Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.4, No.22, 2013

Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks

Benson Adesina Adegoke PhD
Institute of Education, University of Ibadan, Nigeria
E-mail of the corresponding author: doctoradegoke@yahoo.com

Abstract
In this study, the author examined the comparability of item statistics generated from the frameworks of classical test theory (CTT) and 2-parameter model of item response theory (IRT). A 60-item Physics Achievement Test was developed and administered to 724 senior secondary school two students (Age 16-18 years), who were randomly selected from 16 senior secondary schools in Ibadan Educational Zone I, Oyo State, Nigeria. Results showed that item statistics obtained from both frameworks were quite comparable. However, item statistics obtained from IRT 2-parameter model appeared more stable than those from CTT. Moreover for item selection process, IRT 2-parameter model led to deletion of fewer 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 more reliable items than in the CTT model being currently used and ultimately 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

The persistent students' low level of achievement and enrolment in physics continues to attract the attention of major stakeholders in science education. Researchers in physics education have also tried to isolate causes of this low level of achievement and proffered ways of improving students' performance in physics. For example, Adegoke (2011) suggested the use of multimedia instruction; Azar and Şengülec (2011) suggested the use simulated experiments in teaching physics practical. Despite all the suggestions of researchers in physics education, very little improvement in achievement and students enrolment has been recorded. However, little attention has been paid, by researchers in physics education, to the in-depth analysis of the items contained in the test papers especially the objective tests. A thorough study of the process in which the test items were developed as well as their psychometric properties may suggest ways of improving students' achievement in physics.

In educational measurements, there are two main frameworks by which a test and the items it contains can be studied. These are Classical Test Theory (CTT) and Item Response Theory (IRT).

Classical Test Theory

This theory tries to explain the link between the observed score, the true score, and the error score. Within that theoretical framework, models of various forms have been formulated. The most common model is known as "classical test model". It is a simple linear model linking the observable test score (X) to the sum of two unobservable (or often called latent) variables, that is, true score (T) and error score (E).

$$X = T + E$$

There are two unknowns in the equation (X and E) and this makes it not easily solvable unless some simplifying assumptions are made.

The assumptions in the classical test model are: (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel test are uncorrelated.

Although the major focus of classical test theory is on test-level information, items statistics that represent *item difficulty* (often denoted *p*) and *item discrimination* (often denoted *D* or *r*), have been developed and are also important parts of the model. These two major item statistics are used in item analysis and item selection in the development of achievement tests.

Item Difficulty

The success rate of a pool of examinees on an item is used as the index for the *item difficulty*. Symbolically, it is given as

$$\text{Item difficulty index } (p) = \frac{Number\ of\ students\ who\ got\ the\ item\ right}{Total\ number\ of\ students\ who\ tried\ it}$$

Item Discrimination

The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as *item discrimination*. There are several methods being used in CTT to assess *item discrimination*. These include: (a) finding the difference in the proportion of high achieving and low achieving students who score the item correctly and (b) biserial correlation or point-biserial correlation between a dichotomously scored item and the scores on the total test.

The use of the difference between the proportion of high achieving examinees that scored the item correctly and the proportion of low achieving examinees that scored the item correctly necessitates splitting the examinees into two groups. However, instead of splitting the groups into 50-50, usually the 27 % of the two contrasting groups are (Kelly, 1939; Courville, 2004). Mathematically, Item Discrimination index (D) is given as $(p_u - p_l)$, with p_u being the proportion of correct responses for the upper group and p_l, being the proportion of correct responses for the lower group. Its values range from -1 to +1. A positive index indicates that a higher proportion of the upper group answered the item correctly, while a negative item discrimination index indicates that a larger proportion of the lower group answered the item correctly. Although this method gives a fairly stable index of discrimination, it is problematic in that the process of its computation ignores so much data. In fact, it omits the data of a lot of people (e.g. 46% of the examinees) and ignores information regarding the exact scores in the high achieving group and in the low achieving group.

To correct the problem, the point biserial correlation, $r_{pb.j}$ for item *j*, a computationally simplified Pearson's *r* between the dichotomously scored item *j* and the total score *x* has been advocated (see Hambleton, & Jones 1993).

