The 41st Annual Conference of the International Association for Educational Assessment (IAEA), University of Kansas, Lawrence Kansas, USA, 11-10-2015

Theme: **The Three Most Important Considerations in Testing: Validity, Validity, Validity**

Subtheme: **Using Technology to Improve Validity**

Title: Evaluation of Validity of Computer Based Test Items in National Open University of Nigeria

Abstract

Multiple Choice Items (MCI) are one of the most commonly used Computer Based Assessment (CBA) instrument for assessment of students in educational settings especially in Open and Distance Learning (ODL) with large class sizes. The MCI making up the assessment instruments need to be examined for quality which depends on its Difficulty Index (DIF 1), Discrimination Index (DI), and Distractor Efficiency (DE) if they are to meaningfully contribute to validity of the students' examination scores. Such quality characteristics are amenable to examination by item analysis. Hence, the objective of this study is to evaluate the quality of MCI used for CBA in the National Open University of Nigeria (NOUN) as formative assessment measures by employing expost facto research design. One foundation course in School of Education of the University was used for the study. The aim is to develop a pool of valid items by assessing the items DIF 1, DI and DE and also to store, revise or discard items based on obtained results. In this cross-sectional study, 240 MCI taken in four (4) sets of CBA per semester per course in 2012 – 2014 academic years were analysed. The data was entered and analysed in MS Excel 2007. The results indicated items of "good to excellent" DIF I and "good to excellent" DI, Efficient Distractors (DE) and non functional distractors (NFD). Also established were items with poor DI. This study emphasized the selection of quality MCI which truly assess levels of students learning and differentiate students of different abilities in correct manner in NOUN thereby contributed to improving the validity of the test items.

Keywords: Difficulty index, discrimination index, distractor efficiency, multiple choice items, non functional distracter, validity of test scores

Introduction

MCI are considerably widely used as a means of objective measurement. This is because of the many dominant advantages associated with this form of test format. Apart from the fact that it can easily be used to overcome the challenges of large class sizes by being amenable to computer administration and objective scoring of test items, it also aids in timely compilation and release of examination results (Okonkwo, 2010). In addition, they can be

used for diagnostic as well as formative purposes and can assess a broad range of knowledge. Hence, the National Open University of Nigeria (NOUN) uses MCI administered to students using the computer for her formative assessment of students learning outcomes. This computer based assessment (CBA) accounts for 30% of the student's grade in each of the courses offered by students of NOUN.

Test items generally have guideline for writing them as well as for testing the test items. Amongst the guidelines for option development is that which deals with the number of options to be written for each item (Amrahi & Baghaei, 2011). Haladyna, Downing and Rodriguez (2002), in their taxonomy of multiple choice items writing guidelines suggested 43 guidelines of which 10 are concerned with general item writing, 6 are related to stem development, and 20 refer to option development. There is clearly an important concern in MCI writing as indicated by its attraction of 20 guidelines. Traditionally, it is recommended to use four or five options per item in order to reduce the effect of guessing by providing the examinees with plausible distractors as possible. Thus, most classroom achievement tests as well as international standardized test usually follow the rule of four options per item. NOUN also uses four options MCI.

According to the recent understanding of validity, validation is the joint responsibility of the test developers and the test users. Whereas, the test developer is responsible for providing relevant evidence and a rationale in support of the intended test use, the test user is ultimately responsible for evaluating the evidence in the particular context in which the test is to be used (AERA, APA and NCMC, 1999). Both functions are simultaneously performed by NOUN academic staffs.

As earlier stated, test items generally have guidelines for writing them as well as for testing the test items. The test developers for National Open University of Nigeria (NOUN) Computer Based Assessment (CBA) used for formative continuous assessment are always concerned with the 'what' of testing – the content of the test items. This is usually achieved by focusing efficiency on the course content with the aid of table of specific or test blue print developed using NOUN house style in almost all testing situations. Of course, the 'how' of testing is already predetermined as Multiple Choice Item (MCI) because of its advantages earlier enumerated. Thereby satisfying the two main issues of concern to test developers – "the what and the how of testing". Although care of content validity (qualitative item analysis) of the MCI have been taken care of by the use of test blue print during the item writing development stage, it is still vital to establish the quantitative (how of testing) item analysis in order to fully build in quality in the test items. Hence, this study is focused on item analysis via the quantitative perspective.

