The Role of Item Analysis in Detecting and Improving Faulty Physics Objective Test Items

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

Abstract
Results of candidates' level of achievement in physics tests being conducted by public examining bodies show that candidates are not doing well in these public examinations. One of the reasons for the low level achievement in physics may be due to faulty test items that are being administered to candidates each year. The use of faulty items may be as a result of not conducting thorough analysis of these items before they are administered to candidates. In this study, therefore, the author examined the role of item analysis in detecting faulty items and in improving the quality of physics objective test items.

Physics Paper 2 (PP2) of 2012 West Africa Examination Council consisting of 50 multiple choice items was administered to 900 senior secondary school three students (Age 16-18 years), who were randomly selected from 16 senior secondary schools in Ibadan Educational Zone I, Oyo State, and Irewole Local Government Area, Osun State, Nigeria. Specifically the difficulty and discriminating indices of each of the 50 items in PP2 were determined. Analyses of the students' responses to the 50-item PP2 were carried out using both Classical Test Theory and Item Response Theory. Results of the analysis showed that some of the items were faulty in that their difficulty indices were very high and some items could not discriminate between high achievers and low achievers.

These results imply that public examining bodies should carry out thorough analysis of their test items before they are administered to their candidates. More importantly the results of the analysis should be communicated to physics teacher so that they can use the information to clarify some concepts that may be confusing to students in physics.

Keywords: Difficulty index, Discriminating index, Item response theory, Classical test theory, Physics objective test.

1. Introduction

Probably, most physics teachers as well as item writers for public examining bodies construct the tests they give secondary school students with little thought that a detailed analysis of the test items might result in their improvement. This is because a cursory look at some past question papers of major examining bodies reveals that year in year out; the same faulty items are repeated with little or no modification to the stem as well as the options. It is quite unfortunate that some physics teachers instead of developing their own test items go ahead to use the faulty past questions of the public examining bodies, in their school-based examinations. No wonder, the same poor results in public examinations are being recorded among secondary school students.

What are the characteristics of good test items? Good items should have optimum difficulty level and be able to discriminate between students who have learned more than those who have learned less. In general items that will yield reliable measurements are good items (Field, 2006). After all, without reliability, there is no hope for validity, and we must have valid measurements or we will be deceiving ourselves. A pertinent question arises here: How reliable and valid are the items that public examining bodies use in their secondary school certificate examinations? To what extent do these items discriminate between students who have learned more than others and those who have learned less? How difficult are these test items? Answers to these questions will help in determining the quality of these items and consequently improve them.

The major objective of this study, therefore, is to analyse item by item 2012 WAEC physics objective test. In order to do this, focus will be on difficulty and discrimination levels of each item. The results of these analyses will help physics teachers as well as item writers realise the need for analysing the responses of the examinees even at the end of the examination with a view to improving the quality of the test items.

Item analysis, according to Haladyna (1999) is a method of reviewing items on a test, both qualitatively and statistically, to ensure that they all meet minimum quality-control criteria. The difference between qualitative review and statistical analysis is that the former uses the expertise of content experts and test review boards to identify items that do not appear to meet minimum quality-control criteria. Such qualitative review is essential during item development when no data are available for quantitative analysis. A statistical analysis, such as item analysis, is conducted after items have been administered and real-world data are available for analysis. However, the objective of qualitative and statistical review is the same – to identify problematic items on the test. According to Varma (2014) items may be problematic due to one or more of the following reasons:

a) Items may be poorly written causing students to be confused when responding to them.

b) Graphs, pictures, diagrams or other information accompanying the items may not be clearly depicted or may actually be misleading.

c) Items may not have a clear correct response, and a distractor could potentially qualify as the correct answer.

d) Items may contain distractors that most students can see are obviously wrong, increasing the odds of students guessing the correct answer.

e) Bias for or against a gender, ethnic or other sub-group may be present in the item or distractors.