It is computed as

$$r_{pb.j} = \frac{(\mu_j - \mu_x)}{\sigma_x}\sqrt{\frac{p_j}{q_j}}$$

Where

μ_j is the mean total score among examinees who have responded correctly to item *j*, μ_x is the mean total score for all examinees, p_j is the item difficulty index for item *j*, $q_j = (1 - p_j)$ and σ_x is the standard deviation of the examinees' total score. This method is better than the latter in that in its computation, scores of the examinees are used.

Item Analysis and Item Selection

Item analysis within the framework of classical test theory consists of (a) determining sample-specific parameters and (b) deleting items based on the statistical criteria set. A poor item is identified by an item difficulty value that is too high (p> 0.70) or too low (p < 0.30), or a low item discrimination, such that$r_{pb.j}$≤ 0.20. In test development, generally items are selected on the basis of these two characteristics. However, the choice of level of difficulty and discrimination is usually governed by the purpose of the test and the anticipated ability distribution of the group for whom the test is intended. For example, norm-referenced test are developed to differentiate between examinees with regard to their competence in physics. That is, such test is designed to yield a broad range of scores maximising discriminations among all examinees taking the test. Therefore, items are usually chosen to have a medium level and narrow range of difficulty.

Despite the popularity of classical item statistics as an integral part of standardised test and measurement technology, it is fraught with so many limitations (Hambleton & Jones, 1993; Ojerinde, 2013). According to Hambleton and Jones (1993) the major limitations of CTT are: (a) the person statistics (i.e., observed score) is (item) sample dependent, and (b) item statistics (i.e., item difficulty and item discrimination) are examinee sample dependent. Therefore, the estimates of CTT are not generalizable across populations. An awareness of the shortcomings of CTT and the potential benefits offered by IRT has led some examining bodies in Nigeria, for example Joint Admissions and Matriculation Board (Ojerinde, 2013) to start working within the IRT framework.

Item Response Theory

Under item response theory, the primary interest is on the item-level information in contrast to the CTT's primary focus on test-level information. IRT is a modelling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance (Hambleton & Jones, 1993; Hambleton, Robin, & Xing, 2000).

The most commonly used IRT models are built off a single ability parameter (θ), which is very similar to the CTT total-test true score (*X*). In fact, the relationship between the observed score and the ability parameter is the same relationship as the observed score and true score. However, in contrast to classical test theory, item response models are lauded (Ojerinde, 2013; Wang and Hanson, 2001; Wiberg, 2004) for their ability to generate invariant estimates of item and person parameters. That is, theoretically, IRT ability estimates (θ) are "item – free" (i.e. would not change if different items were used) and the item difficulty statistics are person free (i.e. would not change if different persons were used).

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. Moreover, software for analysing it is not yet readily available.

3- parameter model

$$P(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

Where c_i= guessing parameter of the item *I*, a_i= discrimination parameter of item *i* commonly known as item slope, b_i= difficulty parameter of item *i* known as item location parameter, θ (Theta) = the ability level of a particular examinee

The parameter *c* is the probability of getting the item correct by guessing alone. By definition, the value of *c* does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. The parameter *c* has a theoretical range of 0 < = *c*< = 1.0, but in practice, values above 0.35 are not considered acceptable, hence the range $\theta \leq c \leq 0.35$ is usually adopted when the 3-parameter model is used.

The difficulty parameter also known as location parameter, denoted by *b*, is defined as the point on the ability scale at which the probability of correct response to the item is 0.5. The theoretical range of the values of this parameter is - ∞ < = *b*< = + ∞. However, typical values have the range - 3 < = *b*< = + 3.

The item discrimination parameter also known as slope parameter, denoted by *a*, is the slope of the tangent line of the item characteristics curve at the point of the location parameter. Typical values of *a* range - 3 < = *b*< = + 3. While most test items will discriminate in a positive manner (i.e., the probability of correct response increases as the ability level increases), some items have negative discrimination. In such items, the probability of correct response decrease as the ability level increases from low to high.

2-parameter model

When the guessing factor *c* is assumed or constrained to be zero the 3-parameter model is reduced to the 2-parameter model for which only item location and item slope parameters need to be estimated.

$$P(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

1-parameter model

If another restriction that stipulates that all items have equal and fixed discrimination, is imposed, then parameter *a* becomes a constant rather than a variable, and as such, it is not estimated and the IRT model is reduced to one-parameter model.

