Journal of Educational System
Volume 3, Issue 2, 2019, PP 31-41
ISSN 2637-5877

Comparison of the Selection of Items Using Classical Test Theory and Item Response Theory Based on Sample Sizes

Chinelo Blessing Oribhabor[1*], Judith Hannah Osarumwense[2]

[1]Department of Education, Faculty of Arts and Education, University of Africa, Toru-Orua, Bayelsa State, Nigeria
[2]Department of Educational Evaluation and Counselling Psychology, Faculty of Education,University of Benin, Benin City, Edo State, Nigeria

***Corresponding Author:** Chinelo Blessing Oribhabor, Department of Education, Faculty of Arts and Education, University of Africa, Toru-Orua, Bayelsa State, Nigeria, Email: chiblessing42004@yahoo.co.uk*

ABSTRACT

The study examined the comparability of item statistics of classical test theory and item response theory of the selected items for the several sample sizes (from N=200 to N=1000). A 60-item multiple choice Mathematics test which was developed by Edo State Ministry of Basic Education Board, Nigeria was used to gather data from the various randomly selected sample. Results showed that more items were selected from the 60-item Basic Education Mathematics Examination for the various sample sizes when item response theory-based item statistics estimates were used than when classical test theory-based item statistics estimates were used, it was also found that the item discrimination indices of classical test theory and item response theory were comparable for different sample sizes while the item difficulty indices of classical test theory and item response theory were not comparable only on the very small sample of 200.

Keywords: *Classical Test Theory, Item Response Theory, Sample Sizes, Item Difficulty, Item Discrimination.*

BACKGROUND OF THE STUDY

The essence of developing test is to construct a test that have a desired quality by selecting the appropriate items, not minding the type of tool that was used. There are two main test theories, such as classical test theory (CTT) and item response theory (IRT). According to Greg (2009), classical test theory is a body of theory and research regarding psychological testing that predicts/explains the difficulty of questions, provides insight into reliability of assessment scores, and helps in representing what examinees know and can do. The essential basis of classical test theory (CTT) is that many questions combine to produce a measurement (assessment score) representing what a test taker knows and can do. CTT works well for most assessment applications for reason such as its ability to work with smaller sample size (example, 100 or less), and that it is relatively simple to compute and understand the statistics (Greg, 2009). In classical test theory, statistics such as item difficulty index, item discrimination index, indices of reliability, and validity which are used for interpreting test scores are sample dependent. This means that classical test theory has the limitation of sample dependency for estimating the test item parameters of item difficulty and item discrimination. Adedoyin and Adedoyin (2013) opined that classical test theory of item parameter estimates using heterogeneous samples generally result in higher estimates of item discrimination indices as measured by point-biserial correlation coefficients, whereas item difficulty estimates rise and fall with high and low ability groups of examinees.

The limitation of classical test theory has paved way for increase in attention for item responses theory. According to Adedoyin and Adedoyin (2013), item response theory item statistics depend to a great extent on the characteristics of the examinee sample used in the analysis. As the name indicates, IRT focuses primarily on the item-level information in contrast to the CTT's primary focus on test-level information. Also, the focus on estimating an item characteristic curve of item response theory for each item provides an integrative, holistic view of the performance of each item that is not readily available when using CTT can quantity the total-sample difficulty or discrimination for an

item, it lacks an effective means for simultaneously combining and showing this information in an easily-used format. One major factor that affects the stability and accuracy of model parameters is the sample sizes used to estimate the items. He and Weheadon (2012) have it that the probabilistic nature of item response theory reveals that sample size is an important factor that affects the accuracy and stability on the estimation of model parameters. Demars (2003) also studied the effect of sample size on parameter estimation for polytomous items with the NRM and found out that the magnitude of the variation between sample estimates decreases with increasing sample size.

Hulin, Lissak and Drasgrow (1982) opined that the sample sizes (577 ad 589) are adequate for the complexity of the two-parameter logistic model based on Monten Carlo simulation study. Hulin and colleagues simulated binary item responses (1= correct, 0= incorrect) for various sample sizes (200, 500, 1000 and 2000) and suggested that a sample size of 500 was adequate for a 2PL model. Baruch (1980) opined that the larger the sample, the smaller the standard error of the item's characteristics. The index of difficulty of an item in the population measured by the percentage of correct responses, the item-total score correlation in the population and other items' parameters can be estimated more accurately when a larger sample is employed. Hula, Fergadiotis and Martin (2012), in their study of identifying the most appropriate item response theory measurement model for aphasia tests found out that with small and medium sizes, an augmented one-parameter logistic model was the most accurate at recovering the known item and person parameters and no model performed well at any sample size.

