East African Scholars Journal of Psychology and Behavioural Sciences
Abbreviated Key Title: EAS J Psychol Behav Sci
ISSN 2663-1865 (Print) | ISSN 2663-6751 (Online)
Published By East African Scholars Publisher, Kenya
Volume-6 | Issue-2 | Mar-Apr-2024 | DOI:10.36349/easjpbs.2024.v06i02.002

OPEN ACCESS

Original Research Article

Item Statistics of Multiple Choice Physics Achievement Test Using Classic Test Theory and Item Response Theory

Fagbenro W. Ayoola[1*], Abdullahi Ibrahim[1]
[1]Department of Science Education, Federal University Wukari, Wukari

Article History
Received: 06.02.2024
Accepted: 19.03.2024
Published: 24.03.2024

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: This study examine the comparability of item statistics generated from the frameworks of classical test theory (CTT) and 2-parameter model of item response theory (IRT). A 40-item Physics Achievement Test was developed and administered to 600 senior secondary school two students, who were randomly selected from 12 senior secondary schools in Taraba State, Nigeria. Results showed that item statistics obtained from both frameworks were relatively similar. However, item statistics obtained from IRT 2-parameter model looked balanced than those from CTT. In addition, for item selection process, IRT 2-parameter model retained more items than CTT model. This result implies that test developers and public examining bodies should integrate IRT model into their test development processes. Through IRT model, test constructors would be able to generate stable items than in the CTT model used at present and at the end, the test scores of examinees will be more reliably estimated.

Keywords: Item Statistics, Item Analysis, Item Response Theory, Classical Test Theory, Physics Achievement Test.

INTRODUCTION

At the background of every educational measurement are testing and assessment of abilities of learners, often referred to as measures of ability. Educational measures could focus either on measure of aptitude or measure of achievement. These tests attempt to measure one or several hypothetical constructs that are typically unobservable, known as latent trait. According to Baker and Kim, 2004, examples of latent traits include intelligence and arithmetic ability. It is very essential that tests demonstrate consistency in measurement when measuring latent traits, and this is called reliability. Practically, reliability relates to test consistency and exists either as test-retest consistency, parallel or alternate form reliability or internal consistency. Theoretically, reliability relates to the proportion of the total variance in an obtained score that is due to true variance as opposed to error variance (Cohen & Swerdlik, 2010; Schmidt & Embretson, 2013). Internal consistency has to do with the degree to which items within a test are correlated with each other, either by comparing each item with every other item on a test or scale (inter-item correlation, item total correlation, Cronbach's alpha) or by comparing two halves of a single test (split-half reliability). Parallel and alternate-form reliability estimates involve comparing two parallel or alternate test forms to determine the degree to which they correlate with each other, using their means and variances. Test-retest consistency is the degree to which pairs of scores from the same people on two different administrations of the same test are correlated. It is very important to evaluate the reliability of an instrument to determining whether a measure is psychometrically sound.

The reliability of a test impacts the standard error of measurement (SEM). Standard error of measurement is an estimate of the variability expected for observed scores when the true score is held constant (Dudek, 1979). The variability in observed scores occurs because few tests are perfectly reliable. The larger the SEM, the lower the reliability. Therefore, much effort is taken to minimize error because it results in a large SEM, which leads to lower test precision and questionable test validity. However, SEM is typically estimated based on internal consistency reliability (Slick, 2004). The SEM is used to generate confidence intervals, which is the range of scores that is likely to contain the true score (Cohen & Swerdlik, 2010). Apart from the SEM, which is tied to the inherent imperfections of a given test, there are other sources of error in obtained test scores. These include examinee factors (e.g., fatigue, lack of motivation, reactivity to the testing situation, and guessing),