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b_i)}}$$

1-parameter model is also known as Rasch model, named after the researcher who did pioneer work in the area.

In practical situations, according to Baker (2001), it seems 2-parameter model is the best, at least, for norm-referenced tests. This is because in addition to the difficulty index of each item that can be obtained from the Rasch model, the test developer can also determine the discriminating power. This feat can also be performed using the 3-parameter model. However, as pointed out in the earlier section, the value of *c* i.e. the guessing parameter (3-parameter model) does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing (Baker, 2001).

As further noted by Baker, a significant side effect of using the guessing parameter *c* is that the definition of the difficulty parameter is changed. Under the 1-parameter and 2-parameter models, *b* is the point on the ability scale at which the probability of correct response is 0.5. But under the 3-parameter, the lower limit of item characteristic curve is the value of *c* rather than zero. The result is that the item difficulty parameter is the point on the ability scale where:

$$p(\theta) = c + (1 - c)(0.5)$$
$$= \frac{1 + c}{2}$$

According to Baker, this probability is half way between the value of *c* and 1.0. What has happened is that the parameter *c* has defined a floor to the lowest value of the probability of correct response. Thus, the difficulty parameter defines the point on the ability scale where the probability of correct response is halfway between this floor and 1.0.

Although the discrimination parameter ***(a)*** can still be interpreted as being proportional to the slope of the item characteristic curve at $\theta = b$, under the 3-parameter model, the slope of item characteristic curve at $\theta = b$ is actually a (1 – c)/4. As Baker emphasised, while the changes in the definitions of parameters *b* and *a* seem slight, they are important when interpreting the results of test analyses. In this study, therefore, the author's emphasis is on 2-parameter model.

Item Analysis and Item Selection

When employing item response theory, item analysis consists basically of (a) determining sample-invariant item parameters (difficulty index *b*, and discrimination index *a*) using relatively complex mathematical techniques

and large sample sizes; and (b) utilising goodness of fit criteria to detect items that do not fit the specified response model. Poor items are usually identified through a consideration of their discrimination indices (the value of a_i being a low positive or even negative) and difficulty indices (items should be neither too easy nor too difficult for the group of examinees being assessed). The difficulty index is however a function of the average ability of the examinees.

As is the case with CTT, selection of items under IRT models depends on the intended purposes of the test. However, the final selection of items will depend on the information each of the items contribute to the overall information supplied by the whole test. This is usually achieved through test information function.

Item Response Theory versus Classical Test Theory

The relationship between item difficulty and discrimination parameters in the CTT model and two-parameter logistic model were well discussed by Lord (1980). Specifically, Lord (1980) states that under certain conditions, such as examinee performance not being affected by guessing, item-test biserial correlation (*r*)used within the framework of CTT and item discrimination parameter (*a*) of IRT are approximately monotonically increasing function of each other. This relationship may be represented as:

$$a_i = \frac{r_i}{\sqrt{1 - r_i}}$$

Where

a_i=item discrimination parameter value for item *i* used in IRT and r_i = item biserial correlation

The relationship is approximate rather than accurate as a result of the different distributions and assigned scores of the two models. The number of correct score (*X*) of CTT and (θ) of IRT have different shapes and the relationship between (*X*) and (θ) is non-linear. Furthermore, the total test score (*X*) is subject to errors of measurement, whereas the ability score (θ) is not. Lord (1980) also prescribed monotonic relationship between p_i and b_i, however this relationship holds well when all items are equally discriminating.

Statement of the problem

No doubt, the property of invariance of person and item characteristics is very critical for objective measurement. However, as observed by Ojerinde (2013) are these demands sufficient to jettison CTT for IRT? If a very large sample (i.e. N > 500 examinees) is used for the estimation of the item parameters will there be any justification for preferring IRT over CTT? Specifically how close will the item parameters be if large sample of examinees is used for the estimation of item parameters (i.e. difficulty and discrimination indices)? These were the major focus of this study.

Research Questions

The following research questions were addressed.