Item analysis is a process of collecting, summarizing and using information from students' responses to assess the quality of test items (Karelia, Pillai & Vegada, 2013). But, making fair

and systematic assessment of others performance can be a challenging task. A view also expressed by Matlock-Hetzel (1997). Moreover, judgements cannot be made solely on the basis of intuition, haphazard guessing, or custom (Sax, 1989). Hence, evaluators use a variety of tools to assist them in their evaluations. One of the tools frequently used to facilitate the evaluation process is tests. Test is used to assess the effects of instruction on educational programme as is the case of NOUN. It is therefore essential to conduct item and test analyses. Item analysis includes three statistics that can help in the analysis of the effectiveness of test items. These are the difficulty index (DIF I or P), discrimination index (DI) and distractor efficiency (DE).

Item analysis examines how the test items perform as a set. It "investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test" (Thompson & Levitov, 1985, 163). These analyses evaluate the quality of items and of test as a whole (Matlock-Hetzel, 1997). Thus, the analyses invariably validate the test and test items, and can also be employed to revise and improve both items and test as a whole. Item analysis is used to help "build" reliability and validity 'into' the test from the start. Generally, validity is defined as the degree to which a test measures what it is supposed to measure. Whereas, reliability deals with the extent to which a measure is repeatable and stable. That is, the consistency of a measure. There are varieties of techniques for performing item analysis. Item analysis can be both qualitative (what of testing) and quantitative (how of testing). The former focuses on issues related to the content of the test such as content validity. Whereas, the later primarily includes measurement of item difficulty and item discrimination. The later perspective is the hub of this study and it provides framework for conducting the validity of MCI used in NOUN CBA.

According to AERA, APA and NCME (1999, 9) validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. In the past, validity theory considered were different types of validity namely content, construct and predictive. Nowadays item theory has evolved and is defined as unitary concept. Though there are various sources of evidence that may shed some kind of light on different aspects of validity, but they do not constitute different types of validity.

Validity is the adequacy and appropriateness of interpretations and uses of assessment results. A test is valid if it appropriately measures what it is supposed to measure (Miller, Linn & Gronlund, 1995). They opined that validity contains a variety of properties and its influenced by a number of factors which need to be considered before a test and situation is judged valid. These factors are those of the test itself, factors related to learning task and learning situation, factors in test administration and scoring, as well as those related to student response, the nature of the group, the criterion being used, and the nature of the teaching and evaluation and assessments instruments. The validity of a test is critical

because without sufficient validity test scores have no meaning. The evidence one collect and document about the validity of one's test is also adjudged the best legal defence for the examination programme (Professional Testing Inc, 2006). In this sense, validity denotes the meaning of a test score or assessment result. But, validity is generally defined as the degree to which a test measures what it is supposed to measure. There are three basic approaches to validity of tests and measures as shown by Mason and Brumble (1989) already known. These are content validity, and criterion related validity.

The main sources of evidence that might be used to evaluate the validity of an instrument (AERA APA and NCME 1999) are:

- Evidence based on test content, which is an analysis of the relationship between a test's content and the construct it is intended to measure.
- Evidence based on response processes, which requires theoretical and empirical analyses of the responses of test takers in order to provide evidence of the fit between the construct and the nature of performance given by examinees.
- Evidence based on relations to other variables external to the test (such as the scores of other tests measuring the same construct or group (membership variables), which requires an analysis of the degree to which these relationships are considered with the construct underlying the test.
- Evidence based on consequences of testing, which proposes the incorporation of the intended and unintended consequences of test use into the concept of validity.

However, it is import to note that a test cannot be qualified as valid in absolute terms.

Nevertheless, validity is the adequacy and appropriateness of the interpretations and uses of assessment results (Miller, Linn & Gronlund, 2010). A test is said to be valid if it appropriately measures what it is supposed to measure. A critical view of validity (Pedhazur & Schmelkin, 1991) is that it refers to inferences made about scores and not to assessment of content of an instrument. Thus, the conventional view of validity fragmented with respect to content, criterion and construct failed to take into account both evidence of the value implications of score meaning as a basis for actionable items and the social consequences of using the test scores (Messick, 1995). Hence, according to Messick (1995) validity is not a property of the test or assessment but rather it is about the meaning of the test scores. Messick (1998) further argued that social consequences of score interpretations include the value implications of the construct, and this implication must to be addressed by evaluating the meaning of the test score.

Therefore, the purpose of this paper is to evaluate the validity of computer based test used by National Open University of Nigeria going by the quantitative perspective of item analysis. In this regards, the best practices of item analysis and test analysis were employed by using the tools – item difficulty, item discrimination and option distractor efficiency to evaluate the validity of the NOUN CBA test items.