There are three main levels or degrees of analysis of items. The first level reveals how difficult each item is. This serves two purposes. First, it indicates which items are very difficult (less than 20% of the examinees get them correct). These items need revision or else the instruction needs improvement for the future. Second it indicates which concepts need further discussion and elaboration with a given class before proceeding with new material (this is very important in the school-based examinations). Even in examinations being conducted by public examining bodies, the results of the analysis could be sent to physics teachers in schools for them to help students understand better concepts that may appear confusing.

The second level of analysis indicates not only the degree of difficulty of each item, but also the degree to which the people who get the item correct also get other items on the test correct, and the people who get the item wrong also get the other items on the test wrong. This is the discrimination value of the test item.

The third level of analysis indicates not only the difficulty level and discrimination value for each item but what happened to each of the decoy responses. The best items have a pattern of responses in which every alternative is chosen by at least one examinee, the correct alternative is chosen by more of the students who get good scores on the total test than by the students who get poor scores, and the reverse is true for the decoys, i.e., they are more popular for the poor scorers than the good scorers. Any alternatives that do not fit that pattern should be revised before subsequent use of the item in another test.

In educational measurements, according to Baker (2001), 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). In this study emphasis is on IRT this is because of the limitations of the CTT when applied for item analysis. Whereas CTT item statistics depend fundamentally on the subset of items and persons examined, IRT item and person parameters are invariant, thus allowing the researcher to assess the contribution of individual items as they are added or omitted from a test.

Item response theory assumes there is a mathematical function that relates the probability of a correct response on an item to an examinee's ability (See Lord 1980 for a detailed discussion). Many different models of these functional relationships are possible. However for this study, the model chosen was the two-parameter model. In this model, the probability P of a correct response to an item i for an individual with ability Θ $P_i(\theta)$ is

$$P_i(\theta) = \frac{1}{1+e^{-1.7a_{i(\theta-b_i)}}} \qquad \text{….Eqn. 1}$$

Where

a_i = discrimination parameter, b_i is the difficulty parameter, and Θ is the examinee's ability. When employing item response theory, 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). As is the case with CTT, selection of items under IRT models depends on the intended purpose 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 item information and test information functions. An example of item information function is presented in Figures 1.

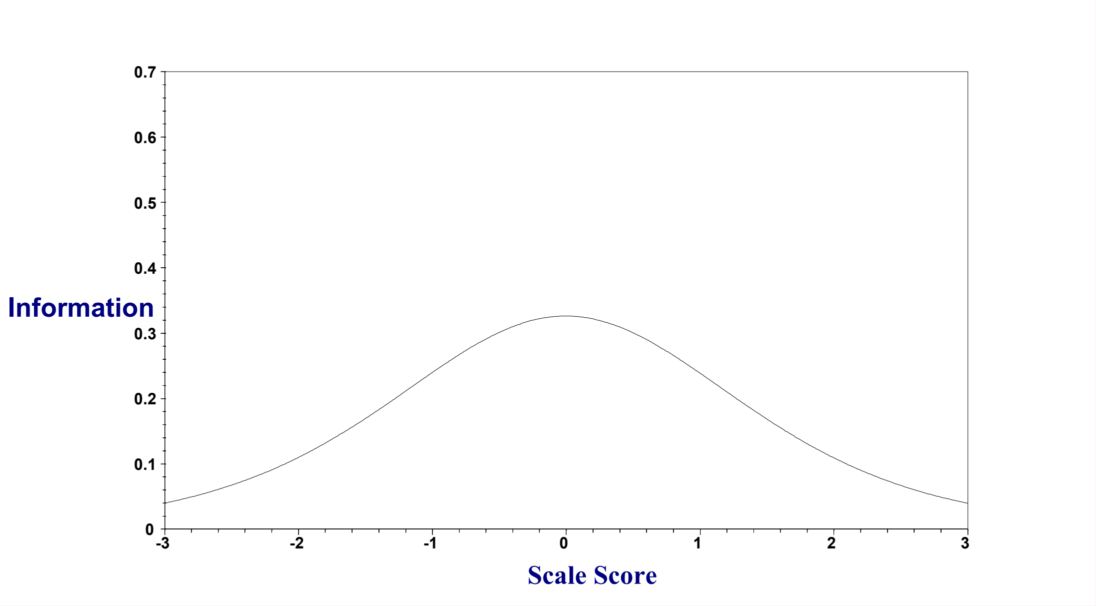


Figure 1: Item Information Function (or curve)

The item information functions show the contribution of the item to the assessment of examinee's ability. In general, items with high discriminating power contribute more to measurement precision than items with lower discriminating power, and items tend to make their best contribution to measurement precision around their *b* value on the ability scale. Maximum information given by an item is estimated at the point where the curve peaks.