1) What are the item statistics of the 60-item PAT using CTT model and 2-parameter model of IRT?
2) How many items survived after the item analysis using the CTT model and 2-parameter model of IRT?
3) How comparable are the CTT-based and IRT-based item discrimination estimates?
4) How comparable are the CTT-based and IRT-based item difficulty estimates?

Methods

Participants

Seven hundred and twenty four senior secondary school two (SSS II) students Age (16-18 years) participated in the study. The students were selected from 16 randomly selected senior secondary schools in Ibadan Educational Zone I, Oyo State, Nigeria. All the students were offering all science subjects including physics. There were 451 boys (62.3%) and 273 girls (37.7%). In each of the schools sampled, intact class of science was used. That is all the students in the class participated in the study. The average age of the students was 16.9 years (SD = 1.3)

Material

One instrument was used for this study. This was Physics Achievement Test (PAT). The initial draft of PAT consisted of 68 items. It was developed by the author of this paper. 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 into consideration. The items were developed from the content of the physics syllabus and physics curriculum for senior secondary one and senior secondary two. In addition, the items were written by following the pattern of WAEC and NECO. That is each item was placed on four-option response mode of A, B, C, and D. Correct response was given a score of 1, while incorrect response was 0. The items covered three main topics in physics. These were mechanics, heat and electricity.

The decision to develop items from mechanics, heat and electricity was taken, because one, analysis of the scheme of work in all the schools that were sampled showed that all the physics teachers had taught mechanics, heat and electricity. Two, there are many formulas and equations which the students have to master in mechanics, heat and electricity. As noted by the Physics Chief Examiners' (WAEC, 2012), many candidates have difficulties

in the use of equations and formulas in test items. Therefore determining the difficulty and discriminating levels of the items using the two frameworks (CTT and IRT) will help test constructors be aware of which framework will be better in determining the psychometric properties of the items drawn from mechanics, heat and electricity, and hence improve the quality of the items.

Steps in the PAT Development

Because, the *PAT* items were just being developed, the following steps were taken in their development.

Step One: Preparation of the test item pool. The draft copy consisted of 68 items. The items were placed on four-point response format A, B, C, and D. For content validity, test blue print was developed. Table 1 shows the test blue print.

Table 1: Test Blue Print

Subject	Behavioural Objectives			Total
	Knowledge	Understanding	Thinking	
Mechanics	11	5	7	23
Heat	10	3	8	21
Electricity	12	7	5	24
Total	33	15	20	68

Moreover, to further ensure the validity of the test items, two secondary school physics teachers who were also examiners with WAEC and NECO read the draft of the test. From their reactions to the 68 items, eight items (two each from mechanics and heat and four from electricity) were deleted on the basis of their poor rendition.

Step Two: The PAT draft copy consisting of 60 items was administered to the 724 students. The administration of the test was during the normal time scheduled for physics on the school official timetable. This was to avoid disruptions to the school official schedules. The time allowed for the students to take the test was 90 minutes. On the average, it took the students about 70 minutes to finish the test. Two research assistants as well as the physics teacher in each school were involved in the administration of the draft copy of the PTA.

Step Three: The item analysis and detection of poor items were carried out to select the final items. The item analyses were carried out using CTT and IRT frameworks. BILOG-MG (Window Version 3.0) with Marginal-Maximum Likelihood Method was used.

Results

Research Question 1: What are the item statistics of the 60-item PAT using CTT model and 2-parameter model of IRT?

Table 2 presents the item statistics of CTT and IRT. The left hand aspect gives the classical item statistics (difficulty p) and discrimination ($r_{pb.j}$) obtained from phase 1 of BILOG. The right hand gives the discrimination (a) and difficulty (b) parameters of item response theory model obtained from phase 2 of BILOG.

