Analysis of Physics Items Using Classical Test Theory Methodology: A Study of Technical Colleges in Enugu State

[1]Leonard Chinaedum Anigbo Ph. D & Mark Eze Ugwoke [2]
[1]Science and Computer Education Department
Enugu State University of Science and Technology (ESUT),
Enugu State
drlcanigbo@gmail.com, ***leoanigbo@esut.edu.ng***
[2] National Business and Technical Examinations
Board (NABTEB), South East Zonal Office,
1, Annang Street, Ogui New- Layout,
P.M.B. 01013, Enugu
markugwoke@yahoo.com

Abstract

The researchers, in this study, analysed physics items using Classical Test Theory methodology for technical colleges in Enugu state. The research design adopted for the study was instrumentation. All physics students in the two purposively selected technical colleges drawn from Enugu and Nsukka educational zones of Enugu State were used for the study. The instrument for data collection, an 80-item Physics Achievement Test (PAT) developed by the researcher was face- and content–validated by three experts and a reliability coefficient of 0.75 was obtained using Kuder-Richardson formular 20 (K-R 20). The instrument was administered to 530 National Technical College three (NTC 3) students of the sampled colleges and their responses were analysed using Xcalibre (item parameter estimation software). The results revealed that some items of the PAT possessed poor psychometric qualities and were not included in the final test. The major recommendation is that examination boards/councils and test development experts should adopt appropriate software for item analysis in the test development processes as a panacea towards ensuring that items of sound psychometric characteristics are included in the final tests.

Introduction

In education and psychology, there are two measurement frameworks by which psychometric properties of items can be studied: the Classical Test Theory (CTT) whose major focus is on test-level information and the Item Response Theory (IRT) whose primary interest is on the item – level information **(**Tschering, 2006; Thorpe and Favia, 2012). The psychometric properties of physics items can be ascertained using item analysis. According to Nworgu (2015), item analysis is the assessment of essential qualities of items in a test. Item analysis is "designed to ensure that items function as they are intended; for example, that criterion – referenced items fairly cover the fields and criteria and that norm-referenced items demonstrate items discriminability; the level of difficulty of the items is appropriate; the test reliability (free of distracters-unnecessary information and irrelevant cues)" (Gronlund and Linn, 1990 : 255). An item analysis will consider the accuracy level available in the answer, the item difficulty, the importance of the knowledge or skill being tested, the match of the item to the programme and the number of items to be included (Cohen, Manion & Morrison, 2009).

Charles Spearman in 1904 through his work in which he established how to correct a correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in making the correlation is the foremost founder of Classical Test Theory (Wikipedia, 2014). Adegoke (2013) reported that Classical Test Theory tries to explain the link between the observed score, the true score, and the error score. It is a simple linear model linking the observable test scores (X) to the sum of two un-observable (or often called latent) variables, that is, true score (T) and error score (E) (Zeng and Wyse, 2009; Ekwonye and Eguzo, 2011; Adegoke 2013 and Ojerinde, 2013).

Mathematically,

$$X = T + E$$

where X represents an observed score

T represents a true score

E represents an error, with the population mean, 0.

The Classical Test Theory (CTT) has been the foundation for measurement theory for decades; it describes how error can influence observed scores, or measurement

(Ojerinde, Popoola, Ojo, and Onyeneho, 2012). The true score "T" reflects the examinee's amount of knowledge which is always contaminated by random errors (Ojerinde, Popoola, Ojo, and Onyeneho, 2012). According to Lord (1980), these random errors can result from several factors such as guessing, fatigue or stress. The observed score, according to him is often called a fallible score because of the error contaminant. The true score is the score that would have been obtained if there were no error in measurement (Lord, 1980).

Ekwonye and Eguzo (2011) defined Classical Test Theory as a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. Generally speaking, the aim of Classical Test Theory is to understand and improve the reliability of psychological tests. Ekwonye and Eguzo (2011) stated that CTT was born only after the following three achievements or ideas were conceptualized: one, recognition of the presence of errors in measurement; two, conception of that error as a random variable; and three, a conception of the correlation and how to index it.