For example, using Item information Function of Figure 1, the maximum information is given at ability level (scale score axis) of Θ = 0, that is the curve peaks at the point 0 on the scale score axis. The amount of information at this ability level is about 0.3.

The item information can also be calculated using

$I_i(\theta)$=$a_i^2 P_i(\theta)Q_i(\theta)$

Where: a_i is the discrimination parameter for item *i*.

$P_i\theta = 1/(1 + EXP(-a_i(\theta - b_i)$

$Q_i(\theta) = 1 - P_i(\theta)$

Θ is the ability level of interest.

Information Function

Since a test is a set of items, the test information (Figure 2) is more appropriate to detect and select items under IRT Framework.

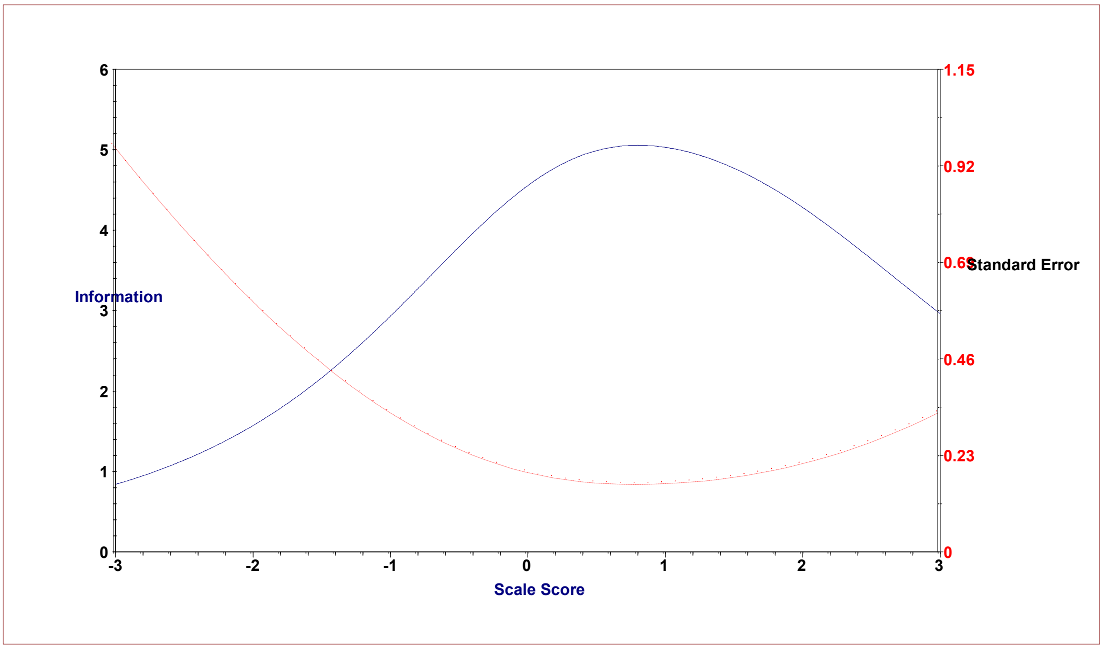


Figure 2: Test Information TIF

The test information at a given ability level is simply the sum of the item information at that level. Consequently, the test information is given by

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta).$$

Where *I (Θ)* is the amount of test information at an ability level of *Θ,*
$I_i(\theta)$ is the amount of information for item *i* at ability level *Θ*,
N is the number of items in the test.
Item and test characteristic functions and item and test information functions are integral features of item response theory models and they are immensely useful in item analysis and test development (Hambleton & Jones, 1993).