Difficulty index (DIF I or P_i) is the proportion of the examinees who answered the item correctly. DIF I or P_i is calculated as follows:

DIF I or $P_i = C_i/N = (H+M+L)/N$

DIF I or P_i : proportion of examinees who answer item i correctly

C_i: Number of examinees who answer item i correctly

H: Number of examinees in high (27%) group who answer item i correctly

M: Number of examinees in middle (46%) group who answer item i correctly

L: Number of examinees in low (27%) group who answer item i correctly

N: Number of examinees who are examined

The difficulty index ranges from 0 to 1. Values close to 0 mean only a few examinees answered the item correctly; values close to 1 means the item was answered correctly by most of the individuals. Hence, the purpose of a test is to have a wide variety in total score, items with values close to 0 or 1 have to be reviewed or may as well be eliminated. Since, they provide relatively little information for discriminating between test-takers. Item difficulty index (DIF I or P_i) can be classified into five categories. The first and the fifth categories are the ones that require special attention (Crocker and Algina, 1986). The categories are: extremely easy (.75 – 1), easy (.55 - .74), moderate (.45 – 54), difficult (.25 - .44) and extremely difficult (0 - .24).

Item discrimination index or Item Discrimination power (DI or D_o) is an index which indicates how well an item is able to distinguish between the more knowledgeable and the less knowledgeable examinees given what the test is measuring. The measure of the level of knowledge is the total score in that test (Nenty, 1985). An item discrimination index is an indication of how much better those who score highly in the entire test perform on that particular item than those who scored poorly on that test. It is used to estimate the extent to which an item helps to discriminate between examinees with high and low performance in a given test. The generally accepted procedure in analyzing a test for item discrimination is to sort the papers from lowest score to highest. Optimal item discrimination is obtained when the upper and lower groups each contain twenty-seven percent of the total group (Richardson, 2002). The groups are used because they maximize differences in normal distribution while providing enough cases for analysis (Wiersma & Jurs, 1990: 145). Therefore, two equal groups using the highest 27% (H) and the lowest 27% (L) scorers are identified, and the intermediate scores of 46% are also identified but are not used in the computation. This is done after grading the test. The item discrimination index is determined by examining the responses to each question by the two extreme groups – highest 27% (H) and the lowest 27% (L) scorers in a test. For each item the DI or D_o is determined by the formula:

DI or D_o = $(P_{i(H)} – P_{i(L)})/n$ or (H-L)/n

Where:

DI or D_o: discrimination index for item i

$P_{i(H)}$: proportion of examinees in the higher tercile on the total score for the test who answer item i correctly

$P_{i(L)}$: proportion of examinees in the lower tercile on the total score for the test who answer item i correctly

H: Number in high group who answer item correctly

L: Number in low group who answer item correctly

n: Number in each group in each group. That is 27% of the test takers (and not the entire test takers)

Generally when students who earn high scores are compared with those who earn low scores, it is expected that more students in high scoring group would answer a question correct more than students from the low scoring group. For very difficult items which no one in either group answered correctly or fairly easy questions which even the students in the low group answered correctly, the numbers of correct answers might be equal for the two groups. However, it is not expected that the low scoring students should answer an item correctly more frequently than students in the higher group. Positive item discrimination index indicates that the item discriminates in the desired direction in favour of the high achievers. Whereas negative item discrimination index means that the item discrimination is against high achiever and indicates a cue that there may be a problem with the way the item was presented on the test or the way the material was taught (or not taught). Such items should be examined for possible ambiguity. Discrimination index ranges from -1.00 to 1.00. According to the value of the index, the discrimination power of any item can be categorised as follow: extreme high (.40-1), high (.30 - .39), moderate (.20 - .29), low (0 - .19) and to discard (< 0). The items that present problems are those located in the last two categories (low or to discard) (Matlock-Hetzel, 1997; Crocker and Algina, 1986).

Distractor analysis (DE) is usually used to examine MCI to determine the effectiveness of the various distrators that were provided. Functional distractors are distrators that are selected by students who failed to choose the correct option to a give item. It is not desirable to have one of the distractors chosen more often than the correct answer. When that happens, it indicates a potential problem (Richardson, 2002) with the item. Either the distractor may be too similar to the correct option (key) and/or they may be something in either the stem or the alternatives that is misleading. When the correct answer is not known to the test takers, and they are purely guessing, their responses would be expected to be distributed among the distrators as well as the correct answer. But, generally an item could have higher percentage of correct responses while still having effective distractors. If one or more distractors are not chosen, the unselected distractors probably are not plausible. Those distractors that are not selected by the test takers should be replaced in subsequent

administration of the tests (Richardson, 2002). The effectiveness of test items may be improved as well as the validity of test scores by selecting and rewriting the items on the basis of item performance data.