There are empirical studies on the comparability of the item and person parameters using CTT and IRT. MacDonald andPaunonen (2002) in their Monte Carlo comparison of item and person statistic found a very high comparability for test scores and difficulty and less comparability for item discrimination. Courville (2005) in his study found high correlations between the CTT and the IRT test item difficulties, it was also found that discrimination indices correlated highly only when the spread of discrimination was large and the spread of difficulty values was small. Adedoyin and Adedoyin (2013) assessed the comparability between classical test theory and item response theory models in estimating test item parameters and found out that the CTT and IRT item difficulty and item discrimination values were positively linearly correlated and there was no statistical significant difference between the item difficulty and item discrimination parameter estimates by CTT and IRT. There is need to take into consideration the sample sizes used in estimating the item statistics parameter estimates for the classical test theory and item response theory. Thus, the study seems to examine the comparability of item statistics parameter estimates of different samples for classical test theory and item response theory.

Problem of the Study

The issue of sample independent and item characteristics are very crucial for objective measurement. Ojerinde (2013) observed and questioned if these demands are sufficient to discard classical test theory for item response theory. When a very large sample is used for the estimation of the item parameters, will it be reasonable to prefer IRT over CTT? How close will the item difficulty and item discrimination parameter estimates when different sample sizes of examinees CTT and IRT are used? The main purpose of this study is to determine the comparability of CTT and IRT in the estimation of test item parameter estimates of the 2013 Edo State Basic Education Certificate Mathematics Examination, Nigeria using different sample sizes of examinees.

Research Questions

The following research questions are raised to guide the study:

- Do items selected based on the sample sizes differ for classical test theory?
- Do items selected based on the sample sizes differ for item response theory?
- Are item selected based on the sample sizes for both CTT and IRT comparable?

Hypothesis

Only research question three was hypothesized.

- There is no significant statistical comparability between the CTT and IRT items selected based on the sample sizes.

Significance of the Study

The findings of the study will increase the empirical knowledge based on CTT and IRT

theoretical frameworks. This study will be of great importance to researchers in educational measurement field, teachers and examination bodies who seek for objective, reliable and valid measurement approach in analyzing, interpreting examination scores.

MATERIALS AND METHOD

Survey research design was employed in the study. The population of this study comprised of all the Upper Basic (JSS3) students in the seven local government area of Edo South Senatorial District, Edo State, Nigeria. Random samples with different sizes were drawn from the population using replacement sampling. Five sample sizes were investigated: 200, 400, 600, 800 and 1000. For each sample size, 10 replicates (repeated samples) were produced. Edo State Basic Education Certificate Mathematics multiple choice examination paper was the instrument used to collect data. The instrument was assumed to be reliable and valid due to its nature as a standardized test administered by the examination board. The instrument was administered to the students with the aid of their Mathematics teachers.

Data gathered were then analyzed using SPSS version 20, Baye's Model Estimator and Microsoft Excel version 20. Classical test theory analysis was done using the Excel version 2010 which generated the item difficulty and item discrimination for the various sample sizes. IRT parameter were estimated using Baye's model estimator program which generated the item difficulty, item discrimination and guessing parameter for the various sample sizes. Dependent t-test was used for testing hypothesis on the comparability of CTT and IRT items selected based on the sample sizes.

RESULTS

Research Question One: Do items selected based on the sample sizes differ for classical test theory?

Table1. *Item discrimination parameter estimates (a-parameter) of CTT for different sample sizes*

Item	N=200	N=400	N=600	N=800	N=1000
1	0.01	0.01	0.02	0.01	0.02
2	0.20	0.24	0.27	0.25	0.27
3	0.16	0.19	0.23	0.21	0.22
4	0.03	0.07	0.14	0.11	0.12
5	0.18	0.17	0.20	0.19	0.19
6	0.14	0.15	0.18	0.17	0.17
7	0.29	0.31*	0.33*	0.31*	0.31*
8	0.28	0.29	0.30*	0.30*	0.31*
9	0.16	0.15	0.17	0.16	0.16
10	0.20	0.22	0.22	0.21	0.21
11	0.11	0.08	0.08	0.08	0.07
12	0.18	0.21	0.24	0.22	0.22
13	0.08	0.08	0.07	0.07	0.07
14	0.15	0.17	0.17	0.17	0.18
15	0.20	0.21	0.19	0.19	0.20
16	0.34*	0.33*	0.31*	0.31*	0.31*
17	0.49*	0.50*	0.44*	0.44*	0.45*
18	0.17	0.20	0.21	0.20	0.20
19	0.23	0.23	0.19	0.20	0.20
20	0.27	0.29	0.27	0.28	0.29
21	0.12	0.13	0.11	0.11	0.12
22	0.21	0.20	0.19	0.20	0.21
23	0.07	0.06	0.05	0.06	0.06
24	0.07	0.07	0.09	0.09	0.09
25	0.20	0.18	0.16	0.17	0.17
26	0.31*	0.29	0.28	0.29	0.29
27	0.35*	0.35*	0.34*	0.34*	0.34*
28	0.26	0.26	0.24	0.25	0.26
29	0.11	0.10	0.09	0.09	0.09
30	0.26	0.28	0.30*	0.29	0.29
31	0.20	0.18	0.17	0.17	0.17
32	0.20	0.21	0.21	0.21	0.21