In this study, the major focus was on thorough analysis (through the IRT and CTT frameworks) of the psychometric properties of physics test items with a view to identifying poor items and suggesting ways of improving their qualities. To guide the study, three research questions were answered. These are:

1) What are the item statistics (difficulty and discrimination indices) of the physics objective test using the CTT model and the 2-parameter model of IRT?

2) Which of the items are faulty on the basis of a) CTT Framework and b) 2-parameter model of IRT?

3) What are the faults inherent in each of the items? And what can be done to improve the quality of the faulty test items?

2. Methodology

Participants

Nine hundred senior secondary school three students (aged 16 – 18 +) who were drawn from 10 senior secondary schools in in Ibadan Educational Zone I, Oyo State, and Irewole Local Government Area, Osun State, Nigeria participated in the study. Among the 900 students who were sampled, 593 (65.9%) were boys and 307 (34.1%) were girls. Their ages ranged between 16 and 18+ years (mean age = 17.8 years; SD = 1.3).

Instruments

One instrument was used. This was 2012 WAEC Physics Objective Test Paper 2 [(PP2) Multiple Choice] for May/June candidates. The PP2 test items were developed by WAEC. PP2 consisted of 50 items and each item was placed on 4-option response format of A, B, C, and D. The test items covered the whole syllabus for senior secondary school physics as prescribed by the Federal Ministry of Education, Abuja, Nigeria.

Procedure

The PP2 was administered to physics students in the sampled school by 10 doctoral students, in the Institute of Education, University of Ibadan, who registered for the Advanced Test Theory. This formed part of their assessment in Couse 813 – Advanced Test Theory. During administration of the test, the physics teacher in each of the sampled school served as an assistant. The time allowed for the PP2 by WAEC was 90 minutes. Because the participants in this study have registered for senior secondary school certificate examination to be conducted in April/May by WAEC, they were therefore given 90 minutes for the test.

Method of Analysis

The item analysis was carried out by using both CTT and IRT frameworks. BILOG-MG (Window Version 3.0) with Marginal-Maximum Likelihood Estimation and Scoring Method was used. In this study, emphasis was on 2-parameter model.

The decision to use 2-PL model was borne out of the fact that in practical situations, according to Baker (2001), it 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. Although this feat can also be performed using the 3-parameter model 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).

3. Results

Research Question One: What are the item statistics (difficulty and discrimination indices) of the physics objective test using: a) CTT model and b) 2-parameter model of IRT?

Table 1 presents the item statistics of the physics object test. Colum 1 and 2 of the left hand of table 1 give the proportion of examinees who answered each item correct (difficult indices) and the values of the point biserial correlation (discrimination indices). These values give the fundamental CTT item statistics. Colum 3 and 4 of the right hand gives the discrimination (*a*) and difficulty (*b*) parameters of the 2-parameter IRT model.

Research Question 2: Which of the items are faulty on the basis of a) CTT Framework and b) 2-parameter model of IRT?

Classical Test Theory Framework

Using the CTT framework, for public examination, an item is considered faulty when the difficulty index is below 0.2 (very hard item) or above 0.70 (very easy item). More importantly, items with negative discrimination are considered faulty. On the basis of these indices the faulty items are:

Difficulty indices: Items 15, 28, 29, 35, 39, 44, and 49 (bold on the table). The difficulty indices are lower than 0.2. When the item difficulty indices of these items are multiplied by 100, the percentage gives the number of examinees who got the item correct. For example, for item 15, the difficulty index is 0.148 this shows that only 14.8% of the total examinees got the item correct. The most difficult item on the test is item 44. Only 8.4% of the total examinees got the item correct.

These items are poor because less than 20% of the examinees got each right.

Discrimination indices: Items 12, 16, 23, 25, 27, 28, 34, 36, 38, and 50. The discrimination indices are very low (negative). The low point-biserial implies that examinees who get these items correct tend to do poorly on the overall test (which indicates anomaly) and that examinees who get the item wrong tend to do well on the test (also an anomaly). This scenario indicates that these items are faulty.

