International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Item Analysis using a Derived Science Achievement Test Data

Ado Abdu Bichi

Department of Arts and Social Sciences Education, Faculty of Education, Northwest University, Kano, Kabuga-Katsina Road, PMB 3220 Kano Nigeria

Abstract*Item analysis is useful in both the development, evaluation of assessments tools and in computing standardized measures of student performance. In item analysis, item statistics (difficulty, discrimination) for each item or question provide a means of assessing the quality of the test items. This study demonstrates the use of the classical item analysis to evaluate the quality of multiple choices Chemistry test items used in Kano state qualifying examination in July 2014. Forty items Chemistry test administered to students at the end of their senior secondary school two (SSII) was subjected to the classical analysis using sample of 530 students, the Statistical Packages for the Social Sciences, version 20 (SPSS 20V) was used to determine the discrimination and difficulty indices of the items and the classification of the items according to their item characteristics. Findings revealed that, out of the 40 items in the test, 12 (30%) items failed to meet the set criteria of item quality and are therefore needs to be revise or improve for further administration. 28 items based on the established standards has been considered as 'good' items. Similarly the results indicate a significant positive correlation between item difficulty and item discrimination indices. It is recommended that, the teacher made Chemistry achievement tests use to examine sciences secondary school students' achievement should be made to pass through all the processes of standardization and validation by conducting psychometric analysis to improve their quality.*

Keywords: Item Analysis, Correlation, Chemistry, Item difficulty, Item discrimination

1. Introduction

Assessment of students is basic to teaching. It is the most important aspect of teaching and learning process, the major objective of teaching is to effect positive changes in students' behaviour. Achievement test is use as a tool to assess the magnitude of these changes. Some teachers use formal measuring tools while others use their subjective impressions developed through their daily encounter with the students. Test is regarded as the most popularly used technique for obtaining information in the school system [31]. [27] described a test task presents a situation that makes it possible to elicit behaviour or performance of individuals and through it determine their knowledge, abilities, skills or feelings. According to [30], test enables teachers and others using it to systematically data for the purpose of making comparisons across individuals, classes, schools, districts or countries. Whereas the teacher made tests developed and used in classroom instruction by the teacher are more popularly with the learners, standardized tests serve certification, quality control and benchmarking purposes [28; 29]

The practice of testing has become increasingly common and a reliance on information gained from test scores has made an indelible mark on our culture. Educational institutions place a high reliance on test performance to make decisions and ensure standards. Recent concern on the proper design, production, administration, analysis, reporting and interpretation of tests has placed an increasing demand on test developers. Obviously, test quality is a high priority for those who make tests, those who take tests, and those who rely on test scores for decision making. Now more than ever, it is critical that tests are efficient and effective at measuring ability and those scores are reliable and precise measures of examinee ability. Criteria used to establish test quality generally focus on the areas of test design, test analysis techniques and test score interpretation. Quality test design is impacted by many elements including format, length, administration procedures, construction, validity and scoring schema [11]. The nature and the quality of information gathered from the achievement test can control the educational development efforts and direct the instruction [34].

The most important characteristics of an achievement test used in assessing students' abilities are its reliability and content validity. [33] for a test to be reliable and valid, a systematic selection of test items with regard to subject content and degree of difficulty is necessary. Moreover, the reliability of the test also depends upon the grading consistency and discrimination between the students of different performance levels. Thus the quality and effectiveness of a test depends upon the individual item. To determine the quality of individual item, item analysis is done after the administration and scoring of the preliminary draft of the test on the selected sample.

According to [33] the two purposes of Item analysis are; firstly, to identify defective test items and secondly, to indicate the areas where the learners have or have not mastered. [33] further stated that, Item analysis measures the effectiveness of individual test item in terms of its difficulty level and its discrimination power i.e to distinguish between high and low achievers in a test. Thus it Item analysis helps in selecting the best test items in the final draft by retaining the good and rejecting poor test items. Similarly, it shows the need to review and modify the items in a test.

According to [24] a major concern in test construction is ensuring the reliability of test items, and one typical step in investigating reliability has been Classical Test Theory (CTT) item analysis. The classical item analysis essentially determines test homogeneity. That is, the more similar the items in a given test, the more likely they measure the same

kind of intended ability and therefore, the higher the reliability [4]. However, CT item analyses statistics do not provide the necessary information on how examinees at different ability levels on the latent trait measured have performed on an item [3]. Therefore, a more robust statistics based on modern test theory has also been widely used in more recent test reliability investigations.