33	0.11	0.13	0.17	0.16	0.15
34	0.14	0.15	0.15	0.15	0.15
35	0.15	0.15	0.17	0.17	0.17
36	0.21	0.21	0.20	0.20	0.20
37	0.21	0.23	0.23	0.23	0.24
38	0.24	0.27	0.29	0.28	0.29
39	0.23	0.27	0.27	0.27	0.28
40	0.24	0.23	0.21	0.22	0.22
41	0.24	0.24	0.21	0.21	0.21
42	0.20	0.18	0.12	0.15	0.15
43	0.46*	0.45*	0.39	0.41*	0.42*
44	0.32*	0.28	0.21	0.24	0.24
45	0.12	0.12	0.12	0.12	0.12
46	0.16	0.15	0.18	0.17	0.16
47	0.11	0.08	0.07	0.08	0.08
48	0.15	0.15	0.15	0.15	0.15
49	0.23	0.26	0.30*	0.29	0.28
50	0.07	0.07	0.08	0.08	0.08
51	0.23	0.21	0.14	0.16	0.16
52	0.23	0.23	0.24	0.24	0.24
53	0.10	0.09	0.03	0.05	0.06
54	0.04	0.06	0.04	0.04	0.04
55	0.08	0.10	0.11	0.10	0.10
56	0.14	0.15	0.16	0.15	0.15
57	-0.01	-0.03	-0.05	-0.04	-0.01
58	0.09	0.02	0.04	0.03	0.08
59	0.08	-0.02	0.01	-0.01	0.05
60	0.01	0.02	0.04	0.03	0.05

Note: (*) means accepted item

Table2. *Item difficulty parameter estimates (b-parameter) of CTT for different sample sizes*

Item	N=200	N=400	N=600	N=800	N=1000
1	0.61	0.60*	0.64	0.63	0.62
2	0.59*	0.58*	0.60*	0.60*	0.59*
3	0.52*	0.50*	0.45*	0.47*	0.47*
4	0.47*	0.47*	0.47*	0.47*	0.47*
5	0.50*	0.48*	0.47*	0.47*	0.47*
6	0.48*	0.48*	0.44*	0.45*	0.45*
7	0.43*	0.40*	0.40*	0.41*	0.40*
8	0.43*	0.43*	0.41*	0.42*	0.42*
9	0.34	0.32	0.28	0.30	0.30
10	0.40*	0.38	0.36	0.37	0.37
11	0.37	0.36	0.40*	0.39	0.38
12	0.34	0.33	0.31	0.31	0.31
13	0.31	0.32	0.32	0.32	0.32
14	0.32	0.31	0.29	0.29	0.30
15	0.37	0.36	0.36	0.36	0.36
16	0.47*	0.48*	0.45*	0.46*	0.46*
17	0.38	0.37	0.43*	0.41*	0.40*
18	0.38	0.38	0.38	0.38	0.38
19	0.32	0.31	0.37	0.35	0.34
20	0.32	0.31	0.29	0.30	0.30
21	0.37	0.36	0.33	0.34	0.34
22	0.35	0.35	0.33	0.33	0.34
23	0.41*	0.41*	0.41*	0.41*	0.41*
24	0.45*	0.45*	0.43*	0.43*	0.44*
25	0.42*	0.41*	0.37	0.38	0.39
26	0.32	0.31	0.29	0.30	0.30