Item Response Theory

Under the item response theory, as expressed in the preceding section, detection of the faulty items is not as straightforward as when CTT is used. Although one may consider difficulty and discriminating indices, the detection of poor items are better carried out using the information which each item contributes to the overall information supplied by the whole test. Figure 4 presents the Test Information Function.

Table 1: Item Statistics of CTT and IRT Models

Item Number	CTT Statistics (Examinees = 900)		IRT Statistics (Examinees = 900)	
	p	r_{pb}	b	a
1	0.644	0.108	-1.804	0.232
2	0.334	0.283	0.870	0.522
3	0.444	0.113	0.858	0.155
4	0.248	0.393	1.153	0.684
5	0.482	0.161	0.147	0.378
6	0.492	0.306	-0.005	0.672
7	0.559	0.304	-0.312	0.594
8	0.294	0.050	2.556	0.207
9	0.396	0.396	0.316	0.949
10	0.616	0.234	-0.730	0.434
11	0.298	0.230	1.310	0.421
12	0.343	-0.080	3.995	0.096
13	0.289	0.011	3.731	0.144
14	0.269	0.302	1.127	0.611
15	**0.148**	0.407	1.702	0.771
16	0.317	-0.066	5.619	0.081
17	0.450	0.063	0.834	0.143
18	0.392	0.217	0.779	0.350
19	0.392	0.107	1.159	0.229
20	0.330	0.130	1.593	0.273
21	0.299	0.169	1.570	0.340
22	0.201	0.260	2.011	0.448
23	0.290	-0.047	5.192	0.102
24	0.248	0.201	1.738	0.411
25	0.280	-0.065	6.578	0.085
26	0.223	0.073	4.724	0.158
27	0.244	-0.040	6.765	0.099
28	**0.186**	-0.071	6.281	0.140
29	**0.168**	0.175	3.482	0.284
30	0.184	0.324	1.874	0.536
31	0.544	0.276	-0.214	0.808
32	0.279	0.178	1.766	0.337
33	0.252	0.160	2.851	0.232
34	0.298	-0.115	6.635	0.076
35	**0.133**	0.396	2.031	0.658
36	0.282	-0.134	-------	-------
37	0.348	0.022	2.634	0.142
38	0.287	-0.087	6.419	0.084
39	**0.190**	0.194	2.388	0.388
40	0.466	0.253	0.183	0.438
41	0.256	0.237	1.757	0.388
42	0.354	0.136	1.245	0.297
43	0.297	0.307	1.002	0.585
44	**0.084**	0.137	4.234	0.354
45	0.368	0.125	1.379	0.239
46	0.291	0.299	1.155	0.511
47	0.268	0.260	1.398	0.472
48	0.252	0.094	3.484	0.188
49	**0.161**	0.314	2.199	0.501
50	0.217	-0.014	5.448	0.118

