Original Research Article

Application Of Classical Test Theory (CTT) in the validation of teacher made Mathematics Multiple Choice Test (MMCT) items

ABSTRACT

Multiple choice items are the most popular item format used in Senior High schools in Ghana and even beyond for almost all subjects which mathematics is not an exception. The quality of the items constructed by the classroom teachers is essential. Stakeholders depend on the results for decision making. Little is known about the quality of the teacher made items. This study therefore sought to find out if the multiple-choice items used by the mathematics teachers meet the criteria for a good assessment instrument using the classical test theory approach. Instrumentation research design was used. A purposive and simple random sample technique were used to select one intact class with a class size of 35 at Baidoo Bonsoe Senior High School in the Ahanta West Municipality of Ghana for the study. A multiple-choice test in core mathematics with items constructed by the mathematics teachers of the school was used. Classical test theory was used to estimate the content related evidence of validity, reliability, difficulty index, and discrimination indices and distracter analysis of the items. It was found that only three items measured high level cognitive thinking, 14 items were observed to have acceptable of difficulty and 28 items discriminated between the upper and lower groups at varying degrees. It also, found that only one out of 10 was without issue of distracters and that reliability of the scores of the test was 0.73. It was therefore recommended that training in test construct should be intensified both whiles the teachers are the training institution and on the job.

Keywords: **Classical test theory, item distracters, item difficulty, item discrimination**

1.0 Introduction

The term assessment is used in every institution in recent times. Every organization, now seeks to examine the worth of either policy, product, staff, students in the case of education and many more. According to Heale and Twycross (2015) and Etsey (2012), assessment is a process of obtaining information for decisions making. This explains why the concept of assessment is used in almost every institution. Where ever there is decision making based on obtained information, there is assessment (Shete, Kausar, Lakhkar & Khan, 2015). The process of obtaining information to make decision about students, programme, policies and curriculum is term as assessment in school. Nitko (2012) therefore defined assessment as a "systematic process of gathering information that is educationally relevant to make legal and instructional decisions about the provision of special services" (pg 99). The definition focuses on education. Nitko continued that, assessment has stages, activity and outcome. The stages are the processes the assessor goes through, the activity is the gathering of the information and the outcome concerns the decision made as a result of the gathered information on the phenomena. Assessment is seen as the process of obtaining or gathering information on the student for decision making. The definition of assessment is limited to obtaining information on the student.

Different test formats; objective and essay, exist for use in the classroom. The objective test comprises multiple choice, true or false, matching, fill in and short answers (Asamoah-Gyimah & Anane, 2018). Generally, objectives test, especially the multiple-choice

format is mostly used because it is essay to score, high content validity, suitable for a large population and susceptible to statistical analysis (Nitko, 2001). As a result, most assessment at the Senior High School (SHS) level both internal and external has a section for objective test which is mostly multiple choice. It is therefore important that the multiples choice items used by the test meet the criteria for a good assessment for accurate decision making on the students.

Psychometric criteria for determining the technical adequacy of measurements are widely established. Criteria derived born out of the fundamental ideas of reliability and validity is particularly important, but given the benefits of many new techniques to evaluation, expanding on their traditional conceptions seems reasonable. Validation, according to Messick as cited in Annan-Brew (2020), entails the establishment of a consequential basis for interpretation of the test score and usage in addition to the more traditional evidential basis. The classical test theory (CTT), the mother of all test theories is the most common method of validation of assessment instrument.

The validation of the items constructed by the classroom teacher is necessary as it provides basis for dependability of the results for decision making. In the Ghanaian classroom, validation of teacher mase test is not given attention. Little is known about the quality of the teacher made items. It therefore important to validate the teacher made test in mathematics to ascertain whether the result could be trusted or not.

Literature Review

1.1 Classical Test Theory and Validation of Assessment Results

According to the classical test theory, any observed test score is a function of two hypothetical components: a true score and a random error. Mathematically, it is expressed as: X = T + E; where X is the observed test score, T is the true score of the individual, and E is the random error. The observed is the score that is seen on the test paper. The true score is the expected value of the observed value of the observed score when the construct is measured repeatedly. The error score is the difference between the individual's observed score and his/her true score (Bichi, 2016). This therefore means that it is the error that distort the equalization of the true score and observed score. When the error is neutralized, individual's score true score and observed will the same when measured repeatedly. Reliability is theoretically defined as the ratio of the variance of the true score to the variance of the observed score (Amedahe & Asamoah-Gyimah, 2015). Mathematically, it is

$p^2_{xx} = \frac{\sigma^2_T}{\sigma x^2}$ expressed as

This implies that reliability tells the extent to which the observed score variance is close to true variance. A perfect reliable test is one with zero error score and that observed score and true score are equal. The reliability of test is +1. As the error increase, the reliability reduces.

The classical test theory considers two factors - content and item characteristics in developing test items (Hamleton, Swaminathan & Rogers as cited in Bichi, 2016). The content related issue is established with expert judgment of content relevance and representation. The other factor considered-item characteristics focus on the difficulty level and the discrimination index of the items. Items that meet difficulty level index (0.30 to 0.70) and discrimination index (0.30 to 0.70 for standardized test) are kept in the item bank. Highly discriminating item is most desirable and level of difficulty depends on purpose for the test. For example, a test for selection would have a high difficulty index greater than 0.80. the focus of the classical test theory is to zero the error so that the true becomes equal to the true. Therefore, sources of errors are considered under sources of error: content, item analysis

(difficulty, discrimination and distracter), reliability and bias. Any, fouls in these sources of evidence introduces error into the observed score.