***Corresponding Author:** Fagbenro W. Ayoola
Department of Science Education, Federal University Wukari, Wukari

examiner factors (e.g., test administration and rapport-building skills, scoring and interpretation mistakes), and environmental factors (e.g., room noise level, lighting, and temperature). A physics achievement test administered to an unmotivated examinee in a noisy room is unlikely to provide an accurate reflection of the examinee's ability. It is assumed that the examinee is putting forth sufficient effort in an environment where extraneous variables (i.e., those not pertaining to physics) are minimized. Such errors could decrease the reliability of a test and result in test bias. Differential item functioning (DIF) is another form of bias in test scores, whereby the probability of endorsing an item is higher for one group than the other, across various trait levels (Swaminathan & Rogers, 1990). In other words, despite two people from different groups having the same latent trait level, they have a different probability of obtaining a correct score on a given item. Closely related to the issue of test bias and DIF is the concept of measurement equivalence, which occurs when there are identical associations between observed test scores and latent trait across different populations (Drasgow, 1984). It is important to evaluate tests for DIF and measurement equivalence to ensure that test findings are accurately interpreted across different samples.

Classical Test Theory

Concepts such as the SEM and reliability estimates are associated with classical test theory (CTT), which posits a set of principles to evaluate the degree to which tests are successful at estimating unobservable variables of interests (DeVellis, 2006; Gulliksen, 1950; Lord & Novick, 1968). In CTT, an observed test score and an observed score variance are functions of a true score and an error score, as well as true score variance and error variance (Spearman, 1907; 1913). CTT is based on a number of well documented assumptions (Schmidt & Embretson, 2013; Zickar & Broadfoot, 2008) guiding its operations. The first assumption is that true scores and error scores are uncorrelated, given that errors are random. Second, a normal distribution of errors can be expected, given that errors are random and due to a combination of several factors. Therefore, the average error score is zero for each examinee in the population and across replications. Third, error scores are not correlated with scores on parallel tests or other test scores. Although CTT acknowledges measurement error, it does not generally allow for different degrees of measurement error for different ability levels (Schmidt & Embretson, 2013). The CTT equation does not take into consideration the content or characteristics of a test item, but the theory references item relationships with other variables. Therefore, equivalent parallel forms of special test equating methods are required when a trait or construct is measured by more than one test. Strictly parallel forms have equal means, variances, and correlations with other variables. Item properties such as item difficulty and discrimination parameters have to be matched across forms. Otherwise, the true score would be dependent on particular sets of items included on a test (Schmidt & Embretson, 2013), whereby a high score would be obtained on a test with easier items, and a low score would be obtained on a test with more difficult items.

However, there are a number of limitations inherent in this theory. First, in CTT, examinee characteristics and test characteristics cannot be separated; each can only be interpreted in the context of the other. This limitation influences test precision in a number of ways. Reliability estimates vary as a function of method used and sample on which they were computed. When a test is "difficult," an examinee will appear to have low ability whereas when the test is "easy," the examinee will appear to have higher ability. In addition, item difficulty is defined in CTT as the proportion of examinees who answer an item correctly (Weiss, 1995). Therefore, item and overall test difficulty depends on the sample of examinees being measured; at the same time, examinee ability estimates depend on the degree to which the test was difficult. Consequently, a test constructed on one group of people with a given trait level cannot be used directly on a different group with a different trait level (Hambleton, Swaminathan, & Rogers, 1991; Schmidt & Embretson, 2013; Weiss, 1995). Second, CTT computes SEM as a constant value. However, a single SEM is not an accurate reflection of a scale. Given that reliability for a given test varies depending on the group with which the individual is measured, different SEMs can be attached to an obtained score. This leads to an illogical situation whereby a person with the same score and responses is evaluated with different "precision," depending on the group to which he/she is compared. Third, given that CTT uses a number-correct score, test scores depend on the difficulty of items included in a test. More difficult tests result in lower average scores, and easier tests result in higher scores. Further, because difficulty levels depend on the sample, this number-correct score is dependent on the group on which the test is normed. The number-correct score artificially limits the number of levels of observed score at which a person can be assessed. For example, a test with 30 items allows only 31 possible scores. Similarly, each item in a scale is given equal weight, regardless of item difficulty, so that a correct answer to an easy item is worth as much as a correct answer to a difficult item. Fourth, item selection procedures advocated in CTT focuses on maximizing reliability. Items are usually selected with .50 difficulty (i.e., 50% of respondents answers correctly), and further item analyses usually results in deleting items with low inter-item correlations. This results in a test with relatively equal difficulty levels and items that are highly discriminating. Such tests are effective at discriminating between upper and lower halves of population but ineffective in discriminating examinees at other levels of traits (e.g., those in the lower 10% of population). Fifth, item parameters are regarded as fixed on a particular test in CTT. The CTT equation does not include test item characteristics or content, even though the theory

underlying the CTT equation refers to item relationships with other variables. Therefore, in order to generalize a true score to other variables or tests, and to make score comparisons, it is necessary for these other variables or tests to have parallel items and it is necessary to use special test equating methods (Schmidt & Embretson, 2013).

