, the larger the standard error of measurement, the less certain is the accuracy with which an attribute is measured.

The standard error of measurement is represented and calculated with the formula:

$$SE_M = S_X\sqrt{1} - R_{xx} \qquad (2)$$

Where: SE_M=standard error of measurement
S_x = standard deviation of test scores
R_{xx}= reliability coefficient
Small SE_M indicates high reliability

The Standard errors of measurement are used to create confidence intervals around specific observed scores [8]. The upper and lower bound of the confidence interval approximate the value of the true score.

The observed score in CTT is assumed to be measured with error. However, in developing measures, the goal of CTT is to minimize this error [9]. The Importance of a test's reliability and calculating the reliability coefficient increases, in that case. If reliability coefficient is known, error variance can be estimated. The square root of error variance is determined as a standard error of measurement and helps to define the confidence interval in order to have a more realistic estimation of the true score [10].

2.2. CTT Statistics and Item Analysis

Classical test analysis utilizes traditional item and sample dependent statistics. These include item difficulty and item discrimination estimates, distractor analyses, item-test inter correlations, and a variety of related statistics. Most of the psychometric analyses have focused on examinee assessment at the test score level, rather than at the item level. Classical test analysis also typically includes a measure for the reliability of scores (i.e., Cronbach's Alpha), difficulty of the test item and Discrimination. Item Analysis is a set of statistical procedures that focus on the selection of items that maximizes score reliability. The major classical analysis statistics are. 1. Difficulty (item level statistic); 2. Discrimination (item level statistic) and 3. Reliability (test level statistic).

i **Item Difficulty**

Item difficulty in classical theory is the first item characteristic to be determined. Item difficulty is simply the proportion of examinees taking the test,

ISSN: 2455-3220
Volume 02 Issue 09
September 2016

who got an item or answer it correctly. The larger the percentage getting an item correctly, the easier the item is. The higher the difficulty value, the easier the item is understood to be. To compute the item difficulty index, divide the number of examinees answering the item correctly by the total number of examinees answering item. An item answered correctly by 75% of the examinees would have a difficulty index or p-value, of .75, whereas an item answered correctly by 40% of the examinees would have a lower item difficulty or p-value, of .40 [11]. The item difficulty is denoted as p and is symbolically given as:

$$P = \frac{R}{N} \qquad (3)$$

Where P = is the difficulty of a certain item
R = is the number of examinees who get that item correct and
N = is the total number of examinees.

A general guideline for the interpretation of an item difficulty index is provided in the following table; see, for example, [12]; [13] [14] among others

Table 1: Item difficulty indices interpretation [14]

Difficulty Index (p)	Interpretation
P ≤ 0.30	Difficult
0.31 ≤ 0.70	Moderately difficult
P> 0.70	Easy

ii **Item Discrimination**

Item discrimination refers to the difference in correct responses between the low and the high scoring students. It is the ability of a test item to discriminate between higher ability and lower ability examinees [12]. For the item difficulty, a group that answered the item correctly, and one that did not is created. This statistic focuses on determining the correct respondents or examinees get the item right or wrong in a test. In essence, the aim of item discrimination is to eliminate or dropped or modified items that do not function well in the tested group [15]. The index of discrimination to determine the discriminating power of an item can be computed using two indices: the item discrimination index, D, and Item discrimination coefficient

a. Item Discrimination Index (D)

This method can be applied to compute a simple measure of the discriminating power of an item using the extreme groups [11]. In calculating the D index, first ranks order the students by their test scores. Next, separate the top 27% of the students and the 27% at the bottom for the analysis. As stated by [13] "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis" The discrimination index, D, is given as

$$D = Pu - Pi \qquad (4)$$

With Pu being the proportion of correct responses for the upper group and Pi being the proportion of correct responses for the lower group, Since its proportion ranges from -1 to +1, a negative index indicates that the larger portion of the lower group answered the item correctly while a positive index indicates that a higher proportion of the upper group got the item correctly [15].

b. Discrimination coefficients

There are two indicators of the item's discrimination effectiveness; these are; point biserial correlation and biserial correlation coefficient. The choice of correlation depends on the kind of question we want to answer. One of the major shortcomings of the discrimination index, D is that, only 54% (27% upper + 27% lower) are used to compute the item discrimination and 46% of the examinees ignored. Similarly, the advantage of using discrimination coefficients in determining the discriminating power over the discrimination index is that every examinee taking the test is used to compute the discrimination coefficients. A point-biserial correlation coefficient (rpbi) is defined by:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} \qquad (5)$$

Where: Mp = whole-test mean for students answering item correctly,
Mq = whole-test mean for students answering item incorrectly,
St = standard deviation for whole test,
p = proportion of students answering correctly
q = proportion of students answering incorrectly [13].