Table 2: Item Statistics of CTT and IRT Models

Item No	*Analysis Using CTT*		*Analysis Using IRT*	
	p	$r_{pb.j}$	a	b
1	0.50	0.49	0.62	0.00
2	0.46	0.60	0.83	0.18
3	0.41	0.33	0.37	0.64
4	0.35	0.52	0.63	0.71
5	0.69	0.34	0.41	-1.24
6	0.49	0.29	0.33	0.05
7	**0.54**	**0.18**	0.23	-0.38
8	0.57	0.21	0.27	-0.65
9	**0.31**	**0.15**	0.22	2.19
10	0.45	0.59	0.76	0.21
11	0.40	0.32	0.38	0.68
12	0.40	0.35	0.42	0.67
13	0.76	0.35	**0.42**	**-1.80**
14	0.75	0.35	**0.41**	**-1.72**
15	0.40	0.63	0.93	0.39
16	0.33	0.40	0.49	1.00
17	0.54	0.47	0.52	-0.02
18	**0.25**	**0.26**	0.31	2.23
19	0.56	0.21	0.25	-0.62

20	0.45	0.32	0.35	0.39
21	0.57	0.58	0.74	-0.28
22	0.46	0.55	0.70	0.17
23	**0.15**	**0.41**	0.53	2.22
24	0.30	0.28	0.30	1.80
25	0.35	0.31	0.33	1.24
26	0.76	0.39	0.54	-1.52
27	0.45	0.57	0.77	0.21
28	**0.28**	**0.51**	0.66	1.05
29	0.50	0.46	0.56	0.01
30	**0.58**	**0.09**	**0.18**	**-1.10**
31	**0.29**	**0.31**	0.39	1.54
32	0.44	0.62	0.80	0.25
33	0.39	0.38	0.45	0.67
34	0.40	0.62	0.81	0.40
35	0.59	0.36	0.43	-0.58
36	0.37	0.41	0.46	0.75
37	0.29	0.43	0.52	1.17
38	**0.47**	**-0.04**	**0.16**	**0.66**
39	0.35	0.23	0.26	1.44
40	0.33	0.18	0.23	1.88
41	0.41	0.27	0.31	0.71
42	0.66	0.31	0.41	-1.06
43	**0.24**	**0.20**	0.25	2.73
44	0.36	0.24	0.28	1.29
45	0.36	0.28	0.34	1.13
46	0.63	0.40	0.48	-0.78
47	0.46	0.24	0.30	0.30
48	**0.28**	**0.25**	0.31	1.91
49	**0.21**	**0.26**	0.31	2.72
50	0.33	0.32	0.35	1.32
51	**0.26**	**0.34**	0.39	1.76
52	0.36	0.30	0.35	1.08
53	0.42	0.41	0.46	0.47
54	0.40	0.27	0.31	0.82
55	**0.20**	**0.23**	0.28	3.03
56	**0.19**	**0.30**	0.36	2.53
57	0.47	0.48	0.60	0.16
58	**0.16**	**0.19**	**0.26**	**4.02**
59	**0.30**	**0.06**	**0.14**	**3.44**
60	**0.48**	**0.14**	**0.21**	**0.26**

Note: Items of special interest are in bold type.

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

Classical Test Theory

On the basis of the criteria set for the difficulty indices (i.e. .30 >*p*> 0.70), items which failed to satisfy the conditions were: 13, 14, 18, 23, 28, 31, 43, 48, 49, 51, 55, 56, and 58. On the basis of discriminating index set (i.e. $r_{pb.j} \leq 0.20$), items 7, 9, 30, 38, 40, 58, 59 and 60 were considered poor. Therefore, on the basis of the level set for difficulty and discriminating indices, 20 items: items 7, 9, 13, 14, 18, 23, 28, 30, 31, 38, 40, 43, 48, 49, 51, 55, 56, 58, 59 and 60 were deleted. The reliability index of the 60 items under CTT was 0.85

Item Response Theory

Under IRT framework, the selection of items depends on the information each of the items contributes to the overall information supplied by the whole test. This requires looking at the test information function. Figure 1 shows the test information function of the 60-item PAT.

Using the test information function, the solid line gives the total information, while the dotted line gives the standard error for a specific ability.