Adegoke (2013) noted that in the equation linking observed score (X) with true score (T) and error score (E), there are two unknowns and posited that these make it not easily solvable unless some simplifying assumptions are made. The three assumptions in the Classical Test Theory are;

(a) that the error score and true score obtained from the same test are uncorrelated (or have a correlation of zero). Hence, the variance of the observed score is expected to be equal to the sum of the variances of the true score and error score; $\delta_x^2 = \delta_T^2 + \delta_E^2$ (Lord, 1980).
(b) that the error terms have an expected mean of zero (or the average error score in the population of examinees is zero (Lord, 1980). Once the error is zero, the observed score is equal to the true score $(X = T)$, $\sum_i^n \frac{E}{N} = 0$; and
(c) that errors from parallel measurements are uncorrelated (Lord, 1980) .
Symbolically, $X//X^1$ if $X_1 = X_2 = T_1 + E_1$

Adegoke (2013) remarked that

although the major focus of Classical Test Theory is on test level information, the reliability can provide a convenient index of test quality in a single number. However, it does not provide any information for evaluating single item. Tschering (2006) noted that item analysis within the classical approach often relies on two statistics: the P- value (proportion) and item-total correlation (point-biserial correlation coefficient). The P-value represents the proportion of examinees responding in the keyed direction, and is typically referred to as item difficulty. The item-total correlation, he continued, provides an index of discrimination or differentiating power of the item, and is typically referred to as item discrimination (Zeng and Wyse, 2009; Ekwonye and Eguzo, 2011). In addition, these items statistics are calculated for each response of the often-used multiple choice item, which are used to evaluate items and diagnose possible issues, such as a confusing distracter.

According to Zeng and Wyse (2009), for dichotomously scored items (1 for correct answer and 0 for incorrect answer), the item difficulty (often denoted by P) is one of the two major item statistics used in item analysis and selection. Zeng and Wyse (2009) stated that the success rate of a pool of examinees on an item is used as the index for item difficulty. Anigbo (2014) defined item difficulty index as the proportion or percentage of the examinees that got the item right or passed the item and gave the mathematical expression as:

$$\text{Item difficulty index (P)} = \frac{\text{Number of examinees who got the item right}}{\text{Number of examinees who tried it}}$$

Anigbo (2014) stated that ideal value for item difficulty index (P) is 0.50. Anigbo (2014) however noted that this value depends on the purpose for which the test is constructed and gave the range considered as appropriate for the difficulty index (P) as 0.30 to 0.70, that is, $(.30 \leq P \leq .70)$. However, for a particular multiple choice item with four options, this study adopted a range of item difficulty index (P) whose P-value is greater than 0.20 and less than 0.95 ($0.20 < P < 0.95$); the higher the P-value, the easier the item and vice versa.

According to Anigbo (2014:52) item discrimination index (often denoted by r_{pb}) is defined as the degree to which the passing or failing of an

item depends on the ones' belonging to the upper or lower group. Anigbo (2014: 52) also defined "item discrimination index as the correlation between the ones' belonging to the upper or lower group and scoring a particular item correctly". The following methods are used in CTT to assess item discrimination (Adegoke, 2013; Anigbo, 2014):

(a) Finding the difference in the proportion of high achieving examinees and low achieving examinees who scored the items correctly; and

(b) Biserial correlation or point-biserial correlation between a dichotomously scored item and the scores on the total test.

Even though the first method is generally accepted in educational measurement for the estimation of item discrimination index, the problem with it is that it omits the data of a lot of examinees (e.g. 46% of examinees) and this problem can be corrected by using the point- biserial correlation, γpb.j for item j; a computationally simplified Pearson's r between the dichotomously scored item j and the total score X (Hambleton and Jones, 1993).