Non response rate (NRR) is the proportion of people who do not answer the item. This rate is obtained from the relation:

$nr_i = 1 - p_i - q_i$

Where:

nr_i : proportion of examinees who do not answer the item i

p_i : proportion of examinees who answer the item i correctly

q_i: proportion of examinees who answer the item i incorrectly

According to the percentage of people who did not answer the item, the non-response rate can be categorised as follows: adequate (0 - .15), acceptable (.16 - .20), tolerable (.21 - .29) and to discard (.30 – 1). In this way, items with non-response rates above .30 have to be discarded or reviewed because most of the examinees may have found the item problematic (Matlock-Hetzel, 1997; Oosterhof 1990; Crocker & Algina 1986). It could be not understandable or too difficult.

Objective

The objective of this study is to determine the quality of MCI used in CBA in the NOUN as formative assessment measure. This was done by employing Expost facto research design. The aim of the study was to bring forth items a pool of valid items by assessing the items difficulty index, discrimination index and distractor efficiency for each of the items. By the identification of items to be stored, revised or discarded based on the obtained results.

Materials and Methods

One foundation course offered by the School of Education, National Open University of Nigeria in 2012 to 2014 academic years was used for the study. The cross sectional research was performed on 240 Multiple Choice Items (MCI) taken in four (4) sets of Computer Based Assessment (CBA) of 80 items per semester. Each set consists of 20 items in each of the 4 consecutive sets of CBA in a semester. A sample of 878 students out of a population of 3909 students who were examined in the course was used for the study. The MCI comprised of “single response type”. All the items had single stem with four options/responses including one key (correct answer) and other three options (incorrect answers/distractors). Each correct response was awarded ½ marks while incorrect response was awarded 0 marks. The score ranged from 0 to 10 per set of 20 items. To avoid possible coping from neighbouring students, the tests were programmed to be computer reshuffled for every individual student taking the test.

Data obtained was entered in MS Excel 2007 and analyzed in the yearly sequence. The scores of 1750 students that took the test in 2012 were entered in MS Excel 2007. The scores were then sorted with scores ranging in descending order from 10 marks to 0 marks for the sample size of 313 students out of 1750 students who took the test in 2012 academic year. One group of 85 students, consisting of higher marks from top was considered as higher ability (H). This group consists of 27% of the sample of 313 students. The other group of 85 students consisting of lower marks from the least score upwards was considered as lower ability (L). This group also consists of 27% of the sample of 313 students. The middle 143 students were extracted centrally from the 1750 students to complete the sample size of 313. Thus, out of the sample of 313 students, 85 were in H group, 85 were in L group while the remaining 143 were in the middle group. The same grouping pattern was adopted and performed for the 2^{nd}, 3^{rd} and 4^{th} sets of CBA for 2012. Also the process was repeated for 2013 and 2014. However, in 2013 a sample of 291 students out of a population of 1232 students that took the CBA in the foundation course were used in the data analysis. Their distributions were higher group (H) 79 students, lower group (L) 79 students and the middle group (M) 133 students. In 2014 a sample of 274 students out of a population of 927 students who took the course were used in the data analysis. Likewise, their distributions were 74 students in the H group, 74 students in the L group and 126 students in the M group. A total of 80 MCI and 320 distractors were analysed for each year. This summed up to a total of 240 MCI and 960 distractors for the 3 academic years under consideration.

Based on the data, various indices like Difficulty Index (DIF I), Discrimination Index (DI), Distrator Efficiency (DE) and Non Functional Distractors (NFD) were calculated with the following formula.

1. DIF I = (H+M+L)/N
2. DI = (H-L)/n
3. DE = (L-H)/n
4. NFD = An item option attracting <5% of the examinees

   Where:

   DIF I: Difficulty index
   DI: Discrimination index
   DE: Distractor efficiency
   NFD: Non functional distrator
   H: Higher achievers who got item correct
   L: Lower achievers who got item correct
   M: Middle achievers who got item correct
   N: Total number of examinees responding to item
   n: Number of examinees in each group

Item analysis was employed on the items entered in the MS Excel 2007 and analysed using the formula for DIF I, DI and DE above.

The difficulty index (DIF I) categories were set to: extremely easy (.75 – 1), easy (.55 - .74), moderate (.45 - .54), difficult (.25 - .44) and extremely difficult (0 - .24).

The discrimination index (DI) was classified into five categories as: extremely high (.40 -1), high (.30 - .39), moderate (.20 - .29), low (0 - .19) and to discard (< 0).

Also, the non-response rate can be categorised as follows: adequate (0 - .15), acceptable (.16 - .20), tolerable (.21 - .29) and to discard (.30 -1). It was intended that items with non-response rates above .30 have to be recommended to be discarded or to be reviewed because most of the examinees may have found the item problematic (not understandable or too difficult). However, in this study, all the items were responded to.