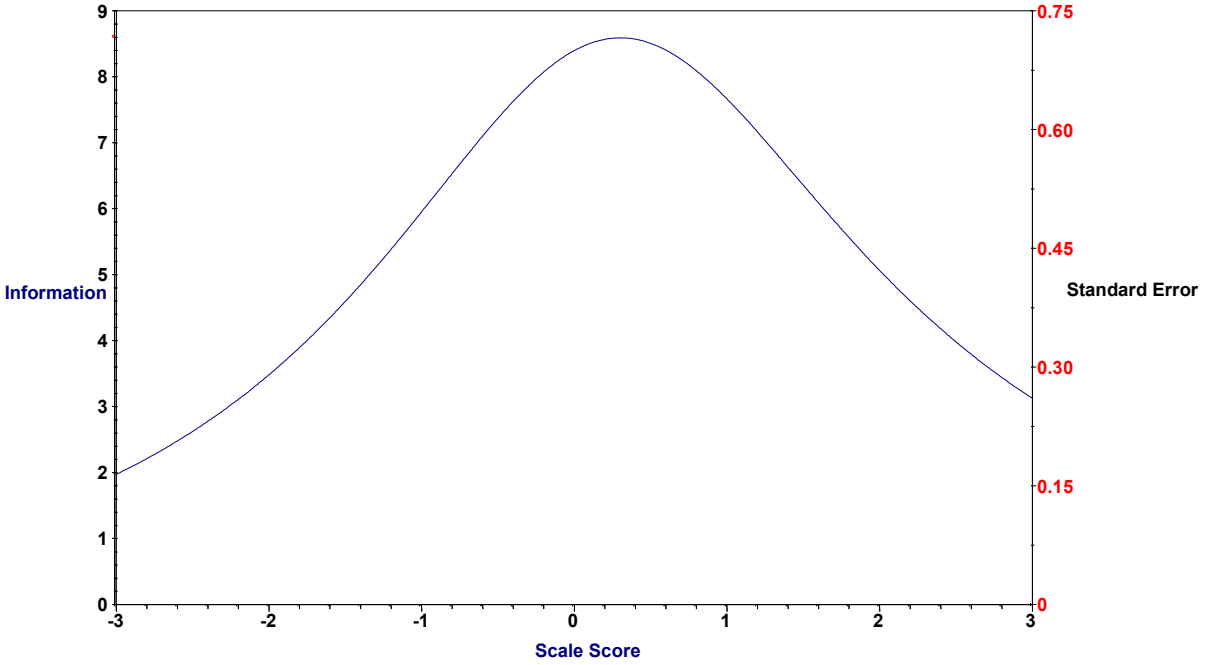

Figure 1: Target Information Function

The test information function as shown in figure 1 shows that the maximum amount of test information was 8.6 at an ability level of 0.25 (i.e. the point at which the curve peaks).

$$SE = \frac{1}{\sqrt{I(\theta)}}$$

Where SE is the standard error of estimation, I is the amount of test information, and θ is the ability level.

From the test information function of Figure 3, items whose difficulty level fall between -1.45 and 2.25 (where the dotted lines cross the solid line) should be included in the final test. On the basis of this, items 13, 14, 26, 43, 49, 55, 56, 58, and 59 which were outside these range were deleted. In addition to this, items with low discriminating indices a lower than 0.20 were also deleted. From table 1, items 30, 38 and 59 were in this category. Therefore using test information function and discriminating index, 11 items were deleted. These were items 13, 14, 26, 30, 38, 43, 49, 55, 56, 58, and 59. The reliability index of the 60 items under IRT was 0.87.

Comparing the item statistics of CTT and IRT, analysis showed that 20 items were deleted by using CTT, while 11 were deleted by using 2-parameter model of IRT. The number of items deleted from using the two frameworks is presented in table 2. These item statistics led to two forms of final draft of PAT: Form A and Form B. Form A (consisted of 40 items) resulted from using CTT while Form B (consisted of 49 items) resulted from using IRT.

Table 3: Items deleted using CTT and IRT frameworks

Model	Number Deleted	Items Deleted
CTT	20	7, 9, 13, 14, 18, 23, 28, 30, 31, 38, 40, 43, 48, 49, 51, 55, 56, 58, 59, and 60
IRT	11	13, 14 , 26, 30, 38, 43, 49, 55, 56, 58, 59
Common items deleted by both models	10	13, 14, 30, 38, 43, 49, 55, 56, 58, 59

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

To answer this question, 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 was carried out. In addition the correlation coefficient of the relationship between of the two parameters was determined. Figure 2 shows the scatter plot of a - values estimated from the two-parameter model of IRT and the r - values of the point biserial correlations of CTT.

The correlation between the a-value and the values of the point biserial correlation r is high and positive (0.970)

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. Theory (e.g. Hambleton& Jones, 1993; Lord, 1980; Wiberg, 2004) states that the correlation coefficient of the relationship between *a* – values and point biserial correlation should be high and positive.