According to Zeng and Wyse (2009), item discriminating power is computed as:

$$\gamma_{pb.j} = \frac{\mu_j - \mu_x}{\delta_x}\sqrt{\frac{P_j}{q_j}}$$

where μ_j = the mean total score among examinees who have responded correctly to item j,

μ_x, = the mean total score for all examinees who attempted the item;

p_j = the item difficulty index (proportion of the examinees that got the item j correct), while

q_j = the proportion of the examinees who got the item j wrong i.e (q_j =1-p_j), and

δx = the standard deviation of the examinees' total score.

Anigbo (2014) stated that this discrimination index (denoted by D or r_{pb}) ranges from -1.0 to +1.0 and that an ideal item should have an r_{pb} of +1.0. Practically, he noted that the range is from +0.30 to +1.0 ($0.30 \leq r_{pb}$

≤ I.0). However, the range for the Pearson point – biserial correlation (r_{pb}) –value used for this study is from positive 0.05 to 1.0 ($0.05< r_{pb} < 1.0$). The negative value of the index implies that more candidates in the lower group scored a particular item correctly. This situation occurs when the item stem or the answer options are so confusing that the clever students are misled to choosing a wrong option; or all the options are either correct or wrong and the choice is based purely on guessing (Anigbo, 2014).

Numerous researchers have confirmed students' poor achievement in physics tests especially in Technical Colleges in Nigeria. They isolated the causes of the problem and suggested ways of improving students' achievement in Physics tests. For instance, while Ugwoke (2014) suggested the use of improvised instructional materials, others opined that motivation of Physics teachers through the provision of incentives such as science or hazard allowance, training and retraining of teachers, provision of scholarship schemes for students that enrol for the subject as well as equipping physics laboratories with required tools, apparatus and materials will enhance their achievement in physics.

However, little attention has been paid by researchers in physics education to the critical issue of quality physics items' (instrument) especially the multiple-choice-test being administered to the students (Adegoke, 2013). Multiple-choice items whose item parameters (i.e. item difficulty and item discrimination power) are not critically analyzed and the item considered good enough for inclusion in a test will yield undesirable responses (i.e. responses not reflecting the true abilities) of examinees when eventually used in the physics test paper (Ebuoh, 2004 ; Anigbo, 2014). Could the neglect of item analysis to ensure that items of sound qualities are included in the test be the cause of the persistent poor achievement of students in physics tests? Could the use of Xcalibre (new item parameter estimation software) to establish the item statistics of Physics Achievement Test provide the solution to the problem? In addition, the researchers are aware that, at post-basic-education level in Nigeria, previous studies on this concept have been exclusively carried out at secondary schools' section to the neglect of technical colleges. The (researchers) have no knowledge of any study of this type ever carried out in technical colleges in Nigeria. These are serious gaps which

this study intends to fill so as to find lasting solutions to students' poor achievement in Physics tests.

Research Questions

The following research questions were posed:

(i) What are the item difficulty indices (p) of the 80 –item PAT?

(ii) What are the item discrimination indices (r_{pb}) of the 80 – item PAT?

(iii) Which items of the PAT are considered good for inclusion in the test?

Methods

The instrumentation research design was employed. According to Garuba (1993), instrumentation studies are studies that are aimed at introducing practices, techniques or instruments for educational practices. In the words of Hsu and Standford (2013), instrumentation design refers to tools or means by which researchers attempt to measure variables or items of interest in the data collection process. This design was adopted because the study is related to the instrument's design, selection, construction, assessment as well as the conditions under which the designated instrument is administered. The area of study was Enugu State that comprised six educational zones namely: Enugu, Nsukka, Obollo-Afor, Awgu, Udi and Agbani zones. The population for this study consisted of all the 580 year III (NTC 3) students of four out of all accredited technical colleges in Enugu State for the 2015/2016 academic session. Purposive sampling technique was employed in selecting two colleges (Government Technical College, Enugu and Government Technical College, Nsukka) with a total of 530 year III (NTC 3) students which constituted the sample. The instrument, developed by the researchers comprised 80 – multiple-choice physics items called Physics Achievement Test (PAT).

Results: The results of the PAT's analysis using Xcalibre are shown on the table 1 below.