Finally, items with non functional distrators (NFD) were considered. Here, NFD in an item is an option(s) other than the correct answer (key) selected by less than 5% (<5%) of the examinees. Alternatively, functional effective distractors are those selected by 5% or more of the participants.

Distractor efficiency (DE) is determined for each item on the basis of the number of NFDs in it and ranges from 0 to 1. If an item contains 3 or 2 or 1 or 0 NFD, then DE will be .33 (or 33.3%), .66 (or 66.6%) or 1 (or 100%) respectively. Items were categorised as poor, good or excellent and actions such as discard/revise or store were proposed based on the values of DIF, DI and DE as suggested.

Result

The result of the items analyses are presented and discussed in the following tables.

Table 1A: 2012 Distribution of Items in relation to DIF I and actions proposed

	Difficulty Index (DIF I)							
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.75 - 1	7	12	11	15	45	56.25	Easy	Revise
.55 - .74	6	3	4	3	16	20.00	Excellent	Store
.45 - .54	4	2	3	1	10	12.50	Very good	Store
.25 - .44	2	2	1	1	6	7.50	Good	Revise & Store
0 - .24	1	1	1	0	3	3.75	Difficulty	Revise
Total	20	20	20	20	80	100.00		

Table 1A showed the difficulty index (DIF I) of the items used for CBA in 2012. Out of the 80 items used in the assessment of the students learning outcomes, 45 were easy, 3 were difficult, while others items are distributed amongst excellent, very good and good as

expected. The excellent, very good and good items are ideal for storing in the question bank while the easy and difficult ones are to be revised to enhance their validity if they are to be used in subsequent CBA test. Revising them would help in increasing their validity as assessment instruments.

Table 1B: 2012 Distribution of Items in relation to DI and actions proposed

				Discrimination Index (DI)				
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.40 – 1	19	16	17	19	71	88.75	Excellent	Store
.30 - .39	0	2	2	0	4	5.00	Very good	Store
.20 - .29	1	1	0	0	2	2.50	Good	Store
0 - .19	0	1	1	1	3	3.75	Poor	Revise/Discard
<0	0	0	0	0	0	0.00	Undesirable	Discard
Total	20	20	20	20	80	100.00		

Table IB revealed the distribution of CBA items used in 2012 in terms of their discrimination index (DI). Out of the 80 items presented to the examinees, only 3 were poor. They failed to discriminate appropriately between the high achievers and the low achievers. This items call for attention and further actions to be taken on them such as revision of the items or discarding them from the item pool so as to increase the validity of the test. The remaining 77 items are ideal and are to be stored for subsequent use in the assessment of students learning outcomes.

Table 1C: 2012 Distribution of Items in relation to DE and actions proposed

				Distractor Efficiency (DE)				
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
0 NFD	10	9	7	6	32	40.00	Excellent	Store
1 NFD	6	7	9	7	29	36.25	Very good	Store
2 NFD	3	3	4	6	16	20.00	Good	Store
3 NFD	1	1	0	1	3	3.75	Poor	Revise/Discard
Total	20	20	20	20	80	100.00		

Table 1C revealed the distribution of items in terms of the number of options that had non functional distrators (NFD) for the year 2012. The result showed that only 3 items had 3NFD while the remaining 77 items had NFD ranging from 0 to 2 as presented in the table. This is a good indication that the items are valid. However, the 3 items with 3NFD should be revised or discarded to improve the validity of the test items thereby improving the quality of the test items.

Table 2A: 2013 Distribution of Items in relation to DIF I and actions proposed

Difficulty Index (DIF I)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.75 - 1	7	10	3	6	26	32.50	Easy	Revise/Discard
.55 - .74	5	6	5	10	26	32.50	Excellent	Store
.45 - .54	5	1	1	1	8	10.00	Very good	Store
.25 - .44	2	1	9	3	15	18.75	Good	Store
0 - .24	1	2	2	9	5	6.25	Difficulty	Revise/Discard
Total	20	20	20	20	80	100.00		

Table 2A revealed the distribution of items in relation to DIF I and the action proposed for CBA in 2013. The easy and the difficult items need to be revised in order to improve on them. The excellent to good items should be stored and reuse because they are valid items.

Table 2B: 2013 Distribution of Items in relation to DI and actions proposed

Discrimination Index (DI)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.40 – 1	17	14	15	13	59	73.75	Excellent	Store
.30 - .39	0	4	4	4	13	16.25	Very good	Store
.20 - .29	0	0	1	1	4	5.00	Good	Store
0 - .19	3	2	0	2	4	5.00	Poor	Revise/Discard
<0	0	0	0	0	0	0,00	Undesirable	Discard
Total	20	20	20	20	80	100.00		

Table 2B showed that out of the 80 items used in the CBA only 5 were poor and should be revised or discarded while 75 spanned between excellent and good and should be stored for future use.