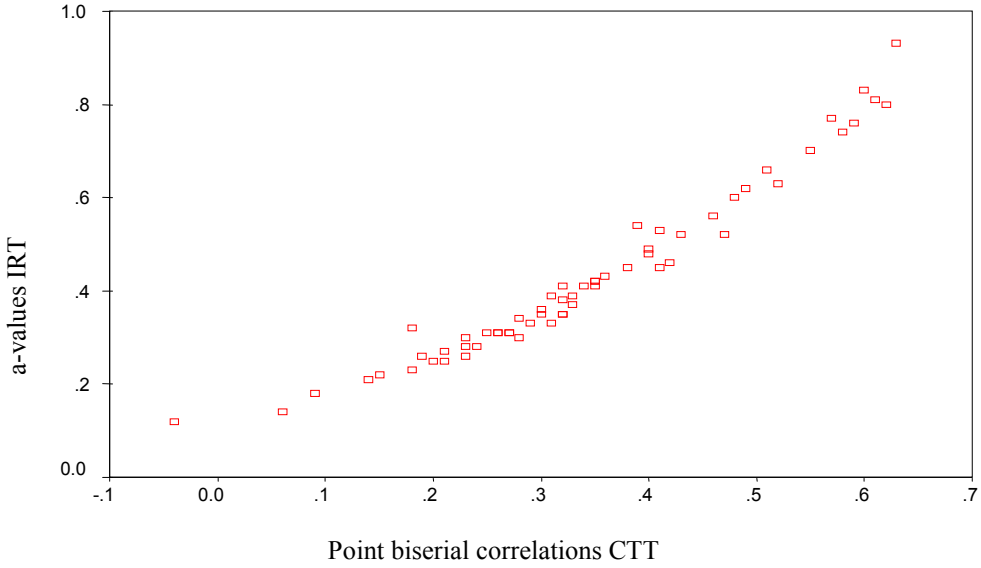

Figure 2: Scatter plot of the relationship between *a* and *r*

Research Question 4: How comparable are the CTT-based and IRT-based difficulty estimates?

To answer this question, a scatter plot of *b* - values estimated from the 2-parameter model of IRT and the *p* - values of CTT was carried out. In addition a correlation of the two parameters was determined. Figure 3 shows the scatter plot of *b* - values estimated from the 2-parameter model of IRT and the *p* - values of of CTT.

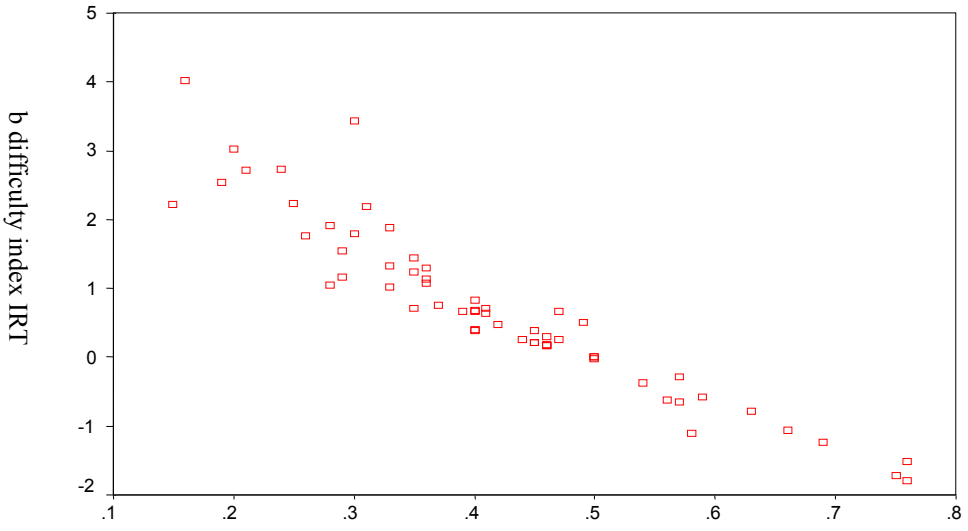

Figure 3: Scatter plot of relationship between *b* and *p*

Theory according to Lord (Hambleton & Jones, 1993) states that the correlation between *p* – values of CTT and b – parameter of IRT should be high and negative. The correlation between the *p* - values and *b* –values is high (-0.942) and statistically significant ($p < .001$). The two item statistics are correlated and therefore a high correspondence exists between the two indices. This shows that as the value of p_i increases, b_i decreases. The results of past studies such as Wiberg (2004) and Stages (2003) laid credence to this.