27	0.27	0.26	0.27	0.27	0.27
28	0.38	0.32	0.34	0.34	0.35
29	0.38	0.36	0.33	0.34	0.34
30	0.37	0.36	0.34	0.35	0.35
31	0.37	0.37	0.35	0.36	0.36
32	0.31	0.30	0.27	0.28	0.28
33	0.37	0.36	0.36	0.36	0.36
34	0.39	0.37	0.35	0.36	0.36
35	0.41*	0.40*	0.38	0.39	0.39
36	0.39	0.37	0.36	0.36	0.36
37	0.45*	0.43*	0.41*	0.42*	0.42*
38	0.42*	0.40*	0.40*	0.40*	0.40*
39	0.36	0.36	0.38	0.38	0.37
40	0.39	0.38	0.38	0.38	0.38
41	0.41*	0.40*	0.36	0.37	0.38
42	0.35	0.32	0.30	0.31	0.31
43	0.38	0.35	0.28	0.31	0.31
44	0.37	0.36	0.37	0.37	0.37
45	0.37	0.33	0.31	0.32	0.32
46	0.39	0.37	0.36	0.36	0.36
47	0.40*	0.40*	0.42*	0.41*	0.41*
48	0.32	0.32	0.32	0.32	0.32
49	0.29	0.27	0.28	0.28	0.27
50	0.30	0.28	0.29	0.29	0.29
51	0.44*	0.42*	0.39	0.40*	0.40*
52	0.44*	0.46*	0.46*	0.45*	0.46*
53	0.47*	0.46*	0.43*	0.44*	0.44*
54	0.49*	0.46*	0.42*	0.43*	0.43
55	0.39	0.37	0.37	0.38	0.37
56	0.48*	0.48*	0.50*	0.50*	0.49*
57	0.70	0.66	0.64	0.68	0.66
58	0.66	0.62	0.61	0.64	0.62
59	0.38	0.43*	0.45*	0.42*	0.44*
60	0.48*	0.47*	0.47*	0.47*	0.47*

Note: (*) means accepted item

Table3. *Summary of the comparison of the number of items selected based on different sample sizes for classical test theory (CTT)*

Sample Size	**Number of Items Selected**	
	Item Difficulty ($0.4 \leq b \geq 0.6$)	Item Discrimination ($a \geq 0.3$)
N = 200	23	6
N = 400	24	5
N = 600	22	7
N = 800	21	6
N = 1000	21	6

Table 3 depicts the item statistics derived from the classical test theory based on different sample sizes. The total number of items selected on the basis of difficulty index for CTT were 23, 24, 22, 21 and 21 for the sample sizes of 200, 400, 600, 800 and 1000 respectively, while the total number of items selected on the basis of discrimination index for CTT were 6, 5, 7, 6, 6 for the sample sizes of 200, 400, 600, 800 and 1000 respectively. This means that item selected based on the sample sizes differ for CTT.

Research Question Two: Do items selected based on the sample sizes differ for item response theory?

Table4. *Item difficulty parameter estimates (b - parameter) of IRT for different sample sizes*

Item	**N=200**	**N=400**	**N=600**	**N=800**	**N=1000**
1	-0.34*	-0.37*	-1.06*	-0.84*	-0.80*
2	-0.21*	-0.49*	-0.66*	-0.69*	0.84*
3	0.82*	0.29*	0.34*	0.34*	0.07*

4	1.85*	1.36*	0.84*	1.20*	0.70*
5	0.76*	0.20*	0.16*	0.14*	-0.14*
6	0.89*	0.11*	0.21*	0.16*	-0.12*
7	0.74*	0.30*	0.22*	0.22*	0.02*
8	0.98*	0.13*	0.15*	0.09*	-0.14*
9	1.71*	0.62*	0.66*	0.63*	0.42*
10	1.23*	0.47*	0.42*	0.47*	0.15*
11	2.32*	0.65*	1.40*	1.68*	1.30*
12	1.78*	0.83*	0.77*	0.87*	0.58*
13	1.89*	0.95*	1.15*	1.09*	0.89*
14	1.61*	1.25*	1.11*	1.09*	0.94*
15	1.20*	0.79*	0.83*	0.80*	0.48*
16	0.34*	0.08*	0.15*	0.12*	-0.08*
17	0.39*	0.07*	0.05*	0.05*	-0.17*
18	0.81*	0.39*	0.51*	0.49*	0.30*
19	0.87*	0.44*	0.38*	0.39*	0.18*
20	1.01*	0.95*	0.95*	0.88*	0.69*
21	1.69*	1.25*	1.65*	1.64*	1.24*
22	1.60*	0.82*	0.90*	0.81*	0.50*
23	1.32*	1.64*	2.06*	1.99*	1.91*
24	1.19*	0.82*	0.96*	0.76*	0.61*
25	0.80*	0.41*	0.31*	0.26*	0.00*
26	0.77*	0.68*	0.49*	0.41*	0.39*
27	0.60*	0.59*	0.43*	0.38*	0.00*
28	0.57*	0.40*	0.48*	0.41*	0.21*
29	0.89*	1.43*	1.65*	1.42*	1.66*
30	0.68*	0.42*	0.38*	0.34*	0.16*
31	0.73*	0.78*	0.85*	0.71*	0.68*
32	0.92*	1.12*	1.07*	0.97*	0.93*
33	1.38*	1.64*	1.25*	1.34*	1.28*
34	1.75*	1.38*	0.89*	0.94*	0.65*
35	1.59*	1.04*	0.77*	0.86*	0.56*
36	1.29*	0.52*	0.36*	0.43*	0.09*
37	0.86*	0.66*	0.49*	0.52*	0.34*
38	1.17*	0.42*	0.22*	0.22*	-0.02*
39	1.29*	0.59*	0.32*	0.35*	0.12*
40	1.23*	0.37*	0.56*	0.57*	0.19*
41	1.05*	0.45*	0.64*	0.63*	0.31*
42	1.17*	1.42*	1.93*	1.58*	1.47*
43	0.58*	0.28*	0.48*	0.37*	0.13*
44	0.76*	0.58*	0.70*	0.62*	0.36*
45	1.55*	1.05*	1.74*	1.74*	1.19*
46	1.06*	1.23*	0.89*	0.93*	0.70*
47	1.60*	1.56*	1.30*	1.48*	1.16*
48	1.30*	1.13*	0.95*	0.94*	0.84*
49	1.42*	1.25*	0.56*	0.61*	0.52*
50	2.73	3.01	2.70	2.89*	3.10*
51	0.89*	0.87*	1.31*	1.12*	0.98*
52	0.56*	0.25*	0.16*	0.16*	-0.10*
53	0.86*	1.01*	1.92*	1.67*	1.07*
54	1.09*	1.24*	1.98*	1.69*	1.57*
55	1.72*	2.03*	1.96*	2.10*	2.08*
56	0.65*	0.38*	0.19*	0.21*	0.05*
57	-1.74*	-1.36*	-1.05*	-1.47*	-1.35*
58	-0.89*	-0.82*	-0.59*	-1.14*	-1.05*
59	2.28*	2.41*	2.37*	2.16*	2.18*
60	1.68*	1.91*	1.84*	2.06*	1.73*

Note: (*) means accepted item.

Table5. *Item discrimination parameter estimates (a - parameter) of IRT for different sample sizes*

Item	N=200	N=400	N=600	N=800	N=1000
1	0.10	0.08	0.09	0.07	0.07
2	0.28	0.46*	0.63*	0.56*	0.60*
3	0.18	0.37	0.45*	0.37	0.47*
4	0.14	0.20	0.84*	0.29	0.87*
5	0.26	0.52*	0.46*	0.44*	0.49*
6	0.27	0.48*	0.54*	0.50*	0.51*
7	0.45*	1.64*	0.90*	0.80*	1.16*
8	0.35	0.70*	0.77*	0.72*	0.79*
9	0.46*	2.81*	2.90*	2.59*	3.14*
10	0.43*	3.33*	2.71*	3.84*	2.21*
11	0.28	0.75*	0.29	0.27	0.32
12	0.39	0.67*	0.71*	0.65*	0.69*
13	0.66*	1.80*	1.34*	1.47*	1.49*
14	0.46*	0.56*	0.70*	0.70*	0.58*
15	0.46*	0.49*	0.45*	0.43*	0.45*
16	0.66*	2.22*	1.23*	1.11*	1.65*
17	1.14*	3.33*	1.73*	1.69*	2.07*
18	0.85*	1.63*	1.29*	1.32*	1.32*
19	0.82*	1.03*	0.64*	0.66*	0.68*
20	0.85*	0.52*	0.52*	0.51*	0.49*
21	0.34	0.35	0.30	0.28	0.30
22	0.40*	0.49*	0.47*	0.46*	0.49*
23	1.07*	0.28	0.21	0.23	0.21
24	0.31	0.31	0.33	0.36	0.30
25	2.78*	2.01*	4.01*	3.40*	1.00*
26	2.58*	1.29*	3.06*	2.98*	1.75*
27	4.17*	0.99*	1.15*	1.24*	1.00*
28	4.90*	0.91*	1.07*	1.18*	0.74*
29	2.86*	0.36	0.37	0.38	0.30
30	3.01*	0.74*	0.86*	0.80*	0.72*
31	2.34*	0.49*	0.50*	0.51*	0.41*
32	1.06*	0.54*	0.60*	0.62*	0.50*
33	0.65*	0.36	0.41*	0.40*	0.36
34	0.34	0.63*	1.31*	1.04*	1.55*
35	0.29	0.73*	1.14*	1.00*	1.10*
36	0.71*	2.64*	4.61*	3.62*	3.46*
37	0.41*	0.42*	1.93*	1.58*	1.04*
38	0.32	0.54*	0.69*	0.60*	0.68*
39	0.42*	0.60*	0.70*	0.66*	0.68*
40	0.43*	2.77*	1.47*	1.31*	2.37*
41	0.53*	2.69*	2.01*	1.96*	2.52*
42	0.73*	0.41*	0.35	0.37	0.34
43	0.83*	0.88*	0.90*	0.87*	0.89*
44	0.77*	0.72*	0.62*	0.65*	0.63*
45	0.40*	0.76*	0.43*	0.43*	0.46*
46	0.48*	0.34	0.43*	0.39	0.38
47	0.31	0.32	0.30	0.27	0.36
48	0.58*	0.47*	0.66*	0.57*	0.49*
49	0.56*	0.50*	0.79*	0.71*	0.64*
50	0.38	0.40*	0.38	0.38	0.29
51	0.45*	0.49*	0.37	0.43*	0.48*
52	0.64*	0.51*	0.58*	0.57*	0.52*
53	0.45*	0.27	0.19	0.50*	0.72*
54	0.21	0.20	0.23	0.19	0.18
55	0.30	0.23	0.24	0.21	0.20
56	0.57*	0.38	0.38	0.40*	0.35

57	0.15	0.10	0.07	0.07	0.07
58	0.25	0.14	0.12	0.11	0.17
59	0.26	0.14	0.12	0.12	0.14
60	0.15	0.13	0.12	0.11	0.12

Note: (*) means accepted item.

Table6. *Guessing parameter estimates (c - parameter) of IRT for different sample sizes*

Item	N=200	N=400	N=600	N=800	N=1000
1	0.19*	0.20*	0.23	0.23	0.23
2	0.16*	0.13*	0.11*	0.10*	0.10*
3	0.16*	0.16*	0.14*	0.14*	0.15*
4	0.15*	0.18*	0.34	0.23	0.36
5	0.16*	0.16*	0.12*	0.12*	0.11*
6	0.15*	0.12*	0.11*	0.10*	0.09*
7	0.12*	0.20*	0.12*	0.11*	0.15*
8	0.13*	0.09*	0.09*	0.07*	0.07*
9	0.16*	0.24	0.22	0.22	0.23
10	0.16*	0.27	0.25	0.29	0.26
11	0.17*	0.23	0.15*	0.15*	0.16*
12	0.14*	0.13*	0.10*	0.12*	0.11*
13	0.21	0.26	0.27	0.27	0.27
14	0.13*	0.14*	0.14*	0.14*	0.12*
15	0.13*	0.10*	0.10*	0.09*	0.07*
16	0.14*	0.23	0.20*	0.19*	0.23
17	0.07*	0.08*	0.14*	0.12*	0.11*
18	0.16*	0.20*	0.21	0.21	0.22
19	0.09*	0.08*	0.07*	0.06*	0.05*
20	0.14*	0.08*	0.06*	0.05*	0.05*
21	0.15*	0.11*	0.10*	0.09*	0.08*
22	0.14*	0.09*	0.07*	0.06*	0.06*
23	0.31	0.18*	0.17*	0.17*	0.17*
24	0.17*	0.16*	0.16*	0.15*	0.13*
25	0.29	0.31	0.29	0.28	0.20*
26	0.18*	0.17*	0.18*	0.17*	0.18*
27	0.11*	0.05*	0.05*	0.04*	0.20*
28	0.19*	0.13*	0.13*	0.14*	0.08*
29	0.29	0.14*	0.14*	0.13*	0.12*
30	0.20*	0.08*	0.06*	0.06*	0.05*
31	0.22	0.11*	0.11*	0.10*	0.09*
32	0.13*	0.10*	0.09*	0.08*	0.07*
33	0.21	0.17*	0.14*	0.15*	0.15*
34	0.18*	0.25	0.27	0.27	0.29
35	0.15*	0.26	0.26	0.27	0.27
36	0.23	0.26	0.24	0.25	0.24
37	0.16*	0.15*	0.28	0.29	0.25
38	0.13*	0.10*	0.08*	0.07*	0.07*
39	0.12*	0.10*	0.09*	0.08*	0.08*
40	0.15*	0.23	0.24	0.24	0.24
41	0.18*	0.27	0.27	0.28	0.28
42	0.17*	0.12*	0.11*	0.10*	0.09*
43	0.10*	0.05*	0.04*	0.04*	0.03*
44	0.12*	0.12*	0.14*	0.14*	0.11*
45	0.15*	0.19*	0.15*	0.17*	0.14*
46	0.14*	0.12*	0.10*	0.10*	0.08*
47	0.15*	0.18*	0.17*	0.17*	0.20*
48	0.13*	0.11*	0.14*	0.11*	0.10*
49	0.10*	0.07*	0.04*	0.04*	0.03*
50	0.18*	0.21	0.19*	0.20*	0.17*

51	0.17*	0.19*	0.18*	0.19*	0.22
52	0.15*	0.13*	0.13*	0.12*	0.10*
53	0.20*	0.17*	0.17*	0.33	0.26
54	0.16*	0.15*	0.15*	0.15*	0.14*
55	0.15*	0.14*	0.13*	0.13*	0.12*
56	0.21	0.16*	0.16*	0.16*	0.15*
57	0.22	0.24	0.25	0.29	0.29
58	0.18*	0.20*	0.19*	0.21	0.19*
59	0.17*	0.15*	0.15*	0.14*	0.16*
60	0.15*	0.15*	0.15*	0.14*	0.16*

Note: (*) means accepted item.

Table7. *Summary of Comparison of the number of items selected based on different sample sizes for item response theory*

Sample Size	**Number of Items Selected**		
	a-parameter (a ≥ 0.4)	**b-parameter (-2.5 ≤ +2.5)**	**c-parameter c ≤ 0.2**
200	38	59	51
400	42	59	47
600	42	59	46
800	42	59	44
1000	42	59	44

Table 7 depicts the item statistics derived from the item response theory based on different sample sizes. The total numbers of items selected on the basis of the discrimination index for IRT were 38, 42, 42, 42 and 42 for the sample sizes of 200, 400, 600, 800 and 1000 respectively; while the total number of items selected on the basis of difficulty index for IRT were 59, 59, 59, 59 and 59 for the sample sizes of 200, 400, 600, 800 and 1000 respectively. This means that the items selected based on the sample sizes do not differ for IRT except for the sample size of 200 for item discrimination value.

Hypothesis One: There is no significant statistical comparability between the CTT and IRT items selected based on the sample sizes.

Table8. *Dependent t-test analysis of comparability between the CTT and IRT items selected based on the sample sizes*

Variable	**Mean**	**SD**	**T**	**df**	**Sig.(2-tailed)**
CTT DF 200 – IRT DF 200	-1.109	1.359	-3.914	22	0.001
CTT DF 400 – IRT DF 400	-0.110	0.555	-0.970	23	0.342
CTT DF 600 – IRT DF 600	-0.065	0.643	-0.472	21	0.642
CTT DF 800 – IRT DF 800	-0.069	0.669	-0.474	20	0.640
CTT DF 1000 – IRT DF 1000	0.183	0.600	1.399	20	0.177
CTT DS 200 – IRT DS 200	-0.032	0.200	-0.390	5	0.713
CTT DS 400 – IRT DS400	-0.231	0.561	-0.919	4	0.410
CTT DS 600 – IRT DS 600	-0.215	0.291	-1.959	6	0.098
CTT DS 800 – IRT DS 800	-0.152	0.267	-1.397	5	0.221
CTT DS 1000 – IRT DS1000	-0.134	0.247	-1.327	5	0.242

Table 8 revealed that there was a statistical significant difference between the item difficulty parameter estimates of 200 sample size by CTT and IRT (t_{22} = -3.914) and it was also revealed that there was no statistical significant difference between the item difficulty parameter estimates of 400, 600, 800 and 1000 sample sizes of CTT and IRT (t_{33} = -0.970; t_{21} = 0.472; t_{20} = -0.474; t_{20} = 1.399 respectively). It was also revealed that there was no statistical significant difference between the item discrimination parameter estimates of sample sizes of 200, 400, 600, 800 and 1000 by CTT and IRT (t_5 = -0.390; t_4 = -0.919; t_6 = -1.959; t_5 = -1.397 and t_5 = -1.327 respectively). This shows that the items selected based on the sample sizes are comparable except for small sample size of 200 for item difficulty of CTT and IRT.

Discussion of Findings

Research question one reveals that items selected based on the sample sizes differ for CTT, while research question two reveals that items selected based on the sample sizes do not differ for IRT except for the sample size of 200 for item discrimination value. It was also found in the study that many items were selected when IRT method was used and this is because IRT approach is sample independent unlike in CTT, where few items were selected and this is as a result of CTT's dependency on sample size. This means that a major limitation of item difficulty and item discrimination parameter estimates under CTT framework is that they are sample dependent. These findings are in agreement with the finding of Ojerinde (2013) who evaluated the comparability of item analysis results of UTME Physics pre-test for classical test theory versus item response theory and found out that the total number of items rejected on the basis of item discrimination index was 19 for classical approach while only 12 were rejected using item response theory model. He also found that 12 items were rejected by CTT approach on the basis of item difficulty by item response theory method. Moreover, Adegoke (2013) who examined who examined the comparability of item statistics found out that in item selection process, IRT 2-parameter model led to deletion of fewer items than CTT model.

It was also found in the study that for the fairly large sample sizes (N =400 to N =1000) used in the study, the CTT-based and IRT-based item difficulty and item discrimination parameter estimates were very comparable. These findings are in agreement with the finding of Magno (2009) who compared the difference between CTT and IRT approach across two samples and test forms in Chemistry and found out that IRT estimates do not change across samples as compared with CTT with inconsistencies and IRT had significantly less measurement errors than the CTT approach. Adedoyin, Nenty and Chilisa (2008) investigated the invariance of item difficulty parameter estimates based on CTT and IRT for varying sample sizes and found out that the item difficulty parameter estimates for IRT were invariant across groups with varying sample sizes. Moreover, Kiany and Jalali (2009) investigated the theoretical and practical comparison of CTT and IRT and found out that item difficulty indexes from CTT were comparable with those from all IRT models and item discrimination indexes from CTT were somewhat less comparable with those from IRT.

Conclusion

From the results of the test items' parameter estimates of CTT and IRT for different sample sizes, it was evident that there was a significant difference between item difficulty parameter of CTT and IRT for sample size of 200. It was also evident that there was no significant difference between CTT and IRT item parameter estimates for sample sizes of 400, 600, 800 and 1000. It can be concluded that both the item parameter estimates of CTT and IRT could be used independently to estimate the test item parameter for different sample sizes except for 200 (small sample size), which revealed that the parameter estimates of the two measurement frameworks were comparable.

Recommendation

Experts should share free IRT software that can be readily used by specialists so as to get consistency in theories.

References

[1] Greg P. Item analysis analytics. Question mark. 2009; 2-19.

[2] Adedoyin OO, Adedoyin JA. Assessing the comparability between classical test theory and item response theory models in estimating test item parameters. Herald Journal of Education and General Studies. 2013; 2(3): 107-114.

[3] He Q, Wheadon C. The effect of sample size on item parameter estimation for the partial credit model. Centre for Education and Research Policy. 2012 Retrieved on from www.cerp. org.uk [Accessed 30th October, 2019].

[4] De Mars C. Sample size and the recovery of nominal response model item parameter. Applied Psychological Measurement. 2003; 27, 275-288.

[5] Hulin CL, Lissak RI, Drasgow F. Recovery of two and three parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement. 1982; 6(3): 249-260.

[6] Baruch N. Item analysis with small samples. Applied Psychological Measurement. 1980; 4(3): 323-329.

[7] Hula WD, Fergadiotis G, Martin N. Model choice and sample size in item response theory analysis of aphasia tests. American Journal of Speech Language Pathol. 2012; 21(2): 38-50.

[8] MacDonald P, Paunonen S. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test

theory. Educational and Psychological Measurement. 2002; 62, 921-943.

[9] Courville TG. An empirical comparison of item response theory and classical test theory item/response statistics. Ph.D Dissertation, Texas: Texas A and M University; 2005.

[10] Ojerinde D. Classical test theory (CTT) versus item response theory (IRT): An evaluation of the comparability of item analysis results. A lecture presentation at the Institute of Education, University of Ibadan, Nigeria; 2013.

[11] Adegoke BA. Comparison of item statistics of physics achievement test using classical test and item response theory framework. Journal of Education Practice. 2013; 4(22): 87-96.

[12] Magno C. Demonstrating the difference between classical test theory and item response theory using derived test data. The International Journal of Educational and Psychological Assessment. 2009; 1(1): 1-11.

[13] Adedoyin OO, Nenty HJ, Chilisa B. Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. Educational Research and Review. 2008; 3(2): 83-93.

[14] Kiany GR, Jalali S. Theoretical and practical comparison pf classical test theory and item response theory. International Journal of Active Learning. 2009; 12(1): 167-197.

Citation: *Chinelo Blessing Oribhabor, Judith Hannah Osarumwense, "Comparison of the Selection of Items Using Classical Test Theory and Item Response Theory Based on Sample Sizes", Journal of Educational System, 3(2), 2019, pp. 31-41.*