Journal of Applied Statistics and Machine Learning
Volume 1; Number 1; 2022 : pp. 1-13

COMPARISON OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY IN THE DISCRIMINATION OF HARD CORE AND PETTY CRIMINALS

Girdhar G. Agarwal[1] and Akash Asthana[2]

[1]*Professor (Retd.), Department of Statistics, University of Lucknow, Lucknow*
[2]*Assistant Professor, Department of Statistics, University of Lucknow, Lucknow*

ARTICLEINFO

Received: 27 July 2021
Revised: 12 August 2021
Accepted: 8 September 2021
Online: 11 January 2022

To cite this paper:
Agarwal, G.G., & Asthana, A. (2022). Comparison of Classical Test Theory and Item Response theory in the Discrimination of Hard Core and Petty Criminals. *Journal of Applied Statistics and Machine Learning*. 1(1): pp. 1-13

ABSTRACT

In many fields of research including Education, Psychology, Social Science, and Marketing Research, data is generally collected through a questionnaire, which is often referred to as the survey instrument. For the scale development of these questionnaires two different approaches, Classical Test Theory and Item Response Theory, are mostly prevalent. In our study a comparison is made between these two theories by the means of item parameters as well as person parameters. Guttman scaling is selected for the classical test theory, whereas one and two parameter models are used for the item response theory. Our data consists of an instrument containing 64 items assessing the adolescent life of hardcore criminals, petty criminals, and community persons. The study reveals that the difficulty parameters of classical test theory are highly negatively correlated with one parameter model of item response theory whereas the discrimination parameters of Classical test theory and two parameter model of Item response theory are uncorrelated. The present study also reveals that the person parameters of classical test theory, and parameters under two models of Item response theory are highly positively correlated.

Keywords: Classical test theory, Item response theory, Item parameters, Person parameters.

1. INTRODUCTION

In disciplines such as education, psychology, social sciences, and marketing research, data is generally collected through a survey instrument consisting of multiple-choice items. Unified quantitative theories are used that describe the behavior of test items and test scores under various conditions. For the scale development of instrument, two different approaches are used: Classical Test Theory (CTT) and Item Response Theory (IRT) (Kothari, 2004; Edwards, 1969).

1.1. Classical Test Theory: An Overview

It is assumed that the gross score (X) has two components (Cappeleri, Lundy, and Hays, 2014):

(i) the person's true score (T) which is the mean of the population of scores if the person were to be given an infinite number of measurement instrument,

(ii) the error (E)

It is assumed that the observed score X equals the true (unobservable) score T plus some error (E)

$$X_i = T_i + E_i$$

In this equation, errors (E_i) are random, independently, and identically distributed and uncorrelated with true score (T_i).

CTT deals with two different types of parameters known as *item parameters* and *person parameters*. Item parameters are used to judge the effectiveness and usefulness of the items for the test (questionnaire) depending upon characteristics considered in the test. The person parameters define the ability of the individuals for that particular trait. There are two item parameters: *item difficulty* and *item discrimination*. The difficulty in answering an item is judged with the help of *item difficulty*. The item difficulty can be defined as the proportion of the individuals who answered the question correctly. Higher is the value of this parameter, the lesser is difficulty of item.

The discrimination between two or more individuals is judged with the help of *item discrimination*. There are two methods for the estimation of item discrimination. In first method, the significant difference between proportion of the responses of the 27 percent of individuals having the higher scores with the response of the 27 percent of individuals having lower scores is used as discrimination parameter (Gulliksen, 1950). In the second method, biserial correlation coefficient between the score of an item (continuous variable) and group of individuals (categorical variable), is used as a measure of discrimination parameter. If the distribution of item scores is dichotomised, the tetra-choric correlation, phi-coefficient or other measures of correlation can also be used as a measure of discrimination parameter (Lord, 1974). In the present study as the item scores are dichotomized and there are three groups, Cramer's V is used as a measure of discrimination parameter.