Table 2C: 2013 Distribution of Items in relation to DE and actions proposed

Distractor Efficiency (DE)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
0 NFD	9	3	11	8	31	38.75	Excellent	Store
1 NFD	8	7	5	6	26	32.50	Very good	Store
2 NFD	2	7	3	3	15	18.75	Good	Store
3 NFD	1	3	1	3	8	10.00	Poor	Revise/Discard
Total	20	20	20	20	80	100.00		

Table 2C revealed the characteristics of the distractors with respect to the non functional ones. Out of the 80 items administered to the students for the CBA in 2013, 8 had 3NFD and should be revised or discarded in order to improve the validity of the items. The remaining 70 items could be store for future use.

Table 3A: 2014 Distribution of Items in relation to DIF I and actions proposed

Difficulty Index (DIF I)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.75 - 1	8	12	11	13	44	55.00	Easy	Revise/Discard
.55 - .74	9	5	5	6	25	31.25	Excellent	Store
.45 - .54	2	0	0	0	2	2.50	Very good	Store
.25 - .44	0	2	4	1	7	8.75	Good	Store
0 - .24	1	1	0	0	2	2.50	Difficulty	Revise/Discard
Total	20	20	20	20	80	100.00		

Table 3A revealed the distribution of items used in CBA in the year 2014 in terms of their difficulty levels. The easy and the difficult items needed to be revised to improve their validity while the rest should be stored and reused.

Table 3B: 2014 Distribution of Items in relation to DI and actions proposed

Discrimination Index (DI)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
.40 – 1	17	14	15	13	59	73.75	Excellent	Store
.30 - .39	1	4	4	4	13	16.25	Very good	Store
.20 - .29	2	0	1	1	4	5.00	Good	Store
0 - .19	0	2	0	2	4	5.00	Poor	Revise/Discard
<0	0	0	0	0	0	0	Undesirable	Discard
Total	20	20	20	20	80	100.00		

Table 3B exposed the distribution of the discrimination index for the items used in CBA for the year 2014. The discrimination index showed that 76 items were distributed from excellent to good. Only 4 items had poor discrimination index and should be revised to improve on their discrimination ability between the high and the low achievers.

Table 3C: 2014 Distribution of Items in relation to DE and actions proposed

Distractor Efficiency (DE)								
Cut of points	CBAI	CBA2	CBA3	CBA4	TOTAL	%	Interpretation	Action
0 NFD	5	3	4	3	15	18.75	Excellent	Store
1 NFD	10	6	8	9	33	41.25	Very good	Store
2 NFD	4	7	6	7	24	30.00	Good	Store
3 NFD	1	4	2	1	8	10.00	Poor	Revise/Discard
Total	20	20	20	20	80	100.00		

Table 3C showed the distribution of non functional distractor per item in the 80 items used in the foundation course under review for NOUN CBA in 2014. Out of the 80 items, 10 had 3NFD and are to be revised or discarded to enhance the validity of the test items. Whereas, the remaining 70 items had 0 to 2 NFD which were spread across excellent to good and are therefore to be stored for subsequent use.

Summary

Tables 4 and figures 1 – 3 below are use to summarise the quality of the CBA used in the NOUN during the period 2012 to 2014 in terms of the Item Difficulty (DIF I), Item Discrimination (DI) and Distractor Efficiency (DE).

Table 4A: Distribution of Items in relation to Item Difficulty (DIF I) and actions proposed

Item Difficulty (DIF I)						
Interpretation / Action	Cut of points	2012	2013	2014	Total	Average
Easy / Revise	.75 – 1	45	26	44	115	38
Excellent / Store	.55 - .74	16	26	25	67	22
Very good / Store	.45 - .54	10	8	2	20	7
Good / Store	.25 - .44	6	15	7	28	9
Difficult / Revise	0 - .24	3	5	2	10	3
	Total	80	80	80	240	80

Table 4A should that item difficulty distribution and the average for the three years under study.

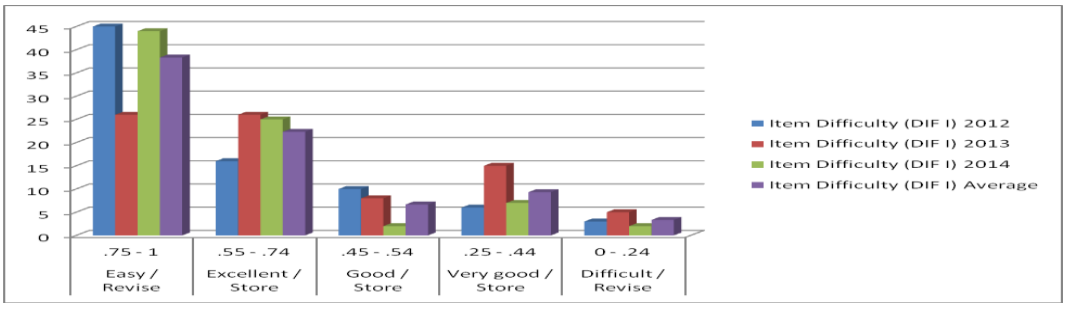

Figure 1: Distribution of Items in terms of their difficulty level for 2012 - 2014

Figure 1 illustrated the distribution of the items in terms of the items difficulty distribution for the three years under study.

Table 4B: Distribution of Items in relation to Item Discrimination (DI) and actions proposed

Item Discrimination Index (DI)						
Interpretation / Action	Cut of points	2012	2013	2014	Total	Average
Excellent / Store	.40 – 1	71	59	59	189	63
Very good / Store	.30 - .39	4	13	13	30	10
Good / Store	.20 - .29	2	4	4	10	3
Poor / Revise/Discard	0 - .19	3	4	4	11	4
Undesirable / Discard	<0	0	0	0	0	0
Total	80	80	80	80	240	80

Table 4B exposed the distribution of the items used in NOUN CBA for the period 2012 – 2014 in the foundation course studied.

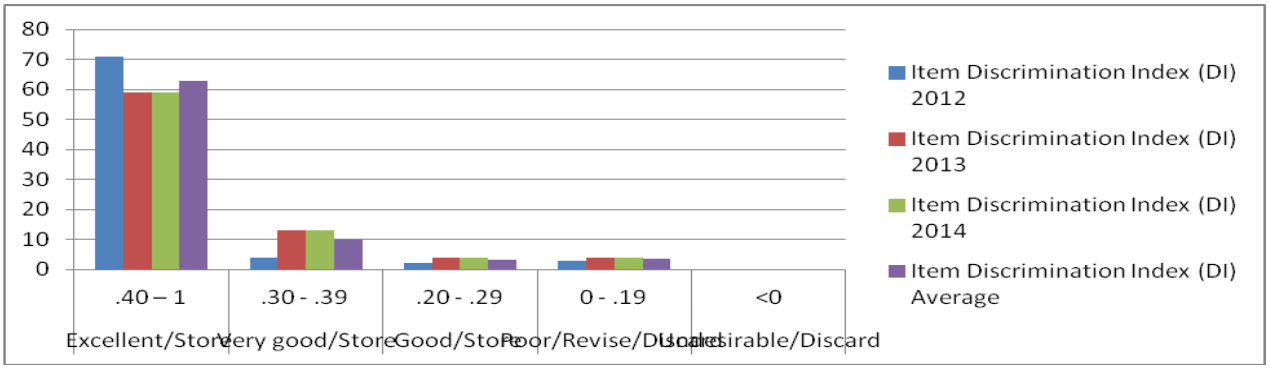

Figure 2: Distribution of Items in terms of their Discrimination Index (DI) for 2012 - 2014

Figure 2 bared the distribution of items in terms of their discrimination between high and the low achievers for the period and the course under study.

Table 4C: Distribution of Items in relation to Distractor Efficiency (DE) and actions proposed

Distractor Efficiency (DE)						
Interpretation / Action	Cut of points	2012	2013	2014	Total	Average
Excellent / Store	0 NFD	32	31	15	78	26
Very good / Store	1 NFD	29	26	33	88	29
Good / Store	2 NFD	16	15	24	55	18
Poor / Revise/Discard	3 NFD	3	8	8	19	6
Total		80	80	80	240	80

Table 4C revealed the distractor efficiencies for the 240 items used for CBA in NOUN for the foundation course and the period under review. The DE is interpreted on the basis of the number of non functional distractors (NFD) per item.

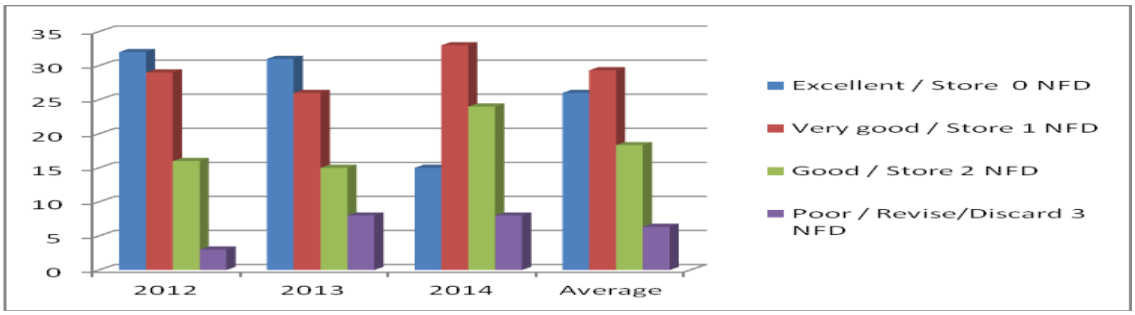

Figure 3: Distribution of Items in terms of their Distractor Efficiency (DE) for 2012 – 2014