Table 1: **Item Statistics of the PAT**

Seq.	Item ID	PAT's Item Statistics		Flags	Decisions
		P	**$r_{pb.j}$**		
1	1	0.626	0.260		**Selected**
2	2	0.557	0.210		**Selected**
3	3	0.338	0.079		**Selected**
4	4	**0.140**	**0.031**	Lp, L $r_{pb.j}$	Not Selected
5	5	0.619	0.167		**Selected**
6	6	0.417	0.290		**Selected**
7	7	**0.134**	0.058	Lp .	Not Selected
8	8	**0.106**	**-0.027**	Lp, L$r_{pb.}$	Not Selected
9	9	0.628	0.200		**Selected**
10	10	**0.162**	**0.070**	Lp, L$r_{pb.}$	Not Selected
11	11	0.292	0.177		**Selected**
12	12	0.592	0.274		**Selected**
13	13	0.275	**0.048**	L $r_{pb.}$	Not Selected
14	14	0.215	0.159		**Selected**
15	15	0.300	0.155		**Selected**
16	16	0.757	0.202		**Selected**
17	17	0.247	**0.017**	L$r_{pb.}$	Not Selected
18	18	0.260	0.094		**Selected**
19	19	0.274	0.050		**Selected**
20	20	0.332	0.109		**Selected**
21	21	0.806	0.171		**Selected**
22	22	0.221	0.056		**Selected**
23	23	0.283	**-0.010**	L$r_{pb.}$	Not Selected
24	24	0.289	**-0.050**	L$r_{pb.}$	Not Selected
25	25	**0.189**	0.190	Lp.	Not Selected
26	26	**0.162**	0.104	Lp	Not Selected
27	27	0.342	0.175		**Selected**
28	28	0.306	0.107	.	**Selected**
29	29	**0.160**	**-0.001**	Lp, L$r_{pb.}$	Not Selected
30	30	**0.157**	**-0.097**	Lp, L$r_{pb.}$	Not Selected

31	31	0.598	0.189		**Selected**
32	32	**0.136**	0.080	Lp	Not Selected
33	33	**0.132**	0.169	Lp	Not Selected
34	34	0.319	**0.017**	$Lr_{pb.}$	Not Selected
35	35	0.313	**0.046**	$Lr_{pb.}$	Not Selected
36	36	0.475	0.108		**Selected**
37	37	**0.185**	0.051	$Lr_{pb.}$	Not Selected
38	38	0.249	0.164	.	**Selected**
39	39	**0.166**	0.116	Lp.	Not Selected
40	40	**0.132**	0.189	Lp	Not Selected
41	41	**0.189**	0.127	Lp.	Not Selected
42	42	0.362	0.239		**Selected**
43	43	0.723	0.153		**Selected**
44	44	0.249	0.076		**Selected**
45	45	0.209	**-0.084**	$Lr_{pb.}$	Not Selected
46	46	0.211	**-0.014**	$Lr_{pb.}$	Not Selected
47	47	0.511	0.140		**Selected**
48	48	0.313	0.157		**Selected** .
49	49	0.247	0.291		**Selected**
50	50	**0.070**	**-0.023**	Lp, $Lr_{pb.}$	Not Selected
51	51	**0.117**	0.054	Lp	Not Selected
52	52	**0.121**	**0.024**	Lp, $Lr_{pb.}$	Not Selected
53	53	**0.068**	0.063	Lp	Not Selected
54	54	0.349	0.239		**Selected**
55	55	**0.175**	0.060	Lp	Not Selected
56	56	**0.162**	0.104	Lp.	Not Selected
57	57	0.389	0.154		**Selected**
58	58	0.643	0.305		**Selected**
59	59	0.655	0.280		**Selected**
60	60	0.385	0.108		**Selected**
61	61	0.653	0.221		**Selected**
62	62	0.675	0.136		**Selected**
63	63	0.345	0.206		**Selected**
64	64	0.330	0.171		**Selected**
65	65	**0.113**	**0.024**	Lp, $Lr_{pb.}$	Not Selected

66	66	**0.172**	0.142	Lp,	Not Selected
67	67	0.391	0.329		**Selected.**
68	68	0.440	0.302		**Selected**
69	69	0.251	0.139		**Selected**
70	70	0.309	0.275		**Selected**
71	71	**0.194**	0.147	Lp.	Not Selected
72	72	0.202	**-0.008**	Lp	Not Selected
73	73	0.225	0.135		**Selected**
74	74	0.366	0.141		**Selected**
75	75	0.347	**-0.043**	$Lr_{pb.}$	Not Selected
76	76	0.219	0.084		**Selected**
77	77	0.215	**-0.016**	$Lr_{pb.}$	Not Selected
78	78	0.202	0.119		**Selected**
79	79	**0.185**	**0.038**	Lp, $Lr_{pb.}$	Not Selected
80	80	**0.075**	**-0.009**	Lp, $Lr_{pb.}$	Not Selected

Key: Hp = High Difficulty, Lp = Low Difficulty, Hr_{pb} = High Discrimination and Lr_{pb} = Low Discrimination.

Research Question 1*:* What are the item difficulty indices (P) of the 80-item PAT?

Table 1 contains item statistics of the PAT showing item difficulty (P) parameters and it revealed that item identity numbers: 4, 7, 8, 10, 25, 26, 29, 30, 32, 33, 37, 39, 40, 41, 50, 51, 52, 53, 55, 56, 65, 66, 71, 79, and 80 possess low item difficulty indices: 0.140, 0.134, 0.106, 0.162, 0.189, 0.162, 0.160, 0.157, 0.136, 0.132, 0.185, 0.166, 0.132, 0.189, 0.070, 0.117, 0.121, 0.068, 0.175, 0.162, 0.113, 0.172, 0194, 0.185, and 0.075 respectively which is less than 0.20 (P < .20). These items are considered too difficult (P < .20). However, none of the items of PAT is too easy (P > .95).

Research Question 2: What are the item discrimination indices (r_{pb}) of the 80-item PAT?

Table 1 contains item statistics of PAT showing item discrimination (r_{pb})

parameters and it revealed that item identity numbers: 4, 8, 10, 13, 17, 23, 24, 29, 30, 34, 35, 45, 46, 50, 52, 65, 72, 75, 77, 79, and 80 possess item discriminating indices of 0.031, -0.027, 0.070, 0.048, 0.017, -0.010, -0.050,

-0.001, -0.097, 0.017, 0.046, -0.084, -.0.014, -0.023, 0.024, 0.024, -0.008, -0.043, -0.016, 0.038, and

-0.009 respectively. These items discriminate very poorly because their item discrimination indices are too low ($r_{pb} < .05$). However, none of the items is discriminating too highly ($r_{pb} >. 1.0$).

Research Question 3: Which items of the PAT are considered good for inclusion in the final test?

Data in table 1 revealed that using the benchmark set for the difficulty indices ($.20 < P < .95$) and the benchmark set for the discriminating indices ($.05 \leq r_{pb} \leq 1.0$) the items with identity numbers listed below satisfy the criteria set for item inclusion in the final test: 1, 2, 3, 5, 6, 9, 11, 12, 14, 15, 16, 18,19, 20, 21, 22, 27, 28, 31, 36, 38, 42, 43, 44, 47, 48, 49, 54, 57, 58, 59, 60, 61, 62, 63, 64, 67, 68, 69, 70, 73, 74, 76, and 78. The summary of the results is that whereas forty – four (44) items of the PAT are considered good for inclusion in the final test, a total of thirty six (36) items of the PAT did not satisfy the criteria set for selection; their item difficulty indices and item discrimination indices were out of the prescribed ranges (benchmarks) adopted for this study and are consequently discarded or removed from the final test.