Discussion and implication of Findings

The major findings of this study were: For the fairly large sample used in this study, 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; Courville, 2004; 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 lay credence on the observation of test experts such as Hambleton and Jones (1993), Ojerinde (2013) that despite the popularity of classical item statistics as an integral part of standardised test and measurement technology, it is fraught with so many limitations. A major limitation of item statistics under CTT framework is that they are examinee and item sample dependent. Therefore, the estimates of CTT are not generalizable across populations. An awareness of the shortcomings of CTT and the potential benefits offered by IRT has led some examining bodies in Nigeria, for example Joint Admissions and Matriculation Board (Ojerinde, 2013) to start working within the IRT framework.

The results of some past studies (e.g. Bechger, Gunter, Huub & Bèguin, 2003; Courville, 2004; **Fan, 1998**; MacDonald & Paunoen, 2002; Stage, 2003) showed little or no superiority of IRT models over CTT models in item parameters estimates. However, it must be noted that in these studies, data for the analysis were obtained from standardised test items. For example, Courville's (2004) work was on data (the samples of the examinees and the items) used in the ACT Assessment Test in the United States of America. In this study, the author's emphasis was on calibration of a newly constructed 60-item test. Item analysis, using the frameworks of CTT and IRT, was carried out to assess the similarities and the differences in the contrasting models. More importantly, the author assessed the number of items that were deleted by each of the contrasting frameworks.

Although the results of some past studies (e.g. Bechger, Gunter, Huub & Bèguin, 2003; Courville, 2004; MacDonald & Paunoen, 2002; Stage, 2003) showed little or no superiority of IRT models over CTT models in item parameters estimates. However, it must be noted that in these studies, data for the analysis were obtained from standardised test items. For example, Courville's (2004) work was on data (the samples of the examinees and the items) used in the ACT Assessment Test in the United States of America. In sharp contrast to Courville's study, in this study, the author's emphasis was on calibration of a newly constructed 60-item test.

The implication of this is that IRT models especially 2-parameter model may be more useful in the calibration of test items. Therefore test developers and public examining bodies should integrate IRT models into their test development processes. Through IRT model, estimation of item parameters will be more reliable. However, because in the overall, item statistics from both IRT and CTT frameworks are comparable in some cases, the author recommends that CTT framework could be used as a complement to IRT.

The extent to which the use of IRT item statistics estimates and IRT item statistic estimates could affect the examinees' scores in the whole test should be examined in future studies.

References

Baker, B. F. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation

Bechger, T., Gunter, M., Huub.,& Bèguin, A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27 (5), 319 – 334.

Courville, T. R. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*. Unpublished Doctoral Thesis, Texas A & M University

Crocker, L. & Algina, J. (1986).*Introduction to classical & modern test theory*. New York: Holt, Rinehart and Winston.

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.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. Tinsley & S. Brown (Eds.) *Handbook of applied multivariate statistics and modelling.* San Diego, CA: Academic Press.

Kelly, T. L. (1939). Selection of upper and lower groups for the validation of test items.*Journal of Educational Psychology*, 30, 17 – 24

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum

Lord, F. M., & Norvick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

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.

Stage, C. (2003).*Classical test theory or item response theory: The Swedish experience* (No. 42). Umea: Kluwer

Academic Publisher

Wang, T., & Hanson, A. (2001).*Development and an item response model that incorporates response time*. A paper presented to the annual meeting of the American Education Research Association in Settle, April.

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

Benson Adesina Adegoke PhD was born on 9 September, 1961. He had taught physics and mathematics at the senior secondary school for about seventeen years before joining the University of Ibadan as a research fellow. He is currently a research fellow at the Institute of Education, University of Ibadan. He teachers research methods and statistics to master's and doctoral students. He is a member of Science Teachers' Association of Nigeria as well as American Psychological Association.