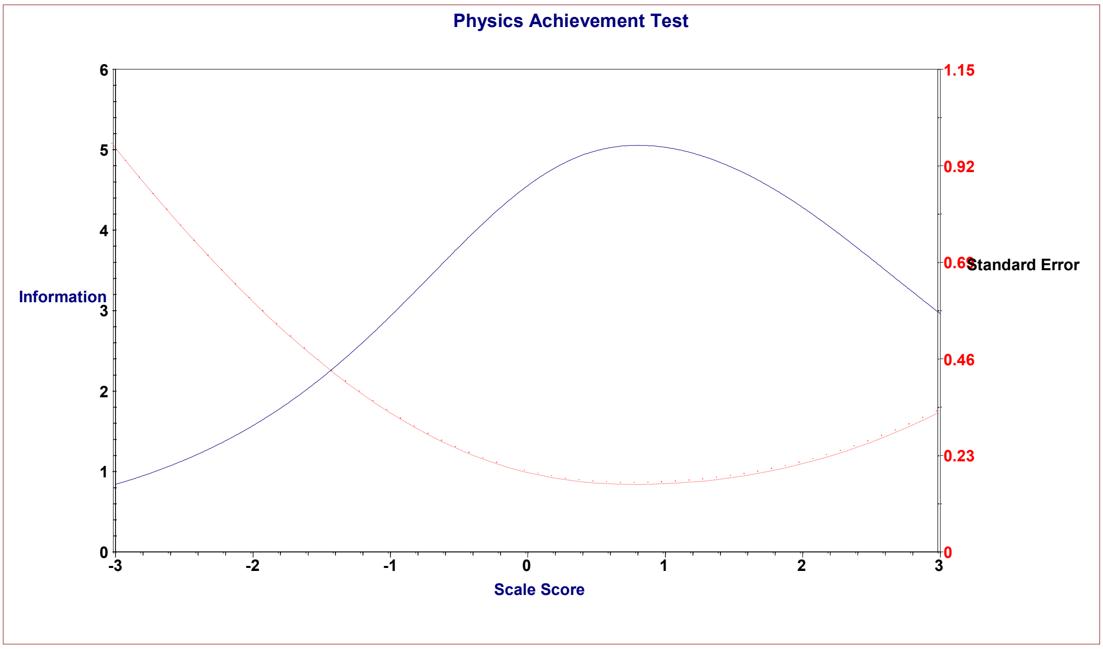


Figure 3: Test Information Function of the 50-item PAT.

The TIF shows that a precise estimate of the ability scale score where the test functions most is at Θ = 1.00 (where the solid line peaks).

Table 2 presents the item information values of each item at the scale score of 1.0, being the point at which the 50-item PP2 functions most.

Table 2: Item Information of each item at Θ = 1.00

Item	IIF	Decision	Item	IIF	Decision
1	0.03	NF	26	0.00	F
2	0.20	NF	27	0.00	F
3	0.01	NF	28	0.00	F
4	0.32	NF	29	0.04	NF
5	0.03	NF	30	0.14	NF
6	0.24	NF	31	0.18	NF
7	0.18	NF	32	0.08	NF
8	0.01	NF	33	0.02	NF
9	0.48	NF	34	0.00	F
10	0.10	NF	35	0.12	NF
11	0.10	NF	36	NC	F
12	0.00	F	37	0.00	F
13	0.00	F	38	0.00	F
14	0.24	NF	39	0.04	NF
15	0.30	NF	40	0.08	NF
16	0.00	F	41	0.08	NF
17	0.00	F	42	0.02	NF
18	0.03	NF	43	0.24	NF
19	0.02	NF	44	0.02	NF
20	0.02	NF	45	0.02	NF
21	0.18	NF	46	0.18	NF
22	0.10	NF	47	0.12	NF
23	0.00	NF	48	0.03	NF
24	0.10	NF	49	0.10	NF
25	0.00	F	50	0.00	F

Key: F = faulty item; NF = not faulty; NC = item not calibrated because item point-biserial value is less than -0.150

Faulty items on the basis of the amount of information contributed to the test information are 12, 13, 16, 17, 25, 26, 27, 28, 34, 36, 37, 38, and 50.

These faulty items contributed highly negligible information to the test information. For example item 12 contributed about 0.001 to the test information

Research Question 3: What are the faults inherent in each of the items? And what can be done to improve the quality of the test items?

Detail analyses of the identified bad items are now carried out to detect the faults inherent in them.

Item 12

When a body is slightly tilted, it is found that its centre of gravity is slightly raised. What is the state of equilibrium of the body?

A. Unstable
B. Stable
C. Neutral
D. Cannot be determined

The stem, the options and the supposed key of item 12 are confusing. In other words the stem of item 12 was not properly couched. The grammar of the stem should be restructured. More importantly, option D is very bad. This is because the focus of the item is for the students to state the type of equilibrium when the position of the centre of gravity of a body increases as a result of tilting the body. So using cannot be determined as an option is not appropriate.

Probably if the item of the test had been properly couched, the failure rate would have been minimal.

Suggestion:

The stem

When a body was slightly tilted, it was observed that its centre of gravity was slightly raised. What was the state of equilibrium of the body?

The options

In future the options should be written as

A. Unstable equilibrium
B. Stable equilibrium
C. Neutral
D. Cannot be determined

Item 25

Which of the following properties of waves is exclusive to transverse waves?