A Point biserial correlation (r_pbi) coefficient ranges from -1 to +1. A high point-biserial coefficient means that students with higher total scores are students selecting the correct response, and students selecting incorrect responses to an item are associated with lower total scores. According to the value of r_pbi, item can discriminate between high-ability and low-ability examinees. Very low or negative point-biserial coefficients help in identifying defective test items [15].

A summary of the widely used [16] criteria and guidelines for categorizing discrimination indices in item and test analysis is used in this study.

Table 2.2: Interpretation of Discrimination Indices [18]

Discrimination Index	Quality of an Item
D ≥ 0.40	Item is functioning quite satisfactorily
0.30 ≤ D ≤ 0.39	Good item; little or no revision is required
0.20 ≤ D ≤ 0.29	Item is marginal and need revision
D ≤ 0.19	Poor item; should be eliminated or completely revised

iii Reliability

There are different means of estimating the reliability of any measure [17]. These methods are explain with the help of the diagram below: adopted from [18]

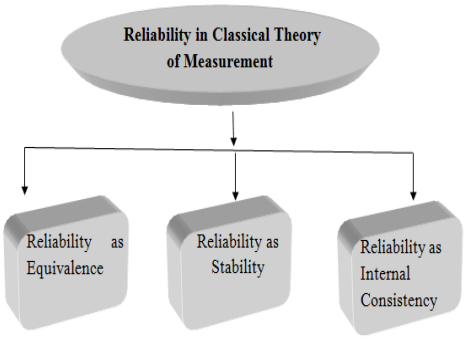

Figure 1: Classical test Reliability

a. Reliability as Equivalence

[19] pointed out that, Reliability as equivalence is of two forms: parallel or alternate form and inter-rater form. Estimating reliability using parallel or alternate form requires the developing two forms of a test or instrument using the same content domain, same number of items, same test specifications, same item format as well as a similar difficulty and discrimination indices.

b. Reliability as Stability

[18] stated that, Test-retest reliability is used to measure the consistency of a test or instruments across time. It is assessed by the correlating the results of tests administered over two or more different periods to the same group of people.

c. Reliability as Internal Consistency

According [17] Internal consistency gives an estimate of the equivalence of sets of items from the same measuring instruments (e.g., a set of questions aimed at assessing students' ability in Mathematics). The internal consistency reliability coefficient provides an estimate of the reliability of measurement, and it is based on the assumption that, items measuring the same behaviour should correlate. Cronbach's Alpha is the most widely used method for estimating internal consistency reliability. Others methods are as Split Half and Kuder-Richardson-20 and 21 (KR-20 and 21).

However In educational research using classical test approach, internal consistency estimates are the easiest to obtain which indicate the extent to which each item correlates with other items. This is measured on a scale of 0-1. The higher the coefficient the higher the item reliability, internal consistency is arrived at by using split-half, Kuder-Richardson-20 and 21 and Cronbach alpha [18].

Split-half

Split-half reliability assumes that the items in an instrument can be split into two matched halves in terms of contents and cumulative degree of difficulty. This is often achieved by assigning all the odd numbered items to one group, and all even numbered items into another. Essentially, testees' marks on one-half are expected to match his or her marks on the other half. The calculation follows by correlating the marks in the odd items with the marks in the even items using Pearson's statistics and corrected for the whole items using.

$$r_{xx} = \frac{2r_{hh}}{1+r_{hh}} \qquad (6)$$

Where: r_{hh}: correlation between the two halves of the test

Procedure:

a) Divide the test into two equal halves
b) Calculate the correlation coefficient between the two halves
c) Calculate the Spearman-Brown reliability estimate

Spearman-Brown formula will give an estimate of maximum reliability that can be expected (upper bound estimate)

Kuder-Richardson-20 and 21 (KR-20 and 21)

[21] develop procedures for determining the homogeneity of items. Probably, the best-known index of homogeneity is KR-20; This KR-20 is arrived at by considering the proportion of correct and incorrect responses to each of the items on a test. The formula for KR-20 is:

$$KR_{20} = \left[\frac{K}{K-1}\right] \times \left[\frac{S^2x - \Sigma pq}{S^2x}\right] \qquad (7)$$

Where;

K = Number of trials or items;
S^2_x= variance of scores;
p = percentage answering item right;
q = percentage answering item wrong and
Σpq = sum of pq products for all k items

However, KR-21 assumes that all items in the test are of equal difficulty and computationally simpler. The formula for KR-21 is:

ISSN: 2455-3220
Volume 02 Issue 09
September 2016

$$KR_{21} = [(K \times S^2) - (\overline{X} \times (K - \overline{X})] \div [(K-1) \times S^2] \quad (8)$$

Where; K = number of trials or items in the test;
S^2 = variance of test and
$\overline{X}$ = mean of test

Cronbach's alpha

The alpha formula is one of the best analyses that can be used to gauge the reliability (i.e., accuracy) of educational and psychological measurements. The formula was designed to be applied in a two-way table of data where rows represent persons (p), and columns represent scores (x), under two or more conditions (i). Because the analysis assesses the consistency of scores from one condition to another, procedures like alpha are known as internal consistency analyses [19]. The reliability was computed with coefficient alpha, defined as:

$$\alpha = (\frac{K}{K-1})(1 - \frac{\Sigma S_i^2}{S_x^2}) \quad (9)$$

Where: k: represent number of items on the test;
Σs_i^2: sum of the variances of the different parts of the test (item i) and
s_x^2: variance of the test scores

Cronbach's α can be shown to provide a lower bound for reliability under rather mild assumptions. Thus, the reliability of test scores in a population is always higher than the value of Cronbach's α in that population. 0.7-0.8 is an acceptable value for Cronbach's α; values substantially lower indicate an unreliable scale [23].

2.3. Item Selection in Classical Test Theory

In classical test theory item analysis consists of determining sample-specific parameters and eliminating items based on the statistical criteria or set standards. A poor item in the entire test is identified by an item difficulty index that is too low ($p<0.30$) or too high ($p> 0.70$), or a low item discrimination indices, such that $rpbi \leq 0.20$ [12]. According to [1] in test development, items are selected on the basis of these two characteristics: item difficulty and item discrimination. An item with the highest discrimination parameters is normally prioritized in item selection, however, the choice of item difficulty and discrimination is usually informed by the purpose of the test and the anticipated ability distribution of the group of people for whom the test is intended. Example, where the purpose of a test is to select a group of high-ability students for the award of a scholarship, here, the items that are quite difficult are generally chosen for the entire population of the test takers.

Example norm-referenced achievement tests are designed to differentiate between examinees with regard to their competence in a particular subject (e.g. Economics). That is; such kind of test is intended to yield a wide range of scores maximising discriminations among all students taking the test. When a test for this purpose is designed, items are generally chosen within a medium level and narrow range of difficulty.

2.4. Advantages of Classical Test Analysis

According to [4] benefits obtainable through the application of proper instructional objectives and item writing using classical test analysis include:

First, Using Classical test theory, analyses can be performed with smaller representative samples of examinees. Secondly, classical test analysis employs relative simple and straightforward mathematical procedures and model parameter estimations are conceptually easy. Thirdly, classical test analysis assumptions are easily met by traditional testing procedures. Because of this it is often referred to as "weak models".

2.5. Limitations of Classical Test Theory

While classical test methods have proven to be very useful and are still widely used among practitioners in test construction and analysis process. [1] mention that, the two classical item statistics; item difficulty and item discrimination that form the cornerstones of many classical test and item analyses are group dependent (depend on the sample). Thus, the *P* and *D* or *r*-values depend on the students' sample in which they are obtained. In terms of discrimination indices, higher values will tend to be obtained from heterogeneous samples and lower values from homogeneous samples. Similarly, in terms of item difficulty indices, higher values will be obtained from the samples examinees of above-average ability and lower values from examinee samples of low or below-average ability [24]. "Such sample dependency relationships reduce the overall utility of these statistics" [4].

Another weakness of classical test theory is that its applications are test dependent or "test-based". Test difficulty directly affects the resultant test scores. Higher knowledge scores are directly associated with tests composed of relatively easy items, and low knowledge scores can be attributed to a test composed of items that are more difficult. The true score model, upon which much of classical test theory is based, permits no consideration of examinee responses to any specific item. Thus, no basis exists to predict how a given examinee will perform on a particular test item [4].This shows that the examinee ability depends on the test item difficulty

[15] wrote that classical test reliability is an indicator of the quality of a set of test scores; hence, reliability

is dependent on characteristics of the group of examinees, in addition to being dependent on characteristics of the test and the test administration. Another limitation of classical test theory is that to compare the performance of different examinees, the examinees must be given the same or parallel items.
Another problem of classical test theory is its inability to provide basis for determining how an examinee in a given population might perform when confronted with test items [25]. Finally, according to [25], classical test theory assumes that the measurement error is the same for all test takers/examinees.
Because of the criticisms heaped upon classical test theory, some test developers have turned to item response theory.

3. Conclusion and Recommendations

Multiple factors such as the psychological state of examinee, environmental factors or test itself affect examinees' scores in each implementation of instrument. Sometimes, each test administration gives different results about the same person. The only valid and reliable constructions of examinations are for interpreting the real aspect of the ability of individual.

As it has been mentioned before, the main purpose of the psychometric process and usage of different measurement approaches or theories is to determine maximum information about an individual. This valuable information is accessible by different methods, if valid, theoretic mathematical background of implementation is used and a reliable atmosphere is satisfied. CTT is a scientific framework which has a pioneer role in educational measurement and psychometric process. Essential rules of this theory are discussed and presented in this study. CTT has served the measurement community for decades; due to its weaknesses IRT has witnessed an exponential growth in recent decades [26]. Therefore, this study presented the main principles of CTT and their effects on the educational measurement process. Besides depicting the simplicity of the CTT model from multiple points of view, various limitations of the model were highlighted. These limitations are detailed in item, person and ability level.

Despite the shortcomings attributed to CTT it is recommended that, Classical test theory approach of item analysis should be maintained in test development and evaluation, because of its superiority and simplicity in the investigation of reliability and in minimizing measurement errors. Secondly, achievement tests used to in examining students' achievement compared to educational standards should be made to pass through all the processes of standardization and validation.

4. Acknowledgement

The author appreciate the effort of the Kano State Government under the visionary governor *Engr. Dr. Rabi'u Musa Kwankwaso, FNSE,* whose fashion and concern for the welfare and educational development of his people introduced the postgraduate scholarship scheme which has offered me the opportunity to achieve what I am celebrating today.

The author dedicated the work to his lovely wife *Maryam Musa*, his children *Khadija* and *Fatima* who despite my absence with long distance remain courageous and always encourage me with prayers, love and goodwill.

5. References

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

[2] Bejar, I. I. (1993). A Generative Approach to Psychological and Educational Measurement. *In N. Frederiksen, R.J. Mislevy,& I.I. Bejar (Eds.), Test Theory for a New Generation of Tests* (pp. 323-357). Hillsdale, NJ: Erlbaum.

[3] Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson and S. L. Hershberger (Eds.), *the new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum, pp. 129-152

[4] Schumacker, R. E. (2010). Classical Test Analysis. *http://appliedmeasurementassociates.com/ama/assets/File/CLASSICAL_TEST_ANALYSIS.pdf*. Retrieved on 13 August, 2014.

[5] Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole Publishing Company.

[6] Traub, R.E., & Fisher, C.W. (1997).On The Equivalence of Constructed Responses and Multiple-Choice Tests. *Applied Psychological Measurement, 1, 355-370.*

[7] Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory using Derived Test Data. *The International Journal of Educational and Psychological Assessment,* Vol.1, Issue 1. Pp. 1-11

[8] Kaplan, R. M. & Saccuzo, D. P. (1997). *Psychological Testing: Principles, Applications and Issues*. Pacific Grove: Brooks Cole Pub. Company

[9] McBride, N. L. (2001). *An Item Response Theory Analysis of the Scales From The International Personality Item Pool and the Neo Personality Inventory-Revised.*

ISSN: 2455-3220
Volume 02 Issue 09
September 2016

Master of Sciences Thesis submitted to the Faculty of Virginia Polytechnic Institute and State University

[10] Erguven, M. (2014). Two Approaches to Psychometric Process: Classical Test Theory and Item Response Theory. *Journal of Education*, *2*(2), 23-30.

[11] Matlock-Hetzel, S. (1997). Basic Concepts in Item and Test Analysis. *Texas A & M University, USA. files.eric.ed.gov/fulltext/ED406441.pdf.* Accessed on 24 June, 2014.

[12] Adegoke, B. A. (2013). Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks. *Journal of Education and Practice, 4*(22), 87-96.

[13] Zubairi, A. M., & Kassim, N. L. A. (2006). Classical and Rasch Analysis of Dichotomously Scored Reading Comprehension Test Items. *Malaysian Journal of ELT Research*, *2*, 1-20.

[14] Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge Mass: Newberry House Publisher.

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

[16] Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement (5th Ed*). Engelwood Cliffs, N.J: Prentice Hall.

[17] Carole L. K., & Winterstein, A. G. (2008). Validity and Reliability of Measurement Instruments used in Research. *American Journal Health-System Pharmacy*, Vol. 65 Dec 1, 2008