1.2. An Overview of Classical Test Theory (CTT)

Since the focus of this study is on item analysis based on CTT, it is important to explore the basic ideas involved in order to fully understand the approach. Classical test theory has been used for decades to determine reliability and other characteristics of measurement instruments. According to [41] CTT is a theory about test scores and focus is on three forms of scores - (i) test score (often called the observed score), (ii) true score, and (iii) error score. [3] emphasised that, CTT attempts to explain measurement errors. In classical test theory, the model of measurement error is based on the correlation coefficient. The correlation coefficient, developed by Charles Spearman, attempts to explain error using two components: a true correlation and an observed correlation.

In CTT the number of correct score is often taken as ability. There are two general factors in measurement while using CTT approach, an observed response (X) i.e. scored obtained by the students on a particular task and a true ability (T) which is the real potential in a student. This relationship in theoretical model known as "Classical Test Model" can be written as;

$$X=T+E$$

Where, E is random error of measurement.

This is a simple linear model that links the observable test score(X) to the sum of two unobservable variables, true score (T) and error score (E).

In the words of [12] "The CTT model is based on the notion that the observed score that test takers obtain from a test is composed of a theoretical un-measurable true score plus some measurement error" as expressed in the model above.

1.2.1. Assumptions of CTT

The major assumptions underlines the CTT 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 the parallel tests are uncorrelated.

According to [17] test analysis using CTT is fairly straight-forward. The most common test score for a test developed using CTT is the number (or percent) of items answered correctly. From purely statistical considerations, test construction using CTT might often consist of selecting those items with the best discrimination (item-total correlation) and which span a range of item difficulties. According to [15] the assumption of classical test theory is that, each individual has a true score which would be obtained if there were no errors in measurement. However, because measuring instruments are imperfect, the score observed for each individual may differ from an individual's true ability. The difference between the observed test score and the true score results from measurement error. Error is often assumed to be a random variable having a normal distribution. Classical test theory (CTT) was the dominant approach until 1953 when Frederic Lord published his doctoral dissertation on Latent Trait Theory. While CTT models test outcomes based on the linear relationship between true and observed score (Observed score = True Score + Error), IRT models the probability of a response pattern of an examinee as a function of the person's ability and the characteristics of the items in a test or survey [43]. Theoretically CTT is simple and easy to apply that is why its test statistics are still commonly used in test construction process, however many researchers e.g. [42] have questioned their utility in the modern era. This is because item statistics such as difficulty and discrimination indices are sample dependent.

1.2.2. The concept of Item Analysis

Item analysis is a process which examines student responses to individual test items in order to assess the quality of those items as well as the quality of the test as a whole [25]. Item analysis enables instructors to increase their test construction skills, identify specific areas of course content which need greater emphasis or clarity, and improve other classroom practices. According to [12]"Item analysis broadly refers to the specific methods used to evaluate items on a test, both qualitatively and quantitatively, for the purpose of evaluating the quality of individual items".He further stated that, *t*he goal is to help its developers to improve the instrument by revising or discarding items that do not meet a minimally acceptable standard. The qualitative review is essential during item development and involves experts who have a mastery of relevant material. Test review boards and content experts cannot always be equipped with the knowledge they require to identify "bad" or "defective" items because of such factors as the multidisciplinary nature of the test content and the demographic characteristics of test takers. The statistical analysis could help to identify problematic items that may have slipped the experts' attention, one way or the other. Thus, the quantitative analysis is conducted after the test/tool has been administered to the test takers. The objectives of both the qualitative and quantitative assessments remain the same – to assess the quality of items. According to [21], "Item analysis investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test." For example, when norm-referenced tests (NRTs) are developed for instructional purposes, such as placement test, or to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses. Similarly, criterion-referenced tests (CRTs) compare students' performance to some pre-established criteria or objectives [25]. Some of the researchers that have contributed immensely to the theory of test item analysis are Galton, Pearson, Spearman, and Thorndike. Generally an item in a test may fail to meet the minimum quality standard, whatever the set standard is. It may be as a result of: (1) the flaws in the question and (2) the flaws in the instruction of the content [12].

1.2.3. Classical Test Item Analysis Statistics

The scope of this paper will only cover the two major statistics of Item difficulty and discrimination using Classical Test Approach as well as the reliability of the test and to find out the relationship between the two item statistics, considering [22] opinion that, problematic items with low point biserial correlation (item discrimination value) may show higher p-value (item difficulty) and that p-value should not be taken as indicative of item quality, only point biserial correlation (item discrimination value) should be used to judge item quality.

Item analysis involved statistics that help in analysing the effectiveness of the items and improving test items or questions. These statistics can provide useful information to determine the validity and accuracy of an item in describing learners or examinees ability from their response to each of the item in a test. The common classical test item analysis statistics are ***(a)*** Item difficulty ***(b)*** Item discrimination ***(c)*** Distractor analysis and ***(d)*** Reliability. This paper therefore will only cover the two major statistics of Item difficulty and discrimination as well as the reliability of the test

a. Item Difficulty (item level statistic)

Item difficulty in classical theory is the first item characteristic to be determined. Item difficulty is a measure of the difficulty of an item [25]. It is simply the proportion of examinees taking the test, who got the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be. To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. An item answered correctly by 85% of the examinees would have an item difficulty, or ***p*** value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or ***p*** value of 0.50 [16].

The item difficulty is denoted as ***p*** and is symbolically given as;

$$p = \frac{R}{N}$$

Where ***P*** = is the difficulty of a certain item. ***R***= is the number of examinees who get that item correct and ***N***= is the total number of examinees. A general guideline for the interpretation of an item difficulty index values is provided in the following table; see, for example, [1]and[24] among others

Table 1: Interpretation of Item difficulty Index [39]

Difficulty Index (p)	Interpretation
$P \leq 0.30$	Difficult
$0.31 \leq 0.70$	Moderately difficult
$P > 0.70$	Easy

b. Item Discrimination (item level statistic)

Item discrimination refers to the percentage difference in correct responses between the low and the high scoring students. It is the ability of an item to discriminate between higher ability examinees and lower ability examinees [1]. Item discrimination statistics focus not on how many people correctly answer an item, but on whether the *correct* people get the item right or wrong. In essence, the goal of an item discrimination statistics is to eliminate items that do not function as expected in the tested group. [2] Two indices can be computed to determine the discriminating power of an item, the item discrimination index, ***D***, and discrimination coefficients.

i. Item discrimination index, D

The method of extreme groups can be applied to compute a very simple measure of the discriminating power of a test item [16]. In computing the discrimination index, ***D***, first score each student's test and rank order the test scores. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. [23]stated that "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis" The discrimination index, ***D***, is given as

$$D = Pu - Pi$$

WhereP_u the proportion of correct responses for the upper group and P_i the proportion of correct responses for the lower group. Since its proportion ranges 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 [2]

ii. Discrimination coefficients.

Two indicators of the item's discrimination effectiveness are point biserial correlation and biserial correlation coefficient. The choice of correlation depends upon what kind of question we want to answer. One of the major shortcomings of the discrimination index, *D* is that, only 54% (27% upper + 27% lower) are used to compute the item discrimination ignoring the 46% of the examinees, however the advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients.

A *point biserial correlation coefficient* (r_{pbi}) is defined by:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

Where, **Mp**= whole-test mean for students answering item correctly (i.e., those coded as 1s), ***Mq*** = whole-test mean for students answering item incorrectly (i.e., those coded as 0s), St = standard deviation for whole test, ***p*** = proportion of students answering correctly (i.e., those coded as 1s), and ***q*** = proportion of students answering incorrectly (i.e., those coded as 0s) [8].

Point biserial correlation (*r_pbi*) ranges from -1 to +1. A high point-biserial coefficient means that students with higher total scores are students selecting the correct response, and students selecting incorrect responses to an item are associated with lower total scores. According to the value of *r_pbi*, item can discriminate between high-ability and low-ability examinees. Very low or negative point-biserial coefficients help to identify defective items [7].

A summary of the widely used [40] criteria and guidelines for categorizing discrimination indices in item and test analysis is used in this study.

Table 2: Interpretation of Discrimination Index [40]

Item Discrimination	Quality of an Item
$D \geq 0.40$	Item is functioning quite satisfactorily
$0.30 \leq D \leq 0.39$	Good item; little or no revision is required

$0.20 \leq D \leq 0.29$	Item is marginal and need revision
$D \leq 0.19$	Poor item; should be eliminated or completely revised

c. Item Reliability (Test level)

There are different means of estimating the reliability of any measure [34]. The statistic that measures the test reliability of inter-item consistency is called the internal consistency reliability coefficient of the test; it refers to the degree to which the items that make up the concept of interest are measuring the same underlying concept. There are different internal consistency measures that can be used[12]. For a test, having multiple-choice items that are scored correct or incorrect, and that is administered only once, the most commonly used method for estimating internal consistency reliability is Cronbach's Alpha and other like Kuder-Richardson formula 20 (KR-20)

1.3. The Problem

In Kano State, Nigeria, the state government usually conducts qualifying examination for all senior secondary school two (SSS II) students, to assess their suitability for sponsorship to write the final examinations being conducted by the two public examining bodies in Nigeria [i.e. West Africa Examination Council (WAEC) and National Examination Council (NECO)]. Chemistry is one of the major science subjects taught at the Senior Secondary School level. The students need to pass the subject at the end of their secondary education at credit level to fulfil the requirements for admission to study natural and physical sciences at the higher institutions of learning in Nigeria. Chemistry is one of the three core science subjects used in the state SS II qualifying examination.

Although the qualifying examination is used as a criteria to sponsor students to write their final examination conducted by WAEC and NECO. These two public examination bodies carried out item analysis of their examination items. However, there has been no evidence that the test items used in qualifying examination scaled through item analysis in selecting the appropriate items. To the point of this paper the non-standardisation of the qualifying examination items may be responsible for the mass failure experience in the outcome of final results of students conducted by the public examination. So just how "good" are the Multiple-Choice Chemistry achievement tests items? How effective are the individual Multiple-Choice Chemistry achievement tests items in predicting the students' overall performance in the whole Chemistry test paper?

1.4. Objective of the Study

The major focus of this study is to analyse the Chemistry achievement test items using CTT frameworks to know the psychometric properties of the test items with a view to identify problematic items and suggest ways of improving their qualities and to find out whether any of the two item statistics can be used to judge the item quality.

1.5. Research Questions

Three research questions were generated to guide the study. These are:

1) What are the item statistics (difficulty and discrimination indices) of the Chemistry Achievement Test?
2) ow many items survived after item analysis on the basis of their difficulty and discrimination values?
3) s there any correlation between item difficulty and item discrimination indices of the Chemistry achievement test?

2. Methodology

2.1. Design

Survey design was employed to collect the relevant data for the study.

2.2. Participants

According to [26], stable estimates of CTT item difficulty and discrimination can be found with a sample size of 150 to 200. Therefore, two hundred (200) senior secondary school two (SSS II) students, age (16-18) from five (5) secondary schools in Kano State, Nigerian participated in the study. The participant consisted of 116 boys (58%) 84 girls (42%).

2.3. Research Instrument

Chemistry Achievement Test (CAT) designed and constructed for assessing the suitability of SSII for government sponsorship to write the final examination conducted by the two examination bodies was adopted in this study. The CAT comprises 40 multiple-choice items with five answer choices/options (A-E). The test items covered the whole senior secondary school Chemistry syllabus prepared for SSCE by WAEC and NECO as well as the Chemistry curriculum developed by the Federal Ministry of Education in Nigeria.

2.4. Data Collection Procedure

The 40 multiple-choice items Chemistry Achievement Test was administered on the sample after receiving specific instruction for the test by the researcher and the subjects' teachers in the sampled schools in July 2014. The test items were scored dichotomously as either correct or incorrect, with correct answer as 1 and incorrect 0.

2.5. Data Analysis

After scoring the data dichotomously, the item analysis of the CAT was carried out. The two psychometric properties of the items (item difficulty and item discrimination indices) were determined. The item difficulty index is calculated as percentage of the total number of correct responses to the test item, using the formula P=R/N, where P is the item difficulty index, R is the number of correct responses and N is the total number of responses, items mean were used. Item discrimination indices was determined using the point biserial correlation coefficient (rpbi) and Cronbach's alpha as a measure of internal consistency reliability of the test were determined by using the Statistical Package for the

Social Sciences (SPSS 20V). Similarly, Pearson Product Moment Coefficient (r) was used to determine the correlation between the two indices at 0.05 level of significance as well as the scatter plot of the relationship.

3. Results and Discussion

Table 3: Summary of Test Statistics

Variables	Test Items
Number of Items	40
Number of Examinees	200
Reliability ***(Alpha)***	.820
Mean Scores	16.67
Standard Deviation	6.66
Mean ***P***	0.420
Mean r_{pbi}	0.290

Table 3 presents the chemistry achievement test summary statistics. The total number of items in the test is forty (40) and the number of students who sat for the test is 200 as presented in the second column of the table. The overall reliability of the test as measured by the Cronbach's Alpha is 0.820, this shows that the test is reliable since the coefficient is high (0.820) greater than the [18] recommended acceptable value of 0.70. Similarly the items mean scores is 16.67 with standard deviation of 6.66. The mean item difficulty (p) is 0.420 and mean item discrimination of the test (r_{pbi})is also 0.290 as presented

Research Question 1:*What are the item statistics (difficulty and discrimination indices) of the Chemistry achievement test?* Table 4 presents the Classical Test Theory model item parameters of the Chemistry Achievement Test. Number of the items in the achievement test (i.e. items 1-40) and the item statistics of Chemistry Achievement Test; Item difficulty (*p*) and Item discrimination (r_{pbi}) generated by using the SPSS.

Table 4: Item Statistics of Chemistry Achievement Test

Item	***(P)***	***(r_{pbi})***	***Item***	***(P)***	***(r_{pbi})***
1	0.32	***0.28***	21	0.44	**0.22**
2	0.72	0.42	22	0.35	0.58
3	0.72	0.36	23	0.46	0.32
4	0.46	***0.23***	24	0.37	0.33
5	0.55	0.32	25	**0.29**	0.37
6	0.44	0.43	26	**0.29**	**0.27**
7	0.34	0.32	27	0.35	**0.23**
8	0.44	0.42	28	0.35	**0.28**
9	**0.19**	**0.01**	29	0.33	0.45
10	0.58	**0.27**	30	**0.20**	***-0.30***
11	0.67	0.42	31	0.52	0.46
12	0.43	0.36	32	**0.23**	***0.05***
13	0.49	0.48	33	0.68	0.53
14	**0.26**	**-0.25**	34	**0.30**	**0.24**
15	0.42	0.37	35	**0.18**	***0.05***
16	0.51	0.55	36	0.55	0.43
17	**0.27**	***-0.02***	37	0.61	**0.29**
18	**0.21**	***0.02***	38	0.45	0.40
19	0.55	0.41	39	0.56	**0.27**
20	0.41	0.38	40	**0.29**	***0.18***

****** Items of special interest are in bold***

NB: P *Item difficulty and* r_{pbi} *Item discrimination*

Research Question 2: *How many items survived after item analysis on the basis of their difficulty and discrimination values?*

Table 5: Distribution of Items based on Difficulty index

Item Difficulty Index (p)	***Total Item***
Easy (*P*> 0.70)	*2 (5%)*
Moderately (0.31 ≤ 0.70)	28 (70%)
Difficult (*P* ≤ 0.30)	10 (25%)

Based on the set standards for interpreting difficulty indices 28 (70%) of the Items were of moderate difficulty, 2(5%) were easy, and 10(25%) were considered difficult. With this rule, 10 items are difficult and can be considered 'poor' or 'faulty' items. In conformity with the rule, 28 out of the 40 items are "good" (moderately difficult) and 2 items can be seen as "fair" (easy). On the basis of the item selection criteria of difficulty indices of (*0.30>P>.070*), 12 items that failed to satisfy the condition are considered 'poor' items (i.e items 2,3,9,14,17,18,25,26,30,32,35,40)

Table 6: Distribution based on Discrimination indices

Discrimination Coefficient	***Total Item***
Very Good (D ≥ 0.40)	13(32.5%)
Reasonably Good (0.30 – 0.39)	9 (22.5%)
Marginal (0.20-0.29)	10 (25%)
Poor (D ≤ 0.19)	8 (20%)

On the basis of discriminating index criteria set, the results indicates that 8 (20%) of the items failed to differentiate between students of different abilities, 10 (25%) items are marginal need to be reviewed, 9 (22.5%) of the items are satisfactory and 13 (32.5%) of the items functions very well. based on the selection criteria of discriminating index (i.e. $r_{pbi} \leq 0.20$), 8 items are 'poor' and failed to satisfy the condition the items can be eliminated completely from the test.

Research Question 3:*Is there any correlation between item difficulty and item discrimination indices of the Chemistry achievement test?*

Table 7: Result of correlation between item difficulty and item discrimination values

Item Statistics	**N**	**Mean**	**S.D**	**r-caldf**	**p**
Item Difficult Index	40	0.420	0.14889		
				0.60* 38	0.00
Item Discrimination	40	0.290	0.19487		

****. Correlation is significant at the 0.01 level (2-tailed).**

Information from table 7 above indicates a significant positive correlation r(38)=0.630, P=0.00 (P<0.05) between item difficulty index and item discrimination value. This shows that the test items on average have 0.420 level of difficulty and 0.290 discrimination values.

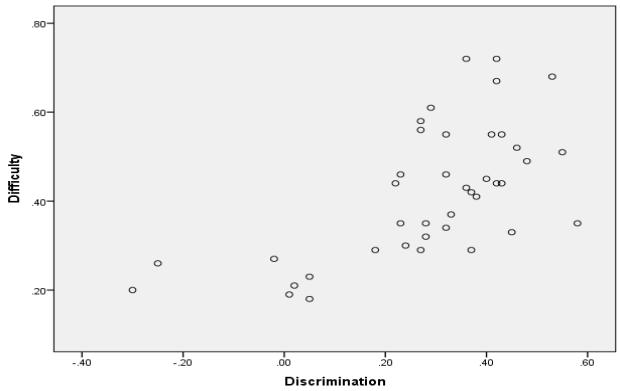

Figure 1: Scatter plot showing relationship between difficulty index and discrimination value of items

Figure 1 provides scatter plot showing correlation between the two variables. Based on this result it can be concluded that there is a significant positive correlation between the two variables (r = 0.630, p = 0.00) with an increasing value of difficulty index, there is also an increase in discrimination index

4. Discussion

The focus of this study is to evaluate the quality of the science achievement test used in assessing students' abilities and to find out whether any of the two item statistics can be used to judge the item quality.

The findings reveals that based on the established standards 28 (70%) of the Items were of acceptable difficulty level, i.e ($0.31 \leq 0.70$). 2(5%) were easy, and 10(25%) were difficult. This finding is consistent with the findings of [36] and [33] whose findings revealed that, majority (75%) and (78%) of the items respectively, was acceptable as far as difficulty was concern. On the basis of item discrimination indices, the results indicates that 8 (20%) of the items failed to differentiate between students of different abilities, 10 (25%) items are marginal need to be reviewed, 9 (22.5%) of the items are satisfactory and 13 (32.5%) of the items functions very well. Considering the [40] set criteria of item selection based on its discriminating index (i.e. rpbi $\leq$ 0.20), 8 items are 'poor' and failed to satisfy the condition the items can be eliminated completely from the test. This denotes that 80% of the test items are in the range of good and very good acceptable discrimination level. This study is also in agreement with that of [36] whose study on evaluating the quality of multiple choice questions (MCQs) in formative examination in Physiology revealed having 75% of the items within acceptable to excellent discrimination.

Considering [22] opinion that, p-value (item difficulty) should not be taken as indicative of item quality, only point biserial correlation (item discrimination value) should be used to judge item quality. Similarly experience has shown that a 'good' item has point biserial correlation above 0.25. The study therefore, concludes that, 12 (30%) of the items (i.e 9, 14, 17, 18, 21, 26, 27, 30, 32, 34, 35, 40) were considered poor and need to be reviewed for further administration. This poor performance of these 12(30%) of items and the students could have been due to poor understanding of difficult topics, ambiguity in wordings of the questions or even inappropriate key, it may also be due to personal variations in students' intelligence level. This findings is consistent with the findings of several studies example; [38], [36] and [33]whose found majority of the items (i.e more than 50%) to be within acceptable level of item difficulty and discrimination.

Similarly, the [22] opinion on emphasis on item discrimination in judging the quality of items, necessitate the examination of the relationship between the two parameters of items difficulty and discrimination to see whether the two parameters provide the same estimates and can be used interchangeably in deciding the item quality or not. The result of the test revealed is a significant positive correlation between the two variables (r = 0.630, p = 0.00) with an increasing value of difficulty index, there is also an increase in discrimination index. This finding agrees with the findings of [36] who have found a slightly significant positive correlation. However the findings disagree with that of [33] and [35] whose findings revealed a moderate negative relationship between the two indices of discrimination and difficulty (r-0.3711) and (r= -0.325) respectively. Similarly, [37] found non-linear the correlation between the difficulty index and discrimination index.

5. Conclusion and Recommendations

Findings of this study emphasises the important role that item analysis play in determining the quality assessment tools especially during test construction as well as validation. The study has been able to establish that an individual item in a test with moderate difficulty and a good positive discrimination power are ideal for a good test. However, an items having zero or negative discrimination power with very low or high difficulty estimates should be completely revise, improve or out rightly rejected. Item analysis results that are generated may be influenced by many factors which include examinees having poor understanding of difficult topics, ambiguity in wordings of the questions or even inappropriate key, instructional procedure applied, it may also be due to personal variations in students' intelligence level. Going by the significant role played by item analysis in evaluating and improving assessment tools or instruments, it is recommended that;

Item analysis should be maintained in test development and evaluation, because of its importance in the investigation of reliability and in minimizing measurement errors.

Secondly, the teacher made Chemistry achievement tests that are used to examine students' achievement compared to educational standards and to assess their suitability for government sponsorship to write the final examination should be made to pass through all the processes of standardization and validation.

Thirdly, any of the two item statistics can be used to judge the quality of the items i.e item difficulty and discrimination indices since the two indices produce almost the same item characteristics.

Lastly, training on test development and construction should be regularly organised for teachers to be more skilful in test construction, marking and grading of students scripts.

References

[1] B. A. Adegoke "Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks".Journal of Education and Practice, Vol.4, (22), 2013.
[2] Courville, T. R. "An empirical comparison of item response theory and classical test theory item/person statistics". Unpublished Doctoral Thesis, Texas A & M University, 2004.
[3] Crocker, L. and J. Algina, Introduction to classical & modern test theory. Holt, Rinehart and Winston, New York, 1986.
[4] A. Davies, Principles of Language Testing, Basil Blackwell Ltd, Cambridge Oxford, 1990.
[5] R.L. Ebel, Essentials of educational measurement (3rd Ed), Prentice Hall, Englewood Cliffs, NJ, 1979.
[6] C. Elvin, Test Item Analysis Using Microsoft Excel Spread-sheet Program. The Language Teacher, 27 (11), pp.13-18, 2003.
[7] M. Erguven, "Two approaches to psychometric process: Classical test theory and item response theory". Journal of Education, 2(2), pp.23-30, 2014.
[8] N. E. Gronlund, and R. L. Linn, Measurement and evaluation in teaching (6th Ed). MacMillan, New York, 1990.
[9] H. Gulliksen, Theory of Mental Tests. Erlbaum, Hillsdale, NJ, 1987.
[10] A. Hotiu, "The Relationship between Item Difficulty and Discrimination Indices in Multiple-Choice Tests in a Physical Science Course". Master in Science Thesis, Charles Schmidt College of Science. Florida Atlantic University, Boca Raton, Florida, 2006.
[11] T. L. Kinsey, "A Comparison of IRT and RASCH Procedures in a Mixed-Item Format Test" Unpublished Doctoral Thesis, University of North Texas, 2003.
[12] Krishnan, V. "The early child Development Instruments (EDI): An Item Analysis using Classical Test Theory (CTT) on Alberta's Data". Early Child Development Mapping (ECMap) Project Alberta, Community-University Partnership (CUP), Faculty of Extension, University of Alberta, Edmonton, Alberta, 2013.
[13] F. M Lord, Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Inc. New Jersey, 1980.
[14] Magis, D. "Influence, Information and Item Response Theory in Discrete Data Analysis", 2007. (Online) available at http://bictel.ulg.ac.be/ETD-db/collection/available/ULgetd-06122007-100147/ [accessed on July 24, 2014]
[15] C. Magno, "Demonstrating the Difference between Classical Test Theory and Item Response Theory using Derived Test Data" The International Journal of Educational and Psychological Assessment, Vol. (1) Issue 1. Pp. 1-11, 2009.
[16] S. Matlock-Hetzel, "Basic Concepts in Item and Test Analysis", 1997, (Online) available at files.eric.ed.gov/fulltext/ED406441.pdf [accessed on June 24, 2014]
[17] A. D. Mead, and A. W. Meade, "Item selection using CTT and IRT with unrepresentative samples", Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA,2010.
[18] J. C. Nunnally, Psychometric theory (2nd Ed.). New York: McGraw-Hill, 1978.
[19] G. Pope, Item analysis analytics part 1: What is Classical Test Theory? 2009. (Online) available at http://blog.questionmark.com/item-analysis-analytics-part-1-what-is-classical-test-theory[accessed on July 5, 2014]
[20] G. Sax, Principles of educational and psychological measurement and evaluation (3rd ed). Wadsworth, Belmont, CA, 1989.
[21] B. Thompson, and J. E. Levitov, "Using microcomputers to score and evaluate test items"Collegiate Microcomputer, 3, pp.163-168, 1985.
[22] S. Varma, Preliminary item statistics using point-biserial correlation and p-values, 2008, (Online) available at http://www.eddata.com/resources/publications/EDS_point_Biserial.pdf[accessed on October 7, 2014]
[23] W.Wiersma, and S. G. Jurs, Educational measurement and testing (2nd ed). Allyn and Bacon, Boston, 1990.
[24] A. M. Zubairi, and N. L. A. Kassim, "Classical and Rasch analysis of dichotomously scored reading comprehension test items",Malaysian Journal of ELT Research, 2, pp.1-20, 2006.
[25] M. Shakil, "Assessing Student Performance Using Test Item Analysis and its Relevance to the State Exit Final Exams of MAT0024 Classes - An Action Research Project". A Paper presented on MDC Conference Day, March 6th, 2008 at MDC, Kendall Campus.
[26] S., B Chang, Hanson, and D. Harris, "A Standardization Approach to Adjusting Pre-test Item Statistics". Paper presented at the annual meeting of the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No.ED 442 838), 2000.
[27] J.O.O.Abiri, Element OF Evaluation and Measurement Techniques in Education. Ilorin: Library and publication committee University of Ilorin Nigeria, 2007.
[28] P.N., Okpala, C. O. Onocha, and O. A. Oyedeji, Measurement and Evaluation in Education. Jattu-Uzairue Stirling- Horden Publishers Nigeria, 1993.
[29] J. E.Ofo, Research Methods and Statistics in Education and Social Sciences. Joja Educational Research and Publishers, Lagos, 1994.
[30] J. Payne, "Contingent Decision Behaviour", Psychological bulletin, 92, Pp.382-402, 1982.
[31] D. Olatunji, and H. O. Owolabi, "Difficulty and Discrimination of Economics Test Items with Various Option Formats among Secondary Schools in Ilorin", Nigeria. Ilorin Journal of Education, Vol. 28 pp.49-63, 2009.
[32] J. C. Nunnally, Educational measurement and evaluation. 2nd edition, McGraw-Hill, New York. 1972.
[33] Suruchi and S. R. Rana, "Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology". Paripex - Indian Journal of Research, Vol. 3(6) pp.56-58, 2014.

[34] K. Carole, and A. Winterstein, "Validity and reliability of measurement instruments used in research", Am J Health-Syst Pharm. Vol 6(5), 2008.
[35] N.K.,Mitra, H.S. Nagaraja, G. Ponnudurai, and J.P. Judson, "The Levels of Difficulty and Discrimination Indices in Type A Multiple Choice Questions of Pre-Clinical Semester 1 Multidisciplinary Summative Tests", International Journal of Science and Medical Education, 3: pp.2-7, 2009.
[36] S.S. Pande, S.R. Pande, V.R. Parate, A.P. Nikam, and S.H. Agrekar, "Correlation between Difficulty and Discrimination Indices of Multiple Choice Questions in Formative Exam in Physiology", South East Asian Journal of Medical Education, 7: pp.45-50, 2013.
[37] S. Sim, and R.I. Rasiah, "Relationship between Item Difficulty and Discrimination Indices in True/False-Type Multiple Choice Questions of A Para-Clinical Multidisciplinary Paper". Annals-Acad. Med. Singapore, 35: pp. 67-71, 2006.
[38] M.R. Hingorjo, and F. Jaleel, "Analysis of One-Best Multiple Choice Questions: The Difficulty Index, Discrimination Index and Distractor Efficiency". J. Pak. Med. Assoc., 62: pp.142-147, 2012.
[39] G. Henning,A Guide to Language Testing: Development, Evaluation, Research. Newberry House Publisher, Cambridge Mass, 1987.
[40] Ebel, R.L. and D.A. Frisbie, Essentials of Educational Measurement. 5th Edn., Prentice Hall, Engelwood Cliffs, New Jersey, 1991.
[41] R. K., Hambleton and R. W. Jones "Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development. Educational Measurement: Issues and Practice, 12(3), pp.38-47, 1993.
[42] B. A. Adegoke"The Role of Item Analysis in Detecting and Improving Faulty Physics Objective Test Items". Journal of Education and Practice, 5(21), pp.110-120, 2014
[43] Le, D.-T. 2013. Applying item Response Theory Modeling in Educational Research. Doctoral dissertation, Iowa State University.

Author Profile

Ado Abdu Bichiis a Graduate assistant at the Northwest University, Kano Nigeria.He received his Bachelor degree in Education and Economics with Upper second Class Honours from Bayero University, Kano Nigeria in 2009. He is currently undergoing his Master degree in Educational Assessment (Psychometric Methods) at the University Sultan Zainal Abidin, Malaysia. His areas of teaching and research interest areEducational and Psychological Test, Measurement and Evaluation, Modern Assessment Practices, Classical Test Theory and Practices, Item Response Theory and Applications in Test Item Development, Research Methods and Statistics in Education