A. Reflection
B. Interference
C. Diffraction
D. Polarization

The key is D. From the CTT framework, the discriminating index is - 0.065. This shows that the item cannot discriminate well between high achievers and low achievers. What is wrong with this item? The options are good but the stem is not good enough. The word “exclusive” might have caused the problem. The use of words whose meaning may not be clear to the majority of the examinees must be avoided. All examinees who did not understand the word “exclusive” might have not answered the item correctly.

Suggestion

The item can be couched as “Which of the following properties of waves is exhibited by transverse waves only?”

Item 27

A ray of light passes from air to water to glass to air. Given that the refractive index for light passing from air to water is $^4/3$ and air to glass is $^3/2$, calculate the refractive index of glass relative to water.

A. 0.50
B. 0.67
C. 0.75
D. 1.13

From the IRT framework, for a candidate to get the item correct, he or she must have ability scale score of 6.765. This is a very difficult item. A thorough examination of the item, however, shows that the stem and the options are okay. It is application item. The results suggest that majority of the students did not do well on the item. Physics teachers need to explain this concept properly.

Item 28

An object is placed at a point X between the focal point F and the optical centre C of a diverging lens. If F' is the focal point on the other side of the lens, the image of the object is formed between

A. F and X.
B. X and C.
C. C and F'.
D. F' and 2F'

Although the stem looks properly written, it would have been better if the second sentence is given in question form. Rather than "If F' is the focal point on the other side of the lens, the image of the object is formed between" it would have been more appropriate if it is written as "If F' is the focal point on the other side of the lens, at what point will the image of the object be formed?

However, I strongly believe that the options are good. This result suggests that many of the students do not really understand the concept of converging and diverging lens. Physics teachers in schools should make these topics clearer to their students. More importantly, the differences in the image as well as the point where images are formed should be made clearer to students.

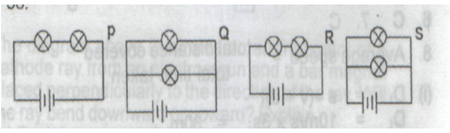
Item 36

Which of the following properties is an advantage of a lead-acid accumulator over an alkaline accumulator?

A. Possesses low internal resistance
B. Can be recharged
C. Has shorter life span
D. Possesses higher emf

The stem of this item is alright. However, the options are problematic. The key is option A, however, option D is also a possible answer. This item has more than one option.

Item 38



Two similar cells are used to light two similar lamps as illustrated in the diagrams above. In which of the circuit diagrams are the lamps brightest?

A. P
B. Q
C. R
D. S

Item 38 tests students' ability to apply knowledge. For a student to get the item correct, he or she needs to know that more current is generated from cells connected in series than cells connected in parallel. Moreover, brighter lights are given by lamps that are connected in parallel than lamps that are connected in series. Therefore among the four circuits, lamps in circuit Q will be brightest.

Analysis of the result shows that the number of the low achievers who got the item is more than the number of high achievers. What is wrong with the item? One, in the diagram, two forms of connection of cells was presented. In diagrams P and Q, cells were connected in series, whereas in diagrams R and S cells were connected in parallel. A major fault in the diagram was that the cells were not labelled and this might have created confusion for the students. Although the symbols for the cells were used, inclusion of E_1 and E_2 in the diagrams to represent the cells might have eliminated the confusion

The item is fairly okay in terms of the stem and the options, however, it appears most of the students did not understand the concepts of combining cells and lamps in series and in parallel. This result points out that Physics teachers need to explain and clarify issues about parallel and series combination of cells, lamps and resistors.

Item 44

An inductor is connected to a 24V, 50 Hz mains supply. If the current through the inductor is 1.5A, calculate the inductance of the inductor ($\pi = {}^{22}/_{7}$)

A. 8.0 X 10^2H
B. 7.5 X 10^0H
C. 1.6 x 10^0H
D. 5.1 x 10^{-2}H

From the analysis, out of the 900 students who were examined only 76 (8.4%) students picked the correct option (D). This shows that it is a very difficult item. In fact it is the most difficult item among the PP2 test items. From IRT framework, the difficulty parameter is 4.234.