Discussions

The results and findings of this study whereby some items of the PAT were discarded were supported by Ojerinde (2013) in his study on item analysis using Classical Test Theory. He found out that that some items possessed poor psychometric qualities (item difficulty and item discrimination) and consequently removed such items from the final test. In a similar vein, Adegoke (2013) who studied the comparison of item statistics of Physics Achievement Test using CTT and 2-parameter IRT model of IRT found out that some items of the PAT possessed unsound psychometric characteristics. He subsequently removed such items from the PAT. Anigbo (2014) similarly, stated that test development specialists must, as a matter of rule, generate more items than are needed in the test due to item mortality during trial testing.

Conclusion.

The Classical Test Theory is a veritable measurement framework in test development processes with this study revealing that forty-four (44) items of the 80-item Physics Achievement Test possess sound psychometric qualities while thirty-six (36) items possess poor psychometric qualities.

Recommendations

1. It is recommended that examining boards and other testing agencies using multiple-choice- test instrument should use Pearson point –biserial (item correlation) to ascertain classical item discrimination and classical item difficulty to guarantee the quality items that constitute tests.
2. Educational institutions and examining boards in Nigeria should sponsor graduate research work in this area of study. They should train and retrain staff to acquire skills and competencies on the use of Xcalibre (item parameter estimation software) and other assessment computer packages through local and international workshops and seminars.

References

Adegoke, B.A. (2013). Comparison of item statistics of Physics achievement test using classical test theory and item response theory frameworks: *Journal of Education and Practice.* (4) 22, 87.

Anigbo, L.C. (2014). *Teachers' handbook on measurement and evaluation for effective teaching and learning*. Enugu: Executive Press Limited.

Cohen, L; Manion, L; & Morrison, K; (2009). *Research Methods in education (6th edition).* New York: Routledge.

Ebuoh, C. N. (2004). *Educational measurement and evaluation for effective teaching and learning*. Enugu: Sky Printing Press.

Garuba, N. L.(1993). *Development of instrument for evaluating practical project in wood-working*. Unpublished Ph.D Thesis, U.N.N.

Gronlund, N. E. and Linn, R. L. (1990). *Measurement and evaluation in teaching* (sixth edition). New York: Macmillan.

Guyer, R., & Thompson, N. A. (2014). *User's manual for Xcalibre item response calibration software*, (version 4.2.2 and later). Woodbury MN: Assessment Systems Corporation

Hambleton, R. K. & Jones R. W. (1993). Comparison of classical test theory and item response theory and application to test development. *Education measurement: Issues and Practice*. 12 (3), 284 – 7.

Hsu, Chia- Chien & Standford, A. Brain (2013). *Instrumentation: Sage Research Methods.* DOI:http://dx.doi.org/10.4135/97814129688

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

Nworgu, B. G. (ed) (2015). *Educational measurement and evaluation: theory and practice*. Nsukka: University Trust Publishers.

Ojerinde, D.; Popoola, K.; Ojo, F.; and Onyeneho, O. P. (2012). Introduction to item response

theory: Application to depression measured. *Psychological Assessment.* 12(3), 354 – 359.

Ojerinde Dibu (2013). *Classical test theory vs item response theory: An evaluation of the comparability of item analysis results.* A paper presented at the Institute of Education, University of Ibadan on May 23.

Thorpe, G. L. and Favia, A. (2012). Data analysis using item response theory methodology: an introduction to selected programmes and applications. *Psychology Faculty Scholarship; Paper 20.* The University of Maine.

Tschering, G (2006). *IRT in item banking; study of DIF items and test construction.* Published M.Sc Thesis Educational Science and Technology University of Twente Euschede, The Netherlands.

Ugwoke, M. E. (2014). Use of improvised instructional materials in teaching Ohm's law: A study of technical colleges in Enugu State. *Journal of Studies in Education.* 3(1) pp.190-195

.Wikipedia (2014). *Classical Test Theory*. Retrieved on August 20 from https://en.wikipedia.orgwiti/classical_test_theory.

Zeng, J. and Wyse, A. (2009). Introduction to Classical Test Theory. *Paper Presentation*. Michigan Department of Education; Office of Educational Assessment and Accountability