Figure 3 illustrated the distribution of the distrator efficiencies (DE) in terms of non functional distractors (NFD) for a period of 2012 -2014 in NOUN CBA for the foundation course under review.

Conclusion

It is obvious that Multiple Choice Items (MCI) are indispensably used as Computer Based Assessment (CBA) instrument for assessment of students in educational settings especially in Open and Distance Learning (ODL) with large class sizes. Nevertheless, the MCI making up the assessment instruments need to be examined for quality which depends on its Difficulty Index (DIF 1), Discrimination Index (DI), and Distractor Efficiency (DE) if they are to meaningfully contribute to validity of the students' examination scores. Hence, the quality characteristics of MCI used in one foundation course in NOUN are examination by item analysis with a view of generating a pool of valid items for storage and to identify those that needs improvement in order enhance their validity.

In this cross-sectional study, 240 MCI taken in four (4) sets of CBA per semester per course in 2012 – 2014 academic years were analysed. The data was entered and analysed in MS Excel 2007. The results indicated that 230 items were of "good to excellent" DIF I and 229 items were of "good to excellent" DI, while 211 items had Efficient Distractors (DE) and only 19 items had non functional distractors (NFD). Also established were items with poor DI. Hence, the study emphasized the selection of quality MCI which truly assess levels of students learning and differentiate students of different abilities in correct manner in NOUN thereby contributed to improving the validity of the test items.

It also recommended that the poor items which did not measure up to the desired quality be revised or discarded to enhance the validity while the valid and quality items be stored for future use.

Recommendations

It is recommended that the National Open University of Nigeria as well as other institutions using Multiple Choice Items in the assessment of students learning outcomes should regularly evaluate the items to determine the quality of the items. The quality items should be pooled and stored in item bank for future use while the poor items should be revised or discarded depending on the problems associated with them. It is also recommended that the exercise should target all courses in which MCI are use as assessment instrument to increase the validity and quality of such instruments.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5).

Crocker, L. and J. Algina (1986) *Introduction to Classical and Modern Test Theory*, New York: Holt, Rinehart and Winston

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

Karelia, B. N., Piillai, A. & Vegada, B. N. (2013). The level of difficulty and discrimination indices and relationaship between them in four-response type multiple choice questions of pharmacology summative tests of year II M.B.B.S students, leJSME, 7(2): 41-46

Mason and Bramble (1989). In Key, J. P. (1997). Research Design in Occupational Education. Except from those materials supplied by different department of the Oklahoma State University. www.okstate.edu/ag/agedom4h/academic/age5980/newspage18.htm

Matlock-Hetzel, S. (1997) 'Basic Concepts in Item and Test Analysis', Texas A. M. University. Available at http://www.ericae.net/ft/tamu/Espy.htm (accessed 1 Dec 2008)

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. American Psychologist, 9, 741-749

Messick, S. (1998). Test validity: A matter of consequence: *Social Indicator Research,* 45, 35-41

Miller, M. D., Linn, R. L., & Gronlund, N. (1995). Measurement and Assessment in Teaching. Pearson Education.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). Measurement and assessment in teaching (10th ed.). Upper Saddle River, NJ: Pearson Education.

Nenty, H. J. (1985). Fundamentals of Measurement and Evaluation in Education. Faculty of Education, University of Calabar, Calabar, Nigeria.

Okonkwo, C. A. (2010). Using e-Assessment to enhance the operational efficiencies of the National Open University of Nigeria (NOUN). Journal of Educational Assessment in Africa, 5, 117 – 138. Publisher: The Association for Educational Assessment in Africa (AEAA).

Oosterhof, A. (1990) *Classroom Applications of Educational Measurements*, Columbus, OH: Merrill

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates. Publishers

Professional Testing Inc. 2006

Richardson, E. (2002). Item and Test Analysis. (Supports PEPE Teacher Indicator 3.2. Alabama Professional Development Modules. Alabama Department of Education. Institute for Assessment.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thompson, B. & Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer,* 3, 163 – 168.

Wierma, W. & Jurs, S. G. (1990). Educational Measurement and Testing (2nd ed.) Boston, MA: Allyn and Bacon