A thorough examination of the item shows that the item is well structured and the options are okay. A major problem is that this item comes from topics that are usually treated last in senior secondary school physics.

Moreover many physics teachers do not usually treat this topic and when taught the topic is usually not properly taught. This points out that physics teachers should try as much as possible to teach this concept early enough (preferably during the first term of SS3 school year).

Item 50
Which of the following radiation emitted in radioactive decay has momentum, a fairly high penetrating power and is deflected by a magnetic?
A. Alpha particle
B. Beta particle
C. Gamma radiation
D. X-radiation
For this item, the discriminating index is negative (- 0.014). This suggests that the item may be faulty. Thorough analysis of the options shows that the item has more than one correct answer. Both options A and B are correct. Items such as these should be avoided. An item that has more than one correct answer is unlikely to discriminate well.

4. Discussion and Conclusion

Results of this study have shown that some of the items contained in PP2 were either too difficult were not good enough to discriminate between who were high achievers and low achievers. From the IRT framework, some items have difficulty parameter whose values are very high. This indicates that they are hard items. According to De vellis (1991) and Hambleton and Jones (1993) such items though may be perfectly alright but having many of them in a test may not be very okay for public examining bodies whose focus should be on norm-referencing.

Most norm-referenced achievement tests ought to be designed in such a way as to differentiate examinees with regard to their competences in the measured areas: That is, the test is designed to yield a broad range of scores maximising discriminations among all examinees taking the test. When a test is designed for this purpose, items are generally chosen to have a medium level and narrow range of difficulty. Therefore, public examining bodies ought to key into this.

In this study using CTT framework, and with p values set such that $p < 0.200$ as being the optimum difficulty level, seven items were considered as being difficult items. Under the IRT framework, items whose difficulty parameters were more than +3.000 were considered very difficult items. Using this parameter, 13 items were considered as being hard. For secondary school students whose population distribution is assumed to be normal in shape, these levels of difficulty are high.

Using the discriminating indices of the CTT framework, some of the items were found to have negative discriminating values. Certainly a negative discrimination value is suggesting that we should look carefully at the item to see why the better students should have more trouble with it than the weaker students. This suggests that these items could not discriminate between high achievers and low achievers. According to Field (2006), Varma (2014) those who got these items might have probably guessed before they got it right.

5. Implications of Findings

The implications of the findings are many. One, it is probable that public examining bodies may not be thorough in the analysis of the trial testing of these items before they are finally administered to candidates. Two, physics teachers may not be doing enough in the classroom. Therefore, I wish to suggest that public examining bodies should endeavour to carry thorough analysis of their test items. Items that are faulty from the low values of difficulty and discriminating indices should be revised. More importantly information gathered from the analysis should be passed to physics teachers. Physics teacher in should use such information to improve the quality of teaching. Emphasis should be on clarifying confusions that may likely arise in some concepts in physics. Therefore, there is the need for physics teachers to properly explain confusing concepts in physics.

References

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

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park: Sage Publications.

Emeke, E. A. & Adegoke, B. A. (2006). Determinants of students' cognitive achievement in senior secondary school physics: How important is test response mode? African Journal of Educational Research, 10, 1 & 2, pp. 25 – 29.

Field, A. (2006). *Research Methods II: Reliability Analysis.* Retrieved August 5, 2006 from The University of Sussex Web site http://www.sussex.ac.uk/Users/andyf/reliability.pdf

Haladyna. T. M. (1999). *Developing and validating multiple-choice exam items, 2nd ed*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Varma, S. (2014). Preliminary item statistics using point-biserial correlation and p-values. Morgan Hill, CA: Educational Data Services, Inc. Available on line: http://www.eddata.com

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 was born on the 9th September, 1961. He holds Bachelor's Degree in Education and Physics, Master's and Doctoral Degrees in Educational Evaluation. He had taught Physics and Mathematics at the Secondary School Level for 19 years before he joined the Institute of Education, University of Ibadan, in 2015, as a Research Fellow. He is now a Senior Research Fellow. He teaches Statistics and Research Methods as well as Advanced test theory to higher degree students. His current research focus is on tests and measurements.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar