Trends in Educational Studies Journal (TRESJ)
Volume 14, Issue 1
January-June, 2022
University of Port Harcourt
ISSN: 0795-3186

Item Analysis of Adopted Cognitive Achievement Test and Performance of Primary Four Pupils in Jere, Maiduguri, Nigeria

Rhoda Emmanuel Camble [(1)], Abubakar Hamman-Tukur [(2)] & Mohammed Bularafa Waziri [(3)]
Department Of Education, University Of Maiduguri [(1, 2, 3)]

Keywords:
Item Analysis, Item Difficulty, Item Discrimination, Achievement Test, Performance, Curriculum

Abstract

Achievement tests are one of the most important aspects of teaching-learning process and the two most important characteristics of an achievement test are its reliability and content validity. Objectives of the study were to determine the reliability of the Adopted Cognitive Achievement Test, find out the item difficulty level and discrimination power of individual test item. Survey research design was adopted for this study. The study covered three primary schools (a private school, a public school and an Internally Displaced persons' (IDP) camp school). The name of the instrument for this study is Analysis of Adopted Cognitive Achievement Test (ACAT) comprised three tests namely verbal, nonverbal and quantitative. Cronbach Alpha formula was used to determine reliability index. On the whole, the ACAT can be said to have high reliability. Univariate ANOVA was used to find the difference in performance based on gender, age and school. All the items in tests 1, 2, 3 and 4 have difficulty levels greater than 0.25 thus all the items are significant (Brown, 1983). The discrimination power of items in test ACAT were all positive except for few (6) implying that the items were correctly answered by upper group pupils. Eight-three (%) pupils who wrote verbal ACAT, 94% of the pupils who wrote nonverbal ACAT and 85% of the pupils' equivalent IQ in quantitative ACAT all fall below average. The correlation between Academic performance and ACAT revealed no significant relationship with correlation coefficient 0.028 at 0.736 level of significance ($p < 0.05$). Only age had significant difference in performance on ACAT.

Introduction

Achievement tests are one of the most important aspects of teaching– learning process and the two most important characteristics of an achievement test are its reliability and content validity. For a test to be reliable and valid, a systematic selection of 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. Ebel(1972), stated that "Item analysis indicates the difficulty level of each item and discrimination between the better and poorer examinees.

According to Brown and Frederick (1971), Item analysis has two purposes: First, to identify defective test items and secondly, to indicate the content the learners have or have not mastered. Item analysis measures the effectiveness of individual test item in terms of its difficulty level and power to distinguish between high and low scorers in test .Thus it helps in selecting and retaining the best test items in the final draft of the test rejecting poor items and also show the need to review and modify the items. Item analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items. Difficulty index (P) and discrimination index (D) are two parameters which help to evaluate the standard of multiple choice questions (MCQ) used in an examination, with abnormal values indicating poor quality.

Designing MCQs is a complex and time consuming process in a multidisciplinary integrated curriculum. MCQs are used mostly for comprehensive assessment at the end of a term or academic session and provide feedback to the teachers on their educational action. Having constructed and assessed a test, a teacher needs to know how good the test questions are and whether the test items were able to reflect students' performance in the course related to learning. Because of their versatile character, MCQs are the most commonly used tool for assessing the knowledge capabilities of primary school pupils.

One of the major concerns in the construction of test items for an examination is ensuring the reliability of the test items. The item statistics can help to determine those items that are good and those that need improvement or deletion from a question bank. It allows any aberrant item to be given attention and reviewed. The findings of this paper have significance for student teachers and test developer. They should be very careful while selecting items.

Han, Kees-Jan and Denny (2014) defined 'Intelligence as what the Intelligence Test Measures'. Intelligence tests are classified in two ways - individual/group, aptitude/achievement. Aptitude tests measure the potential development of the testee, like a tool for predicting future performance. Achievement tests, on the other hand, measure mastery of a specific domain, and thus evaluate current performance, for example primary 4 pupils end of third term examination results. Abstract tests are mainly aptitude in type, while quarterly examinations in schools are mainly achievement in type. Intelligence tests also differ in how they are being administered, whether by individual or by group. Individual tests consider the interaction between the tester and the testee, and thus are more customized and personal. On the other hand, group tests are more economical, saving time, money and effort.

Reliability refers to the stability and consistency of scores the intelligence test produces. The Kuder-Richardson twenty (KR 20), as it was also called, soon became a classic and is index of the internal consistency of the test (Kimmo, 2000). "Internal consistency" refers to consistency of

students' responses across the items on the test. KR 20 can be thought of as a measure of the extent to which the items on a test provide consistent information about a students' level of knowledge of the content assessed by the test. Assuming that all the items on a test relate to a single content domain, we would expect students with a very high level of knowledge of the domain to answer most items correctly and students with a very low level of knowledge of the domain to answer most items incorrectly. The value of KR 20 can range from 0 to 1, with numbers closer to 1 reflecting greater internal consistency. Alternative forms of the formula were suggested e.g. by Horst (1953), but the most famous variation, known as alpha, was presented by Cronbach (1951).According to Freeman (1962), "Difficulty value of an item may be defined as the proportion of certain sample of subjects who actually know the answer of an item." It expresses the proportion or percentage of students who answered the item correctly. Item difficulty can range from 0.0 (none of the students answered the item correctly) to 1.0 (all of the students answered the item correctly). Experts recommend that the average level of difficulty for a four-option multiple choice test should be between 60% and 80%; an average level of difficulty within this range can be obtained, of course, when the difficulty of individual items falls outside of this range. If an item has a low difficulty value, say, less than .25, there are several possible causes: the item may have been wrongly keyed; the item may be too challenging relative to the overall level of ability of the class; the item may be ambiguous or not written clearly, there may be more than one correct answer. Index of difficulty for each test item can be calculated as:

Dv = (Ru + Rl) / (Nu + Nl)

Dv = item difficulty

Ru = the number of students in the upper 27% who responded correctly

Rl = the number of students in the lower 27% who responded correctly

Nu = the number of students in the upper group

Nl = the number of students in the lower group

Numerical value of Dv ranges from 0 to +1.00 with an optimal difficulty level of 0.62 (Thompson &Levitov, 1985). Higher the values of the difficulty index, easier the item is. According to Brown (1983) – In item difficulty, if most students answered an item correctly then the item was an easy one. If most students answered an item incorrectly then it should have been a difficult one. So, the items answered correctly by 100% or 0% of the examinees are insignificant.

Item discrimination is the degree to which students with high overall exam scores also got a particular item correct. It is often referred to as Item Effect, since it is an index of an item's effectiveness at discriminating those who know the content from those who do not.

Dp= (Ru - Rl)/ 0.5 N

Dp – discrimination power

N – total number of correct responses

Ru = the number of students in the upper 27% who responded correctly

Rl = the number of students in the lower 27% who responded correctly

Numerical value of discrimination power may range from -1.00 to +1.00. Ebel and Frisbie (1986) gave the following rule of thumb for determining the quality of items with respect to their discrimination index. If a test item is correctly answered by upper group students and incorrectly by lower group students then the item is said to have positive discrimination power. If a test item is correctly answered by lower group students and incorrectly by upper group students then the item is a negative discriminator. Item with negative discrimination decreases the validity of test and thus must be discarded. If a test item is answered correctly by equal number of upper and lower group students then its showing zero discrimination.

Mitraetal; (2009) reported poor correlation of discrimination index with difficulty index. Sim and Rasiah (2006) found the maximum discrimination power of items with the difficulty index .40 to .74.

Statement of the problem

Some basic forms of item analysis must have been carried out routinely by the designers of the Adopted Cognitive Ability Test (ACAT). However, the researcher has no idea of the reliability and validity of the tests, thus the need to subject the test to analysis. Hence the present research study was taken up with an objective to analyze the quality of the multiple choice questions used in the ACAT of primary four pupils, find out relationship between the item difficulty and item discrimination indices of these MCQ items in the ACAT. What is the relationship between scores on ACAT and the pupils' academic performance? Is there difference based on gender, age and school on pupils' performance in ACAT.

Objectives

The objectives of the study were:

1. To determine the reliability of the Adopted Cognitive Ability Test (ACAT).
2. To find out the item difficulty level and discrimination power of individual test items.
3. To find out the relationship between degree of item difficulty and corresponding power of discrimination of test items.
4. To develop norms for the ACAT
5. To determine the inter-relationship among the subtests contained in the ACAT.
6. To establish relationship between scores on ACAT and the pupils' academic performance in end of third term examinations.
7. To determine the difference based on gender, age and school on performance in ACAT.

Research Questions

The following research questions were asked:

1. What is the Reliability index of ACAT?

2. What are the difficulty level and the power of discrimination of the items on the ACAT MCQ Test
3. What is the out the relationship between degree of item difficulty and corresponding power of test items?
4. What is the norms for the ACAT development?
5. What is the inter-relationship among the subtests contained in the ACAT.
6. What is the relationship between scores on ACAT and the pupils' academic performance in end of third term examinations?
7. What is gender, age and school performance difference in ACAT ?

Hypotheses

The following hypotheses guided the study:

1. There is no significant relationship between degree of item difficulty and corresponding power of discrimination of test items.
2. There is no significant relationship between scores on ACAT and the pupils' academic performance in end of third term examinations.
3. There is no significant relationship among the subtests contained in ACAT.
4. There is no significant difference across gender, age and school in performance of primary four pupils on ACAT.

Methodology

Research Design: Survey research design was adopted for this study

Population The study covered three primary schools (a private school, a public school and an Internally Displaced persons' (IDP) camp school). The study was delimited to primary four pupils of the said schools in the 2020/2021 academic season.

Sample and Sampling Technique: Random sampling method was adopted to select a sample of 146 primary school pupils Sample included the students of both genders who came from a public, private and internally displaced persons' camp (IDPs) in Jere, Maiduguri Metropolis.

Research Instrument

Analysis of Adopted Cognitive Achievement Test (ACAT) comprised three tests namely verbal, nonverbal and quantitative. Verbal test had four subtests each having a set of 15 questions (a total of 60 items). Nonverbal had three subtests: subtests 1 and 2 both had 35 questions each while subtest 3 had 8 questions (a total of 78 items). Quantitative test also had three subtests: subtests 1 and 2 had 15 questions each while subtest 3 had only 11 (a total of 41 items). The ACAT had a total of 179 items in all. Instructions were made clear on the test and a separate answer sheet was used to fill the answers to the three different tests.

Researcher herself conducted the test on a sample of 146 pupils from public, private and IDPs' schools. Pupils were given as much time as they required to complete the tests. Pupils were instructed to fill their answers only in the answer sheet provided.

The procedure involved the following steps:

Determining the reliability index of each subtest. Identification of upper 27% and lower 27% examinees having highest and lowest scores in rank order respectively on the total test.

Calculation of each item, of the proportion of the examinees attempting it correctly. The difficulty level (p) and discrimination index, (D) were calculated using afore mentioned formulae. Raw scores were converted to IQ scores. Performance on ACAT was correlated against academic performance of primary 4 pupils. Univariate ANOVA was used to find the difference in performance based on gender, age and school.

Validity and Reliability of the Instrument: The instrument was validated for face and content validity by professionals in the field of curriculum planning and development and testing and measurement in the Department of Education, University of Maiduguri. Split-halve reliability was used to test the reliability of the instrument among 20 pupils in a school not included in the study. Cronbach Alpha formula was used to determine reliability index for the full test, r = 0.71, 0.8 (see table 1).

Method of Data Analysis

Descriptive statistics of frequency counts and percentages were used to descript demographic characteristic of the respondents, while Univariate ANOVA was used to find the difference in performance based on gender, age and school was used to test the hypotheses at $p< 0.05$ level of significant.

Results

The results of data analyses are presented as follows:

Table 1:
Reliability Index of ACAT Subtests

Variable	Cronbach Alpha	Standardized Alpha
Verbal test I	0.659	0.652
Verbal test II	0.699	0.697
Verbal test II	0.766	0.765
Verbal test IV	0.488	0.474
Non Verbal test I	0.823	0.818
Non Verbal test II	0.778	0.763
Non Verbal test III	0.425	0.414
Quantitative test I	0.710	0.710
Quantitative test II	0.799	0.799
Quantitative test III	0.663	0.641

Note: HR = highly reliable, NR = Not reliable

Table 1 on reliability of ACAT showed that only verbal test 4 and nonverbal test 3 had reliability index of 0.488 and 0.425 respectively. Of the ten subtests in ACAT, only two had low reliability index. On the whole, the ACAT can be said to have high reliability.

Table 2:

Item difficulty and discrimination index of verbal ACAT

	TEST 1		TEST 2		TEST 3		TEST 4	
Q/NO	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)
1	0.81	0.12	0.72	0.08	0.63	0.65	0.65	0.53
2	0.51	0.76	0.74	0.08	0.55	0.70	0.62	0.47
3	0.51	0.86	0.55	0.12	0.55	0.82	0.45	0.95
4	0.50	0.71	0.58	-0.02	0.62	0.66	0.54	0.67
5	0.37	0.41	0.37	-0.14	0.54	0.70	0.49	0.52
6	0.44	0.71	0.28	-0.11	0.54	0.68	0.27	0.29
7	0.33	0.44	0.36	-0.17	0.63	0.56	0.26	0.62
8	0.42	0.64	0.47	0.14	0.72	0.36	0.49	0.62
9	0.50	0.89	0.37	0.04	0.42	0.50	0.45	0.59
10	0.40	0.56	0.38	0.21	0.72	0.36	0.55	0.80
11	0.50	0.69	0.08	0.17	0.56	0.56	0.56	0.44
12	0.27	0.35	0.29	0.05	0.37	0.63	0.54	0.55
13	0.47	0.41	0.45	-0.03	0.51	0.49	0.47	0.55
14	0.54	0.58	0.32	-0.20	0.54	0.57	0.36	0.57
15	0.58	0.56	0.31	0.13	0.42	0.34	0.32	0.72

Table 2 shows the difficulty level and discrimination powers of verbal tests 1, 2, 3 and 4. It can be seen that all the items in tests 1, 2, 3 and 4 have difficulty levels greater than 0.25 thus all the items are significant (Brown, 1983). The discrimination powers of items in test 1 are all positive implying that the items were correctly answered by upper group pupils. For test 2, items 4, 5, 6, 7, 13 and 14 had negative values implying lower group pupils answered the items correctly. The items in tests 3 and 4 have positive discrimination powers. This implies items were answered correctly by upper group pupils.

Table 3:
difficulty and discrimination index of nonverbal test 1 and 2

	TEST 1		TEST 2		TEST 3	
Q/NO	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)
1	2.05	-1.47	0.63	0.27	0.26	0.86
2	0.74	0.29	0.45	0.42	0.39	0.41
3	0.38	0.14	0.42	0.81	0.50	0.73
4	0.53	0.57	0.47	0.81	0.50	1.03
5	0.42	0.53	0.47	0.61	0.16	1.11
6	0.53	0.57	0.46	0.47	0.39	0.75
7	0.24	-0.42	0.46	0.81	0.16	0.83
8	0.49	0.70	0.43	0.58	0.47	0.68
9	0.26	0.52	0.54	0.74		
10	0.38	0.57	0.51	0.74		
11	0.57	0.63	0.37	0.86		
12	0.55	0.61	0.45	0.66		
13	0.43	0.67	0.41	0.78		
14	0.41	0.67	0.47	0.86		
15	0.37	0.44	0.49	0.46		
16	0.42	0.50	0.42	0.56		
17	0.58	2.00	0.45	2.00		
18	0.28	0.50	0.37	0.56		
19	0.55	2.00	0.34	2.00		
20	0.68	2.00	0.41	2.00		
21	0.39	0.44	0.30	0.33		
22	0.24	2.00	0.21	2.00		
23	0.34	2.00	0.39	2.00		
24	0.25	0.48	0.29	0.65		
25	0.55	2.00	0.18	2.00		
26	0.34	2.00	0.28	2.00		
27	0.47	0.41	0.38	0.16		
28	0.26	2.00	0.28	2.00		
29	0.20	2.00	0.14	2.00		
30	0.08	0.47	0.12	0.13		
31	0.20	2.00	0.30	2.00		
32	0.32	2.00	0.32	2.00		
33	0.47	0.14	0.26	0.15		
34	0.08	2.00	0.21	2.00		
35	0.29	2.00	0.25	2.00		

Table 3 shows the difficulty level and discrimination powers of nonverbal tests 1, 2 and 3. It can be seen that all the items in tests 1, 2 and 3 have difficulty levels greater than 0.25, thus all the items are significant (Brown, 1983). The items in nonverbal ACAT tests 1, 2 and 3 all have positive

discrimination powers except for items 1 and 7 in test 2. This implies all items were answered correctly by upper group pupils.

Table 4:
Item difficulty and discrimination index of quantitative ACAT

	TEST1		TEST2		TEST3	
Q/NO	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)	Dif.(p)	Dis.(D)
1	0.56	0.00	0.61	0.58	0.54	0.60
2	0.68	0.45	0.50	0.60	0.42	0.38
3	0.64	0.53	0.58	0.62	0.83	0.22
4	0.46	0.41	0.56	0.58	0.49	0.70
5	0.54	0.66	0.61	0.47	0.60	0.53
6	0.51	0.59	0.51	0.46	0.08	-0.09
7	0.61	0.53	0.53	0.53	0.13	0.22
8	0.56	0.43	0.58	0.51	0.53	0.60
9	0.63	0.51	0.46	0.51	0.50	0.63
10	0.60	0.45	0.56	0.55	0.51	0.62
11	0.60	0.33	0.50	0.40	0.19	0.31
12	0.39	0.34	0.56	0.58		
13	0.53	0.34	0.49	0.36		
14	0.58	0.44	0.47	0.53		
15	0.58	0.48	0.43	0.41		

Table 4 shows the difficulty level and discrimination powers of quantitative tests 1, 2 and 3. It can be seen that all the items in the test have difficulty levels greater than 0.25, thus all the items are significant (Brown, 1983). The items in tests 1, 2 and 3 have positive discrimination powers. This implies all items were answered correctly by upper group pupils.

Table 5:
Relationship between Item difficulty (p) and discrimination power index **(d)**

Subtest	Correlation Coefficient	Remark
Verbal test 1	-0.129	No correlation
Verbal test 2	-0.129	No correlation
Verbal test 3	-0.154	No correlation
Verbal test 4	0.022	Weak positive correlation
Nonverbal test 1	-0.476	Moderate negative correlation
Nonverbal test 2	-0.439	Moderate negative correlation
Nonverbal test3	-0.361	Weak negative correlation
Quantitative test 1	0.170	No correlation
Quantitative test 2	0.487	Moderate positive correlation
Quantitative test 3	0.533	Moderate positive correlation

Pearson correlation coefficient "r" calculated as seen in Table 5 showing a moderate negative relationship between values of p and D. This negative correlation signifies that as the difficulty index increased discrimination index also increase but to an optimum value only after which discrimination power decrease with the increase in difficulty level. This suggested that the easier items (>.80) or too difficult items (<.20) poorly discriminate between the superior and inferior examinees. (Angoff, 1984). The difficulty indices and discrimination indices were most often reciprocally related. This is the case in verbal tests 1, 2, and 3 as well as in nonverbal tests 1, 2 and 3. However, it is not the case in verbal test 4, quantitative tests 1, 2 and 3. Therefore the null hypothesis is rejected for verbal and nonverbal ACAT but rejected for quantitative ACAT.

Equivalent IQs of the raw scores on ACAT

One of the popular scales in use is the Cattell's scale which uses a standard deviation of 24. According to Cattell's scale, IQ is classified as:

Over 160 - Genius Level
140 - 159 - Highly Intelligent
120 - 139 - Above Average
100 - 119 – Average
90 - 99 - Below Average

Table 6:
Summary of Equivalent IQ of raw score in ACAT

Equivalent IQ	VERBAL	NONVERBAL	QUANTITATIVE
Over 160	0	0	0
140 – 159	0	0	0
120 – 139	19	9	20
100 – 119	44	69	44
90 – 99	42	25	32
Below 90	41	40	40

Table 6 is a summary of the equivalent IQ of the raw scores obtained in ACAT. It can be seen that most of the pupils fall within the IQ range of 99 and 120. Some of the pupils scores fall below IQ 90 (verbal- 41, nonverbal- 40 and quantitative- 40). There were 19, 9 and 20 pupils above average IQ (120-139) in verbal, nonverbal and quantitative respectively. 44, 69 and 44 pupils fall within average IQ (100-119) in verbal, nonverbal and quantitative respectively. Verbal, nonverbal and quantitative recorded 42, 25 and 32 respectively for below average IQ (90-99). 83%of pupils who wrote verbal ACAT had below average IQ, 94% of the pupils had below average performance in nonverbal ACAT and 85%of the pupils' equivalent IQ in quantitative ACAT fall below average.

HO_2: There is no relationship between scores on ACAT and the pupils' academic performance in end of third term examinations.

Table 7:
Relationship between ACAT and the Pupils' Academic Performance

Variable	Mean	Standard Deviation	Correlation Coefficient	Sig p<0.05
Academic Per.	61.027	26.479		
ACAT	89.53	47.242	0.028	0.736

The correlation between Academic performance and ACAT revealed no significant relationship with correlation coefficient 0.028 at 0.736 level of significance ($p < 0.05$). The null hypothesis is therefore accepted that there is no relationship between scores on ACAT and the pupils' academic performance in end of third term examinations.

HO_3: There is no significant relationship among the subtests contained in ACAT.

Table 8:
Relationship between ACAT Subtests (Correlations)

		VBV	VSC	VBS	VBA
VBV	Pearson Correlation	1	.330**	.439**	.482**
VSC	Pearson Correlation	.330**	1	.555**	.477**
VBS	Pearson Correlation	.439**	.555**	1	.629**
VBA	Pearson Correlation	.482**	.477**	.629**	1

		NOV	NVT	SET
NOV	Pearson Correlation	1	.537**	.031
NVT	Pearson Correlation	.537**	1	-.077
SET	Pearson Correlation	.031	-.077	1

		QTR	NBS	QET
QTR	Pearson Correlation	1	.545**	.415**
NBS	Pearson Correlation	.545**	1	.521**
QET	Pearson Correlation	.415**	.521**	1

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

Table 10 answers the hypothesis on relationship among subtests in ACAT. It shows that there is significant relationship among the subtests in verbal ACAT (VBV-VSC .330, VBV-VBS 0.439, VBV-VBA .482 etc). For nonverbal ACAT, significant relationship is seen between NOV-NVT .537 only with no relationship between NOV-SET and NVT-SET. For quantitative ACAT, the results reveal significant relationship among all three subtests: QTR-NBS 0.545, QTR-QTE 0.415 and QET-NBS .521.Correlation significant at .01 level. Thus, for verbal and quantitative ACAT, the null hypothesis is rejected but for nonverbal ACAT, it is accepted.

Table 9:
Difference in Gender, Age and School in Performance of Primary Four Pupils on ACAT

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	26850.176	1	26850.176	13.809	.000
	Error	132030.631	67.902	1944.441[a]		
QUANTITATIVE	Hypothesis	75732.895	31	2442.997	1.401	.217
	Error	34886.135	20	1744.307[b]		
NONVERBAL	Hypothesis	91940.333	42	2189.056	1.255	.297
	Error	34886.135	20	1744.307[b]		
VERBAL	Hypothesis	74330.994	39	1905.923	1.093	.428
	Error	34886.135	20	1744.307[b]		
AGE	Hypothesis	34927.842	8	4365.980	2.503	.046
	Error	34886.135	20	1744.307[b]		
School	Hypothesis	457.049	2	228.525	.131	.878
	Error	34886.135	20	1744.307[b]		
Gender	Hypothesis	22.791	1	22.791	.013	.910
	Error	34886.135	20	1744.307[b]		

From Table 9 it can be seen that only age had significant difference in performance on ACAT. There was no difference seen in the performance of pupils in ACAT based on gender and school. Thus the null hypothesis is accepted; that there is no significant difference based on gender and school in performance on ACAT, but it is rejected in respect of age because the analysis revealed significant age difference in performance on ACAT.

Discussion

The reliability index of the subtests on ACAT showed high internal consistencies. Most of the test items also fall within acceptable ranges for difficulty and discrimination as put forward by Brown (1983) and Ebel and Frisbie (1986). Most of the test items in ACAT had positive values inferring that the questions were mostly answered by upper group pupils. The analysis revealed that there is reciprocal relationship between item difficulty and discrimination power. Developing and administering Multiple Choice Questions on the content knowledge of research methods in education helps teacher educators in moulding future teachers. Hitherto item analysis is an important phase in the development of a test or instrument.83% of pupils who wrote verbal ACAT had below average IQ, 94% of the pupils had below average performance in nonverbal ACAT and 85% of the pupils' equivalent IQ in quantitative ACAT fall below average. Thus general IQ of pupils in ACAT is below average. The correlation between Academic performance and ACAT revealed no significant relationship with correlation coefficient 0.028 at 0.736 level of significance ($p < 0.05$).The analysis revealed that there was significant relationship among the subtests in verbal ACAT (VBV-VSC .330, VBV-VBS .439, VBV-VBA .482 etc). For nonverbal ACAT, significant relationship was seen between NOV-NVT .537 only with no relationship between NOV-SET and NVT-SET. For quantitative ACAT, the results reveal significant relationship among all three subtests: QTR-NBS

.545, QTR-QTE .415 and QET-NBS .521.There was no difference seen in the performance of pupils in ACAT based on gender and school.

Conclusion

The principle function of an instrument used in any educational research is to infer student's capacities and it offers information on which to base the making of correct decisions. Developing and administering Multiple Choice Questions on the content knowledge of research methods in education helps teacher educators in in moulding future teachers. Hitherto item analysis is an important aspect of assessment. There was no significant relationship between primary 4 pupils' academic performance and performance in ACAT. This could be as a result of difference in quality of the two tests since the reliability and validity of the third tem examinations were unknown. There was also no difference in gender and school but there was difference in age. Therefore it could be that age is an important factor in performance of primary four pupils.

Recommendations

Teacher Made Tests in primary schools should be subjected to item analysis to ensure quality items are used for internal assessments. This may enhance performance of the pupils in external assessments like the ACAT.

References

Angoff W. (1984). Scales, Norms and Equivalent scores. Educational Measurement Origins. Theories and Explications 2, 121.

Blood, D.F. and W.C. Budd, (1972). Educational measurement and evaluation. New York: Harper and Row.

Brown, and Frederick G. 1981, "Measuring Classroom Achievement", Halt Richard and Winston, U.S.A.", c, pp. 101-110, 224p.

Brown, F. (1983), "Principles of Educational and Psychological Testing," Third edition, NY: Holt, Rinehart & Winston. Chapter 11.

Ebel, R.L.(1972) Essentials of Educational Measurement (1st Edition). New Jersey: Prentice Hall.

Ebel, R.L., & Frisbie, D.A.(1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Freeman Frank S. (1962).Theory and Practice of Psychological Testing, New Delhi: Oxford & Ibh publishing

Han L. J. van der Maas, Kees-Jan Kan & Denny Borsboom (2014). Intelligence Is What the Intelligence Test Measures. Seriously.Journal of Intelligence, 2, 12-15.

Kimmo Vehkalahti (2000). Reliability of Measurement Scales.Statistical Research Reports 17, Finnish Statistical Society.

Mitra,N.K., Nagaraja, H.S., Ponnudurai, G. & Judson, J. P. (2009).The levels of difficulty and discrimination indices in type A multiple choice questions of Pre-clinical Semester 1 multidisciplinary summative tests. IeJSME, 3, 1, pp. 2-7.

Si-Mui-Sim and Rasiah R.I. (2006) Relationship between item difficulty and discrimination

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