Item Response Theory

Given the inherent limitations of CTT, it is not surprising that an alternative test development theory has been proposed. Modern psychometric theory is also known as item response theory (IRT) or latent trait theory and has been defined as a model-based measurement in which trait level estimates depend on both the test-taker's responses and the properties of the items that were administered (Embretson & Reise, 2000). Fundamental to IRT is the concept of a link between item responses and the trait (known as theta, θ) measured by the scale (Drasgow & Hulin, 1990). IRT identifies item parameters like item difficulty, item discrimination, and guessing. Item difficulty, denoted by b or β, is defined as "the point along the θ continuum where individuals have a fifty percent chance of a positive response" (Drasgow & Hulin, 1990, p. 582). Item discrimination, denoted by a or α, describes how well an item can differentiate between examinees having abilities below an item location and examinees with abilities above the item location. Item discrimination is defined by the steepness of the item characteristic curve (Drasgow & Hulin, 1990). The guessing parameter, denoted by c or γ, takes into account instances where people with low θ occasionally endorse an item. The item characteristic curve graphs the relationship between changes in trait level and changes in the probability of a specified response (Cohen & Swerdlik, 2010; Hambleton, Swaminathan, & Rogers, 1991; Holland, 1990). The smaller the slope, the less discriminating the item is, because the item response probabilities (on y-axis) are relatively less responsive to changes in trait level (Embretson & Reise, 2000). The IRT framework encompasses a group of models. For test items that are dichotomously scored, there are three traditional IRT models, known as 3-, 2- and 1- parameter IRT models. The proposed 4-parameter logistic model which incorporates response time and slowness parameter (Wang and Hanson 2001) has not really been formally incorporated into the traditional IRT models. The many benefits of IRT models provide compelling justifications for the use of such models in creating, evaluating, and applying psychological tests (Embretson & Reise, 2000). IRT incorporates techniques for evaluating the applicability of a given test across different subgroups.

IRT or CTT

A number of studies have investigated differences between scores derived through CTT and IRT. Tinsley and Dawis (1977) found that unlike CTT, IRT yielded person ability estimates that were independent of the test item difficulty levels. In other words, when students were administered two tests, their ability was estimated to be higher on the easier test and lower on the more difficult test using CTT-based raw scores. However, students' ability estimates remained relatively constant on both the easier and more difficult tests when derived using IRT-based methods. In their Monte Carlo simulation study of test items and examinees, MacDonald and Paunonen (2002) found that IRT and CTT accurately estimated ability levels of participants (IRT-based ability parameter θ and CTT-based person test score) and test item difficulty (IRT-based β parameter and CTT-based item difficulty P value). However, the IRT-based discrimination parameter, α, demonstrated more consistently accurate estimates than did the CTT-based item discrimination index, particularly in simulated conditions with a large range of item difficulty statistics. However, Ojerinde (2013) observed that these demands are not sufficient enough to abandon CTT for IRT if a very large sample size (i.e. N > 500 testees) is used in estimation of the item parameter. The thrust of this study is to examine how close the item parameters will be if large sample size of testees is used in estimating the item parameters.

Statement of the Problem

Measurement of students' scientific reasoning and knowledge processes is a complex job, yet vital to efficient teaching and learning (National Research Council 2001; 2007). Measurement of students' knowledge is affected by a variety of factors, which has raised the attempt to comprehend how item features and test theories affect inferences about student understanding, as well as the advancement of new tools and procedures to measure valid student achievement (NRC 2001, 2007; Gitomer & Duschl, 2007). For an objective measurement and strive to maintain common metric, the property of invariance of person and item characteristics is very critical, however Ojerinde (2013) observed that the demands is not sufficient enough to abandon CTT for IRT if a very large sample size (i.e. N > 500 testees) is used in estimation of the item parameter. If a very large sample size is used in the estimation of the item parameter, will there be any rationalization for abandoning CTT for IRT?

Research Questions

The following research questions were addressed.

1. What are the item statistics of the 40-item PAT using CTT model and 2-parameter model of IRT?
2. How many items are retained after the item analysis using the CTT model and 2-parameter model of IRT?
3. To what extent are the CTT-based and IRT-based item discrimination estimates comparable?
4. To what extent are the CTT-based and IRT-based item difficulty estimates comparable?

METHODS

The sample comprised 600 senior secondary II Physics students ($male = 378\ and\ female = 222$) drawn from population of Physics students in Taraba State. Multi stage sampling technique was employed to randomly select twelve schools from three Local Government Areas of the three Senatorial Districts of Taraba State, and an intact arm of SS II from each of the sampled schools was used.

Table 1: Sample Frame for Effect of Item Sequence on Physics Achievement

S/N	Selected Local Government Area	No of School	Sample of SSII	Sample of SSII by Sex	
				Male	Female
1	Wukari	4	204	133	71
2	Jalingo	4	243	159	84
3	Bali	4	153	86	67
Total		12	600	378	222

The instrument employed for data collection in this study is Physics Achievement Test. The initial draft of PAT consisted of 60 items. It was developed by the researcher. However, the physics syllabus prepared for SSCE by WAEC and NECO, as well as the Physics curriculum prepared by the Federal Ministry of Education, Abuja, Nigeria was taken used as guide. The researcher used content of the physics syllabus and physics curriculum for senior secondary one and senior secondary two in developing the items. Each item was placed on four-option response mode of A, B, C, and D as used by WAEC and NECO. The theme covered by the item is electricity.

Items were developed from electricity analysis of the scheme of work in all the schools that were sampled showed that all the physics teachers had taught electricity. Also, as noted by the Physics Chief Examiners' (WAEC, 2012), many candidates have difficulties in the use of equations and formulas in test items and there are many formula and equations which the students have to master in electricity.

Test blue print, illustrated in table 2, was developed to ensure the content validity of the test. The thought processes were limited to knowledge, comprehension and application because of the age of the students, reduction of tedium, and inability of the multiple choice objective question to accommodate learning outcome such as ability to; articulate explanations, display thought processes, furnish information, organize personal thought, perform a specific task, produce original ideas and produce examples.

Table 2: Table of Specification

S/N	Content	Knowledge	Comprehension	Application	Total
1	Electric charge	2	2		4
2	Current in a simple circuit		1	4	5
3	Potential difference	1	2	2	5
4	Resistance	2	2	2	6
5	Series circuit	1	1	3	5
6	Parallel circuit	1	1	3	5
7	Electric power	2	1	2	5
8	Electric energy	1	2	2	5
	Total	10	12	18	40

Also, opinions of panel of qualified experts in Physics and Education Evaluation were sought in deciding the appropriateness of the items to give logical validity index of 0.77.

The final draft copy of PAT consisting of 40 items was administered to the 600 students. The test was administered immediately after the school official hour. The researcher was assisted by four research assistants. The time allowed for the students to take the test was 60 minutes. On the average, it took the students about 50 minutes to finish the test. The items were marked, 1 is given for correct answer and zero for wrong answer.

The IRT and CTT frameworks were used in carrying out the item analysis to select the final items. The BILOG-MG (Window Version 3.0) with Marginal-Maximum Likelihood Method was used.

Research Question 1: What are the item statistics of the 40-item PAT using CTT model and 2-parameter model of IRT?

As illustrated in table 3, the item statistics using CTT framework gives the classical item statistics of difficulty (p) and discrimination (r) on the left while the item statistics using IRT frameworks gives the discrimination (a) and difficulty (b) on the right.

Table 3: Item statistics using CTT and IRT Framework

Analysis using CCT			**Analysis using IRT**	
Item no	p	r	a	b
1	0.69	0.34	0.41	-1.24
2	0.49	0.29	0.33	0.05
3	0.27	0.18	0.23	-0.38
4	0.57	0.21	0.27	-0.65
5	0.25	0.15	0.22	2.19
6	0.45	0.59	0.76	0.21
7	0.40	0.35	0.42	0.67
8	0.76	0.35	0.12	-1.80
9	0.75	0.35	0.19	-1.72
10	0.54	0.47	0.52	-0.02
11	0.26	0.18	0.31	2.23
12	0.56	0.21	0.25	-0.62
13	0.46	0.55	0.70	0.17
14	0.16	0.11	0.53	2.22
15	0.30	0.28	0.30	1.80
16	0.27	0.19	0.66	1.05
17	0.51	0.46	0.56	0.01
18	0.19	0.09	0.18	-1.10
19	0.29	0.21	0.39	1.54
20	0.45	0.62	0.80	0.25
21	0.38	0.41	0.46	0.75
22	0.30	0.43	0.52	1.17
23	0.23	-0.04	0.16	0.66
24	0.36	0.23	0.26	1.44
25	0.67	0.31	0.41	-1.06
26	0.24	0.17	0.25	2.73
27	0.37	0.24	0.28	1.29
28	0.47	0.24	0.30	0.30
29	0.28	0.12	0.31	1.91
30	0.21	0.23	0.31	2.72
31	0.34	0.32	0.35	1.32
32	0.26	0.19	0.39	1.76
33	0.37	0.30	0.35	1.08
34	0.41	0.27	0.31	0.82
35	0.21	0.23	0.28	3.03
36	0.19	0.18	0.36	2.53
37	0.48	0.48	0.60	0.16
38	0.17	0.19	0.26	4.02
39	0.30	0.06	0.14	3.44
40	0.32	0.14	0.21	0.26

Research Question 2: How many items are retained after the item analysis using the CTT model and 2-parameter model of IRT?

On the basis of the criteria set for the difficulty indices $(0.30 > p < 0.70)$ using classical test theory framework, items which failed to satisfy the conditions were: 3, 5, 11, 14, 16, 18, 19, 23, 26, 29, 30, 32, 35, 36, 38, and 39. Also for the discriminating index (r), the condition set is ($r \geq 0.20$), items 3, 5, 11, 14, 16, 18, 23, 26, 29, 32, 36, 38, 39 and 40 were considered poor. Therefore, using the criteria set for difficulty and discriminating indices, 17 items: items 3, 5, 11, 14, 16, 18, 19, 23, 26, 29, 30, 32, 35, 36, 38, 39 and 40 were deleted. The reliability index of the 40 items under CTT was 0.75.

In IRT framework, the selection of items is a function of the information each of the items contributes to the overall information supplied by the whole test. To effectively do this, there is a need to look at the information function in figure 1. The solid line for the 40 items PAT information function gives the total information while the dotted line gives the standard error for a specific ability.

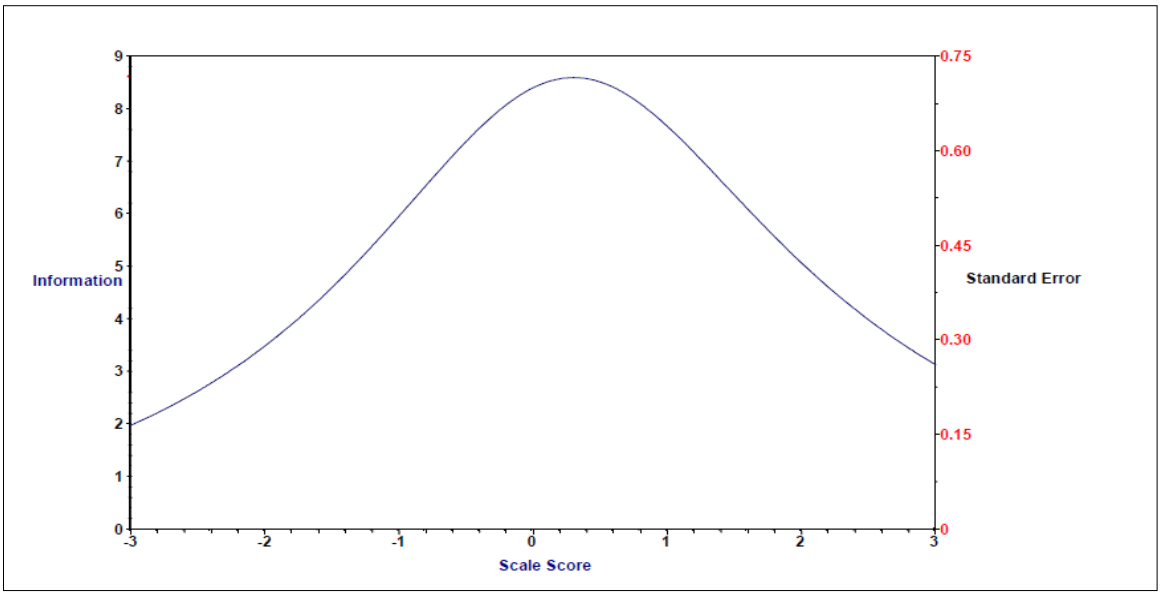

Figure 1: Test information function

On the test information function, the maximum amount of test information was 8.6 at an ability level of 0.25. Using the test information function, items whose difficulty level fall between -1.45 and 2.25 should be included in the final test. Base on this decision, items 8, 9, 38, 39 and 40 were deleted. Also, items with low discriminating indices (a) ($a \geq 0.2$) were deleted. Items 8, 9, 18, 23, 38, 39 and 40 were in this category. Using test information function and discriminating index, 7 items were deleted. These were items 8, 9, 18, 23, 38, 39 and 40. The reliability index of the 40 items under IRT was 0.77.

The item statistics using CTT leads to a final draft of PAT with 23 items because 17 items were deleted, while the item statistics using IRT framework leads to a final draft of PAT of 33 items because 7 items were deleted. Common items deleted by both framework are; 18, 23, 38, 39, and 40.

Research Question 3: To what extent are the CTT-based and IRT-based item discrimination estimates comparable?

Figure 2 illustrated a scatter plot of *a* - values estimated from the 2-parameter model of IRT and the *r* - values of the point biserial correlations of CTT. The correlation coefficient of the relationship between the two parameters was determined. There is a high positive correlation between the *a*-value and the values of the point biserial correlation *r*. The value is 0.737.

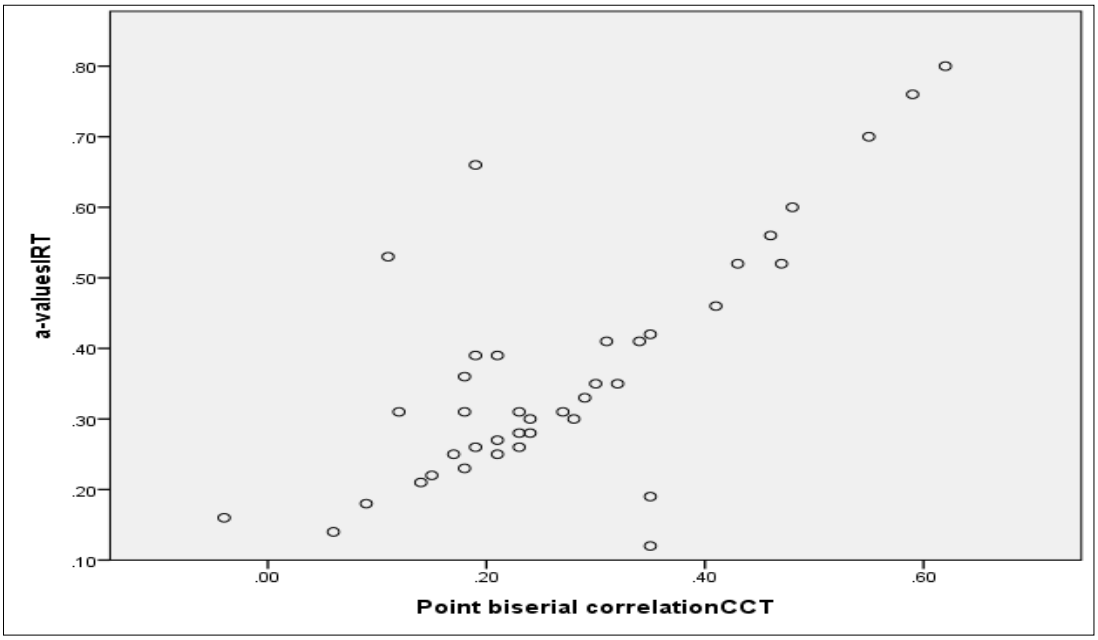

Figure 2: Scatter plots of the relationship between *a* and *r*

And is statistically significant ($p < .001$). These findings show that a high correspondence exists between the two indices. This finding shows that CTT–based discrimination index is comparable with the IRT–based discrimination parameter. The result of this study is in agreement with Wiberg, 2004, which concluded that the correlation coefficient of the relationship between *a* – values and point biserial correlation should be high and positive

Research Question 4: To what extent are the CTT-based and IRT-based item difficulty estimates comparable?

Fugure 3 shows a scatter plot of *b* - values estimated from the 2-parameter model of IRT and the *p* - values of CTT. The correlation coefficient of the relationship between the two parameters was determined.

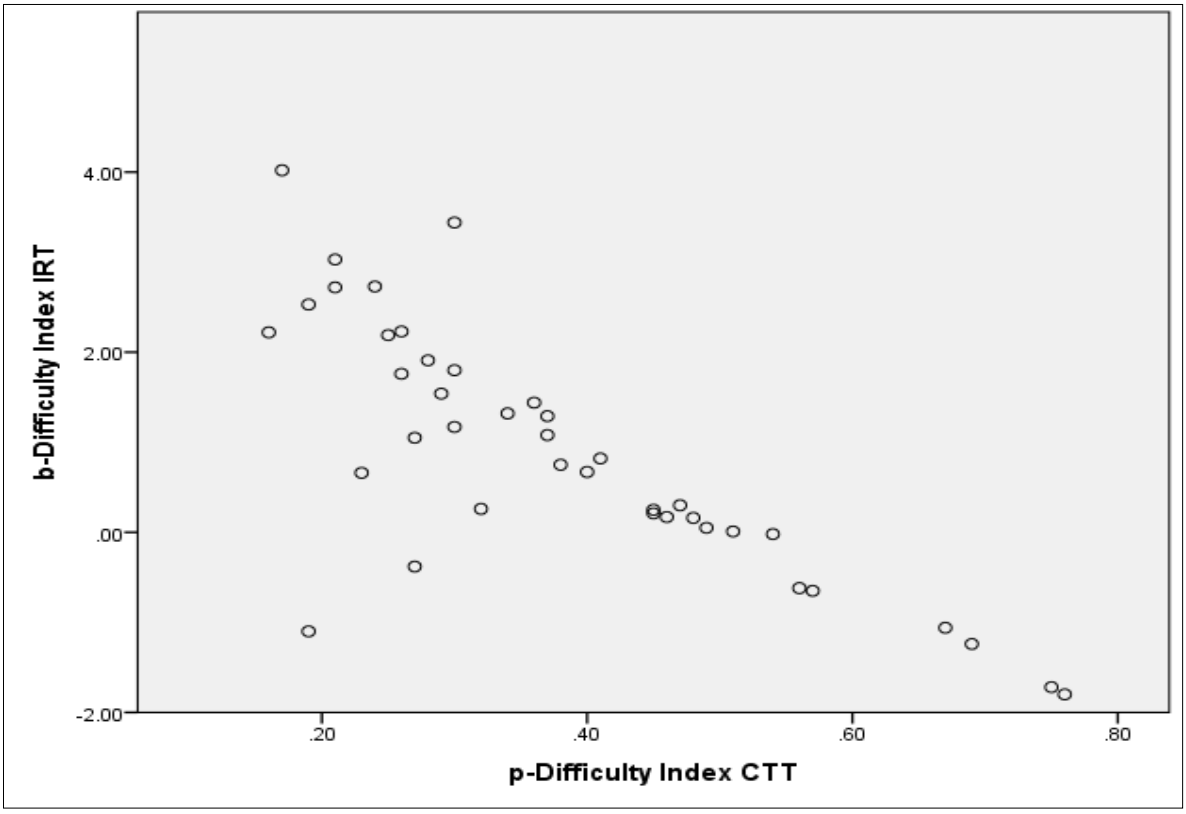

Figure 3

The correlation coefficient between *p* – values of CTT and b – parameter of IRT is high and negative. The value is - 0.794 and is statistically significant ($p < 0.001$). This shows that as the value of *pi* increases, *bi* decreases. The result is in agreement with the result of past studies such as Wiberg (2004) and Stages (2003).

Discussion and Implication of Findings

The major findings of this study is based on the fairly large sample used, the CTT-based and IRT based item statistics estimates were very comparable. These findings were consistent with the earlier studies (e.g. Bechger *et al*., 2003; Adegoke, 2013; Ojerinde, 2013; Stage, 2003, Wiberg, 2004). However, using CTT-based item statistics estimates, more items were deleted from the 60-item PAT than when IRT-based item statistics estimates were used. This finding agreed with observation of test experts such as Hambleton and Jones (1993), Ojerinde (2013).

The result of this study aligned with results of some past studies (e.g. Bechger, Gunter, Huub & Beguin, 2003; Adegoke, 2013; MacDonald & Paunoen, 2002; Stage, 2003) which showed that IRT model has little or no superiority over CTT models in item parameters estimates.

However, because the overall, item statistics from both IRT and CTT frameworks are comparable in some cases, the author recommends that Examining bodies should integrate IRT models into their test development processes. CTT framework could be used as a complement to IRT.

References

- 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.
- American Educational Research Association, American Psychological Association, & National

Council on Measurement in Education. (AERA; 1999). *Standard for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

- Baker, F. N., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques.* Second Edition. New York, NY: Marcel Dekker.
- Bechger, T. M., Maris, G., Verstralen, H. H., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied psychological measurement*, *27*(5), 319-334.
- Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement.* New York, NY: McGraw-Hill.
- DeVellis, R. F. (2006). Classical test theory. *Medical care*, *44*(11), S50-S59.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134-135.
- Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology, Vol. I* (pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.
- Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335-337.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gitomer, D. H., & Duschl, R. A. (2007). Establishing multilevel coherence in assessment. In: Moss, P.A. (Ed.). *Evidence and decision making*. The 106th yearbook of the National Society for the Study of Education, Chicago, 288-320.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications, Inc.
- Hambleton, R., & Jones, R. (1993).Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38 -47.
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577-601.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Weasley.
- MacDonald, P. & Paunonen, S. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 921 – 943.
- Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 321-943.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press, Washington, D.C.
- National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. National Academy Press, Washington, D.C.
- Ojerinde 'Dibu (2013). *Classical test theory (CTT) VS item response theory (IRT): An evaluation of the comparability of item analysis results*. A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May.
- Schmidt, K. M., & Embretson, S. E. (2013). Item response theory and measuring abilities. In J. A. Schinka and W. F. Velicer (Eds.), Research Methods in Psychology (2nd ed.). Volume 2 of Handbook of Psychology (I. B. Weiner, Editor-in-Chief).
- Slick, D. J. (2004). Psychometrics in neuropsychological assessment. In E. Strauss, E. M. S. Sherman, & O. Spreen (Eds.), *A Compendium of Neuropsychological Tests:*
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 161-169.
- Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, 5(4), 417-426.
- Stage, C. (2003). Classical test theory or item response theory: The Swedish experience (No. 42). Umea: *Kluwer Journal of Education and Practice*, *4*(22). www.iiste.org
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning usinglogistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Tinsley, H. E. A., & Dawis, R. V. (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement, 1,* 483-487.
- Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. J. Lubinski and R. V. Dawis (Eds.), Assessing individual differences in human behavior: New concepts, methods, and findings (pp. 49-79). Palo Alto, CA: Davies-Black Publication.
- Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test* (No. 50). Umea: Kluwer Academic Publications
- Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance, & R. J. Vandenberg. (Eds.). Statistical and methodological myths and urban legends: Doctrine, *verity and fable in the organizational and social sciences* (pp. 37-59). New York, NY: Routledge/Taylor & Francis Group.

Cite This Article: Fagbenro W. Ayoola & Abdullahi Ibrahim (2024). Item Statistics of Multiple Choice Physics Achievement Test Using Classic Test Theory and Item Response Theory. *EAS J Psychol Behav Sci*, *6*(2), 11-18.