There are several methods to measure the person parameters under CTT the main ones being the method of paired comparison, the method of

equal appearing interval, the method of successive interval, method of summated rating (also called as Likert Scale) and Guttman scaling (Edwards, 1969).

However, Likert Scale and Guttman scaling are most frequently used for obtaining the person parameters as the scores obtained by these methods are suitable for further mathematical and statistical computations. Guttman scaling is used when the questions are of binary response type whereas Likert scale is used when the questions are of multiple response type (more than two responses).

1.2. Item Response Theory: An Overview

The CTT model is limited in several ways. Item parameters are treated as fixed for a particular test. The CTT model has no allowance for possibly varying item parameters. Thus, the generality of true score is limited to tests with parallel or very similar collections of items. Also, item properties are not directly linked to behaviour. That is, knowing a person's score refers to an overall level relative to a group of persons. Nothing is known about which items the person has likely passed or likely failed. Thus, using item difficulty and discrimination to select items is justified by their impact on various population statistics, such as variances and reliabilities (Hambleton, Swaminathan, and Rogers, 1991).

Item response theory is developed to address these limitations of the CTT, and it is an improvement over CTT. IRT considers three item parameters (at least one in a model) called the item difficulty, discriminating power of item, guessing parameter, and parameter related to subject called as ability parameter (or IQ or trait under consideration) which provides score of subjects on the test. In IRT the estimates for item properties are not population specific, they are meaningful for all the groups of individuals. Moreover, the IRT models causally relate the score of subjects to the item parameters. IRT models are also called as 'Latent Trait Models'. IRT depends upon two basic postulates:

(i) The performance of an examinee on a test item can be predicted by a set of factors called as traits, latent traits, or abilities.

(ii) The relationship between examinees' performance on items and set of traits underlying item performance can be described by a monotonically increasing function, called as 'item characteristic function' or 'item characteristic curve' (ICC). The item characteristic function specifies that as the level of trait increases the probability of the correct response to an item also increases.

1.2.1. Models in IRT

There are two types of models in IRT, one defined for the binary response items and other defined for multiple response items. For the binary response type questions, the most popularly used models are: (i) One Parameter (1P), (ii) Two Parameter (2P), (iii) Three Parameter (3P) models, depending on number of item parameters used in the Model. Some of the important IRT models for multiple response items are Bock's Nominal Model, Samejima's Multiple-Choice Model, Rating Scale Model and Partial Credit Model (Hambleton , Swaminathan, and Rogers, 1991). The functions used to define these models are called 'Item Characteristic Function' and curve plotted for these functions using parameter estimate are called as 'Item Characteristic Curve'.

Lin (2008) has compared the CTT, 1P model and the 2P model of IRT for testing the parallelism of tests using alternate forms of a 60-items test from a pool of 600 items and showed that the CTT approach performed at least as well as the IRT approaches. Champlain (2010) has made a comparison between CTT and IRT and has shown that CTT and IRT are complementary approaches as each provides useful information at various phases of activity. Sharkness and DeAngelo (2011). have made a comparison between the CTT and IRT using Chornbach's Alpha as a measure of internal consistency of questionnaire and factor analysis for item selection under CTT and Graded response model under IRT and obtained that IRT provides much more information as compared to CTT. Solomon, Emaikwu, and Obinne (2020) have used an instrument consisting of 60 multiple choice items of May/June 2008 NECO SSCE Mathematics Paper to compare the CTT, 1P model, and 2P model of IRT by means of item parameters and obtained that the difficulty parameters of CTT and IRT are negatively correlated whereas discrimination parameter of CTT and 2P model of IRT are positively correlated. In these studies, the comparison is made by using either the person parameters or the item parameters. No researcher has considered both the parameters in a single study. The main objective of the present study is to compare the CTT and IRT by using both the item parameters and person parameters; the Guttman method of scaling under CTT is compared with the 1P model and 2P model of the IRT.

2. METHODOLOGY

In the present study the data is collected through a questionnaire, constructed for studying the adolescence life of the criminals, consisting of 64 questions related to the adolescence life events. All the questions are of binary response type. The survey is conducted over 750 individuals

divided into three groups: (i) Experimental group (EG), (ii) Control–I group (C1), (iii) Control–II group (C2), with each group consisting of 250 respondents.

2.1. Subjects

2.1.1. Experimental group

The experimental group consists of inmates charged for major offences (murder, attempt to murder, kidnapping, rape, forgery, robbery, dacoity) under specified criminal sections at least two times on different occasions with their cases being admitted by the courts of law for trial.

2.1.2. Control-I group

The individuals of this group are the prisoners charged for less serious offences (theft, house breaking, bribery, dishonestly receiving stolen property, hurt, rash driving, journey without ticket, gambling, and unlawful possession of arms) under specified sections at least two times on different occasions with their cases being admitted by the courts of law for trial.

2.1.3. Control-II group

These are the neighbours of individuals of the experimental group, belonging to the same age-group and socio-demographic background, having no evidence of specified criminal behaviour and willingness to cooperate.

2.2. Measurements

The questionnaire consists of sixty-four questions (Table 1) related to life of subject during childhood. The questionnaire is developed to identify the causes related to adolescent life of subjects that might have motivated them for committing a particular crime.

Table 1: Description of items of questionnaire

Item No.	*Item Description*	*Item No.*	
1.	Feeling insecure during childhood	33.	Bed wetting
2.	Area of insecurity (behavioural or physical)	34.	Temper tantrums
3.	Spent of spare time (with family or with friends)	35.	Panic attacks
4.	Unnecessary roaming	36.	Run away from school

contd. table 1

Item No.	*Item Description*	*Item No.*	
5.	See/ read violent/ sex films	37.	Run away from work
6.	See / read religious /social/ historical films / serials	38.	Aggressive and violent behaviour to others
7.	No means of amusement	39.	Aggressive and violent behaviour towards things
8.	Age of peers (same or not same)	40.	Attitude towards elders
9.	Habits of peers (good or bad)	41.	Run away from house
10.	Concurrence social / religious work	42.	Pity thefts
11.	Concurrence social/religious work by friends	43.	Often telling a lie
12.	Use of toxic drugs by person	44.	Mischievousness
13.	Use of toxic drugs by friends	45.	Abusing behaviour
14.	Death of father during childhood	46.	Cruel behaviour towards living animals
15.	Death of mother during childhood	47.	Shy / timid in childhood
16.	Father left the home in childhood	48.	Over sensitive
17.	Mother left the home in childhood	49.	Interest in impractical things
18.	Absence of father for a long time	50.	Irresponsible behaviour
19.	Absence of mother for a long time	51.	Discourage
20.	Family disputes during childhood	52.	Over conformist
21.	Competition in between siblings/ family members	53.	Day dreaming
22.	Acceptance of person by family members	54.	Cruel behaviour
23.	Physical/mental disease or handicappers in family	55.	Emotional gradient behaviour
24.	Physical / mental disease or handicap during childhood	56.	Zeal / determine
25.	Death in family	57.	Optimistic
26.	Compelled to leave home	58.	Over ambitious
27.	Deprivation	59.	Aggressive / irritating
28.	Natural disasters	60.	Anxiety
29.	Sudden changes in economic condition	61.	Cheerful
30.	Quarrelsome married life	62.	Emotional
31.	Nightmares	63.	Revolutionary
32.	Walk while sleep	64.	Dissatisfied

The questions numbered 1 to 13 are related to habits of subject. Questions numbered 14 to 30 are related to the incidents that potentially leave high impact on subject. Questions numbered 31 to 40 are related to behavioural trait of subjects and questions numbered 41 onwards are related to the personality traits in subject. All the questions are of binary response type (yes or no). The questions with special categories are mentioned with questions in brackets.

2.3. Statistical Analysis

As all the questions of the questionnaire are of binary response type the Guttman scaling of CTT and 1P and 2P models of the IRT are used. The item parameters using CTT and person parameters using Guttman scaling are computed using MS Excel. The item parameters and person parameters for IRT models are computed using R(2020), a software environment for statistical computing and graphics.

2.3.1. Guttman Scaling

Guttman scaling is also known as 'Cumulative Scaling' or 'Scalogram Analysis'. This method was proposed by Guttman L. in 1944. The purpose of Guttman scaling is to establish a one-dimensional continuum for a concept required to be measured. Unidimensional scales represent those scales in which a subject with more favourable attitude score than another subject must also be just as or more favourable to every item in the continuum than other subjects. In other words, a subject with higher rank than other subject should also have same or higher rank on every item in the set than the other subject (Kothari, 2004). The scale value is obtained by just summing their individual responses. The unidimensionality of the scale is tested by Goodenough's method (Gulliksen, 1950).

2.3.2. 1P Model

The 1P model is one of the most widely used models. It is also called the Rasch model in honour of its developer Rasch (Hambleton, Swaminathan, and Rogers, 1991). In this model it is assumed that the distribution of the trait is Logistic. The only parameter in the model is *item difficulty*, which is the characteristic of item that influences the performance of the subject. The 1P model is given by:

$$P_i(x) = \frac{e^{(x-\beta_i)}}{1+e^{(x-\beta_i)}}$$

where, $i = 1, 2, 3, \ldots, n$.

$P_i(x)$: probability that a randomly chosen subject with ability x answer i^{th} item correctly,

x : ability of the subject,

β_i : difficulty of the i^{th} item.

In this model the subject with low trait has zero probability of answering positively to the item.

2.3.3. 2P Model

Lord (Lin, 2008) was the first person to develop the two parameters item response model based on the normal distribution. Birnbaum used the logistic function in place of normal function for the two parameters model (Lin, 2008). Logistic function is more convenient to work with than the normal function. The 2PL model uses two parameters, called as item difficulty and item discrimination parameter (or discriminating power of item). The 2PL model is given by:

$$P_i(x) = \frac{e^{\alpha_i(x-\beta_i)}}{1+e^{\alpha_i(x-\beta_i)}}$$

where, $i = 1, 2, 3, \ldots, n$,

$P_i(x)$: probability that a randomly chosen subject with ability x answers i^{th} item correctly,

x : ability of the subject,

β_i : difficulty of the i^{th} item,

α_i : discriminating parameter of the i^{th} item.

The discriminating parameter of item is useful in differentiating between two groups. An item with high discriminating parameter (> 1) is useful than the item with small discriminating parameter (< 1). The range of discriminating parameter is $(-\infty, +\infty)$, but the items with negative discriminating parameter are eliminated from the ability test or modified while measuring the psychological traits, as they show that probability of answering correctly decreases as the ability increases.

3. RESULTS

Most of the individuals in each group are male (above 95%). Nearly 80% in each group are below 33 years of age. A little more than half of the individuals in each group are inhabitants of an urban area (52%) and nearly one-third from each group are employed in a farming occupation. Nearly half of the subjects in each group are either illiterate or had education below high school. The number of married subjects in the experimental group (51.6%) is less compared to community people (64%). The comparison of socio-demographic characteristic is made using chi-square test (Hogg, 2006). The three groups are found similarly distributed with respect to all the socio-demographic characteristics except the marital status (Table 2).

Table 2: Socio Demographic Profile of Sample

Variables	*Subject Type*			*Chi Square*
	Experimental	*Control-II*	*Control-I*	*(p-Value)*
Sex				
Male	239 (95.6%)	238 (95.2%)	239 (95.6%)	0.062 (0.970)
Female	11 (4.4%)	12 (4.8%)	11 (4.4%)	
Age				
Less than 33 years	200 (80.0%)	197 (78.8%)	205 (82.0%)	5.694 (0.223)
Between 33-43 years	44 (17.6%)	50 (20.0%)	36 (14.4%)	
More than 43 years	6 (2.4%)	3 (1.2%)	9 (3.6%)	
Domicile				
Rural	120 (48.0%)	120 (48.0%)	120 (48.0%)	< 0.001(> 0.999)
Urban	130 (52.0%)	130 (52.0%)	130 (52%)	
Occupation				
Currently unemployed	34 (13.6%)	34 (13.6%)	33 (13.2%)	7.091(0.717)
Farming	91 (36.4%)	99 (39.6%)	96 (38.4%)	
Service	21 (8.4%)	16 (6.4%)	12 (4.8%)	
Business	37 (14.8%)	47 (18.8%)	42 (16.8%)	
Self employed	57 (22.8%)	42 (16.8%)	53 (21.2%)	
Other	10 (4.0%)	12 (4.8%)	14 (5.6%)	
Marital Status				
Married	129 (51.6%)	160 (64.0%)	147 (58.8%)	15.125(0.004)
Unmarried	114 (45.6%)	89 (35.6%)	92 (36.8%)	
Others (Widow, Divorced, *etc.*)	7 (2.8%)	1 (0.4%)	11 (4.4%)	
Education				
Illiterate	1 (0.4%)	0 (0.0%)	0 (0.0%)	2.129(0.977)
Below high school	120 (48.0%)	120 (48.0%)	121 (48.4%)	
High school	94 (37.6%)	96 (38.4%)	96 (38.4%)	
Intermediate	22 (8.85%)	21 (8.4%)	20 (8.0%)	
More than intermediate	13 (5.2%)	13 (5.2%)	13 (5.2%)	

Item parameters computed using the CTT shows that item no. 33 is found to be easiest (difficulty = 99.733) whereas item no. 59 as most difficult (difficulty = 11.6) among all 64 items (Table 3). The discrimination for any item is not more than 0.510. The highest discriminating power is obtained for item no. 9 (discrimination = 0.501) whereas least for item no. 54 (discrimination = 0.036). Guttman scale scores lie in the range (31,61) (Table 4). A score of zero means the troublesome life and highest score of 61 means the trouble-free life during childhood (Table 4). Using Goodenough's method, the coefficient of reproducibility is obtained as 0.840 which suggests that the questionnaire is unidimensional.

Using 1P model of IRT item no. 33 (difficulty = –3.520) is obtained as easiest and item no. 59 (difficulty = 1.434) as most difficult. In this model the discrimination power of all the items is fixed up at 1.702. Whereas the 2P model of IRT suggests that item no. 17 (difficulty = –13.957) is obtained as easiest and item no. 60 (difficulty = 13.525) as the most difficult. The highest value of discrimination parameter is obtained for item no. 33 (discrimination = 3.361) and least for item no. 60 (discrimination = 0.004) (Table 3). The person parameters obtained by using 1P models of IRT lies between -2.55 and 2.18. The person parameter for 2P model of IRT ranges between -1.07 to 1.47. The score nearer to -3 represents the troublesome life during childhood and score nearer to 3 represents the trouble-free life during childhood (Table 4).

Table 3: Item parameters using CTT and IRT

Item No.	*CTT*		*1P Model*	*2P Model*	
	Difficulty	*Discrimination*	*Difficulty*	*Difficulty*	*Discrimination*
1	95.067	0.140	-1.760	-5.290	0.586
2	87.200	0.263	-1.122	-2.678	0.794
3	87.333	0.256	-1.130	-2.800	0.758
4	41.067	0.181	0.350	1.165	0.317
5	62.133	0.154	-0.213	-3.870	0.129
6	44.267	0.325	0.263	0.435	0.555
7	50.267	0.301	0.105	-0.031	0.527
8	49.467	0.325	0.126	0.029	0.568
9	47.733	0.501	0.171	0.052	1.300
10	60.667	0.129	-0.173	-1.938	0.227
11	57.733	0.410	-0.093	-0.376	1.139
12	92.400	0.196	-1.480	-3.218	0.861
13	94.400	0.192	-1.679	-3.461	0.907
14	97.600	0.109	-2.208	-4.769	0.838
15	97.467	0.124	-2.175	-4.236	0.948
16	92.933	0.126	-1.528	-5.388	0.496
17	92.933	0.103	-1.528	-13.957	0.186
18	98.933	0.083	-2.698	-6.860	0.691
19	99.467	0.107	-3.110	-3.693	1.726
20	97.333	0.119	-2.143	-5.519	0.688
21	98.933	0.091	-2.698	-4.601	1.091
22	88.800	0.152	-1.216	-5.514	0.385
23	98.133	0.059	-2.361	-6.184	0.672
24	83.467	0.218	-0.934	-3.241	0.526
25	98.667	0.073	-2.564	-4.224	1.145

contd. table 3

Item No.	*CTT*		*1P Model*	*2P Model*	
	Difficulty	*Discrimination*	*Difficulty*	*Difficulty*	*Discrimination*
26	98.000	0.128	-2.319	-3.884	1.134
27	96.667	0.164	-2.006	-3.143	1.264
28	98.933	0.168	-2.698	-2.986	2.006
29	94.667	0.082	-1.710	-9.130	0.320
30	68.000	0.214	-0.381	-2.667	0.288
31	99.467	0.060	-3.110	-4.061	1.508
32	96.267	0.184	-1.935	-2.771	1.447
33	99.733	0.103	-3.520	-2.885	3.361
34	92.000	0.244	-1.446	-2.266	1.342
35	92.133	0.238	-1.457	-2.750	1.032
36	84.267	0.241	-0.972	-2.698	0.675
37	92.533	0.110	-1.491	-8.989	0.284
38	83.333	0.331	-0.928	-1.524	1.390
39	84.133	0.269	-0.965	-2.086	0.921
40	91.867	0.275	-1.435	-2.277	1.319
41	71.200	0.391	-0.479	-1.019	1.124
42	82.800	0.390	-0.904	-1.374	1.610
43	46.133	0.120	0.214	1.503	0.103
44	93.200	0.300	-1.553	-1.942	1.933
45	98.400	0.143	-2.454	-2.947	1.799
46	52.667	0.301	0.042	-0.257	0.447
47	59.733	0.262	-0.147	-0.982	0.421
48	87.733	0.330	-1.152	-2.099	1.126
49	96.267	0.171	-1.935	-4.012	0.889
50	97.333	0.108	-2.143	-4.655	0.835
51	29.067	0.227	0.699	2.366	0.390
52	58.667	0.342	-0.118	-0.573	0.694
53	88.133	0.200	-1.176	-4.288	0.487
54	16.667	0.036	1.165	-3.439	0.490
55	12.800	0.041	1.363	-4.082	0.491
56	25.200	0.045	0.827	-2.876	0.391
57	39.200	0.340	0.401	0.744	0.633
58	89.733	0.186	-1.277	-3.518	0.661
59	11.600	0.188	1.434	7.730	0.266
60	22.667	0.129	0.918	13.525	0.004
61	31.733	0.165	0.617	7.893	0.097
62	49.067	0.357	0.136	0.026	0.878
63	87.867	0.250	-1.160	-3.056	0.703
64	53.067	0.451	0.031	-0.170	1.135

Table 4: Summary of scores under Guttman scaling of CTT, 1P and 2P models of IRT

Person Parameter	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Guttman Scale	31.00	61.00	47.8747	4.70778
1P Model	-2.55	2.18	-0.0411	0.86727
2P Model	-1.07	1.47	0.1034	0.38073

The difficulty parameters under CTT are computed as the percentage of individuals responded positively for the item. The correlation between difficulty parameters of the CTT and 1P model and 2P Model of IRT suggests that the difficulty parameters of the CTT and 1P model of IRT are highly negatively correlated whereas the difficulty parameter of CTT is moderately negatively correlated with that of 2P model. The negative correlation between two approaches (CTT and IRT) is due to method of measuring the difficulty parameter. In CTT, the difficulty parameter is biased back ward (high score means easy item), whereas in IRT, the parameter is biased forward (low score means easy item). The difficulty parameter of 1P model is moderately positively correlated with that of 2P model. The item discrimination parameter of CTT is not associated with that of 2P model (correlation coefficient = 0.151), which means that for smaller value of the discrimination parameter under CTT no decision can be made about the discrimination parameter of 2P model of IRT.

Table 5: Correlation between item parameters of CTT, 1P and 2P model of IRT

Parameter	*Variables*	*Correlation*	*p-Value*
Item Difficulty	CTT and 1P Model	-0.942	<0.001
	CTT and 2P Model	-0.662	<0.001
	1P and 2P Model	0.605	<0.001
Item Discrimination	CTT and 2P Model	0.151	<0.001
Person Parameter	CTT and 1P Model	0.911	<0.001
	CTT and 2P Model	0.996	<0.001
	1P and 2P Model	0.913	<0.001

4. DISCUSSION

From the correlation analysis person parameters obtained by Guttman scaling of CTT, 1P model and 2P model of IRT are highly correlated. Hence, it can be concluded that irrespective of the fact whether the item parameters are obtained through CTT or 1P model or 2P model of IRT the person parameters are related and are highly correlated. This study supports the result of study by Awopeju *et al.* [8] that the difficulty parameter of CTT

and IRT are negatively correlated but contradicts the finding that discrimination parameter of CTT and 2P model of IRT are correlated.

References

1. Awopeju, O. A. and Afolabi, E. R. I. (2016). *Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination*. European Scientific Journal vol.12:28. pp: 263-284. ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431
2. Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). *Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures*. Clinical therapeutics, 36(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006
3. De Champlain A.F. (2010). *A primer on classical test theory and item response theory for assessment in medical education.* Medical Education. Vol.44:1. pp: 109-117.
4. Edwards A. L. (1969). *Techniques of Attitude Scale Construction*. Valkis, Feffer and Simons Pvt. Ltd. Bombay (India).
5. Gulliksen H. (1950). *Theory of Mental Test.* chapter 20. John Willey & Sons, New York (USA).
6. Hambleton R. K., Swaminathan H., and Rogers H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications Inc., London (U. K.).
7. Hogg R.V. and Craig A.T. (2006). *Introduction to Mathematical Statistics*, Pearson Education Inc., New York.
8. Kothari C. R. (2004). *Research Methodology, Methods and Techniques*. chapter 5. New Age International (P) Ltd. New Delhi (India).
9. Lin C. J. (2008). *Comparisons between classical test theory and item response theory in automated assembly of parallel test forms.* The Journal of Technology, Learning, and Assessment. Vol. 6:8. pp: 1:42.
10. Lord F.M. and Novick M.R. (1974). *Statistical theories of mental test scores*. Information Age Publishing Inc, USA.
11. R Core Team (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
12. Sharkness, J., DeAngelo, L. (2011). Measuring Student Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys. *Res High Educ* **52,** 480–507. https://doi.org/10.1007/s11162-010-9202-3
13. Solomon, A., Emaikwu S. O, and Obinne A.D.E (2020). *Comparative Analysis of Classical Test Theory and Item Response Theory in Estimating Item Difficulty of BECE Mathematics Objective Items in Makurdi-Nigeria.* World Journal of Innovative Research (WJIR), Volume-9, Issue-2, August 2020 Pages 24-31, ISSN: 2454-8236.