1.2 Content related

This evidence is about the content representativeness and relevance of the assessment results. Content-related evidence of validity is assessed by showing the degree to which the content of assessment results represents the content about which conclusions are to be drawn. The judgement on content relevance focuses on whether tasks included in the assessment are in the test domain definition. The relevance of the assessment results is the extent to which the assessment matches the school's curriculum target. There should be an overlap between the assessment domain and the curriculum. The weight given to each content area should be appropriate to the local curriculum (Nitko, 2004). According to Nitko, to ensure content validity, the items should have the following characteristics: (1) reflect current thinking of the subject matter of what is essential to teach and assess (2) accurately represent the subject matter (3) keyed correctly and (d) contain meaningful and relevant content.

To judge whether as assessment, the content has related evidence to support the interpretation and uses of the assessment results, table of specification is prepared and use (Nitko, 2004). The table of specification is a means of defining the domain for standardized position on achievement test. It contains the major content areas and skills to be assessed and the percentage of tasks content-skills. The presenter then prepared a test specification to guide in the construction of the items to ensure content validity as stated by Amedahe and Asamoah-Gyimah (2003), who said, table of specification is a two-way chart showing the subject matter content and learning outcomes established for the instructions. They further stated that by inspecting the test/table of specification, content validity which measures how representative the scores of a test represent all the domain of learning is determined.

1.3 Item Analysis

According to Nitko (2001), item analysis refers to the process of collecting, summarizing and using information from students' responses to make decision about each assessment task or item. This means that item analysis is focused on critically examining each test item in order to make decision about the item. Etsey (2012) holds a similar view that the purpose of item analysis is to check appropriate difficulty level, irrelevant cues and other defects and distracters effectiveness in multiple–choice items. From the definitions, item analysis is not a one-shot event but series of events and the purpose is to critically examine the responses students provided to each item in order to determine the state of the item in relation to the instructional goals.

According to Bichi (2016), item analysis within the classical approach often relies on two statistics: the P-value (proportion) and the item-total correlation (point-biserial correlation coefficient). The P-value represents the proportion of examinees responding in the keyed direction, and is typically referred to as item difficulty. The item-total correlation provides an index of the discrimination or differentiating power of the item, and is typically referred to as item discrimination (Shete, Kausar, Lakhkar & Khan, 2015). In addition, these statistics are calculated for each response of the oft-used multiple choice item, which are used to evaluate items and diagnose possible issues, such as a confusing distracter.

1.3.1 Item difficulty (P)

According to Liaquat, Asif, Siraji and Maroof (2012), item difficulty means the percentage of students who answer correctly each test item. Item difficulty indices is an indication of the proportion of the examinees who responded to the item correctly. The lesser the proportion, the difficulty the item is. It is calculated by dividing the number of students who answer the

item correctly by the total number of examinees. Mathematically, $P = \frac{R}{T}$ where R = number of students who answer the item correctly and T= total number of examinees. The P index ranges from 0 (when no student correctly answered the item) to 1 (when all students answer the item correctly). This indicates that the smaller the p index, the more difficult the item and the greater the p index, the less difficult the item. Allen and Yen (1979), suggested that an effective or ideal or a good item should have a p-index ranging from 0.30 to 0.07, a more difficulty item should a P-index below 0.30 and an item with a P-index above 0.70 is considered to be too easy.

1.3.2 Item discrimination (D)

As Nitko (2001) puts it, item discrimination (D) is the difference between the fraction of the upper group answering the item correctly and the fraction of the lower group answering the item correctly. The D-index indicates the extent to which the item is able to differentiate between higher achieving students and lower achieving students (Shete, Kausar, Lakhkar & Khan, 2015).

According to Nitko (2001), item discrimination is important because it is able to indicate both the absolute achievement and relative achievement of the students. By absolute achievement, item discrimination is able to determine the level of subject matter a student has accurately learned. Relative achievement means that item discrimination is able to determine the relative rank of each student either in the upper or middle or the lower group. The D-index could be negative or positive and ranges from -1 to +1. It is negative when proportion of lower group answering the item correctly is greater than proportion in the upper group answering the item correctly (Pande, Santosh, Pande, Parate, Nikam & Agrekar, 2013). When proportion in upper group answering the item correctly is greater than proportion in lower group answering the item correctly, a positive D-index is obtained. A good item is one with a D-index greater than 0.30 but high positive D-indices are greatly used by developers of standardized test (Etsey, 2012). D-index is presented as $D = P_U - P_L$, where P_U = proportion in upper group answering the item correctly and P_L = proportion in lower group answering the item correctly.

P_U/P_L = number in upper/lower group answering the item correctly
Total number of students in upper/lower group.

To be able to determine the D-index for each item, examinees are groups based on their total scores. For a small class size, the two groups are created, upper and lower group. For a large sample size, 30 and above, the first and last 30% forms the upper and lower groups respectively. Then the difference in the proportion of the upper and lower groups that answered the item correct is estimated as discrimination index.

1.3.3 Analysis of Distracters

According to Annan-Brew (2020), distracter analysis is an investigation into the keys to a multiple-choice item as whether the options functioned as intended. A distracter is incorrect option which is attracted to the uninformed. The quality of the items depends partly on the effective functioning of the distracters selected by the examinees (Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010). A good distracter should attract at least one examinee. A good distracter must plausibly be attracted to the uninformed examines (Amedahe & Asamoah-Gyimah, 2003). The function of the distracters is to determine whether examinee really knows the correct answer to the item.

1.4 Reliability

As put by Etsey (2012), reliability is the degree of consistency of assessment results. Osterlind (2006), also states that reliability is a proportion and gives an indication of how

much errors are in a set of scores. That is reliability is related to errors in scores. Etsey (2012) further explained that errors are the reasons for the inconsistency in scores. The CTT states that observe scores approaches the true score as error reduces.

1.4.1 Methods of Estimating Reliability

The source of error under consideration gives the different methods for estimating reliability (Liaquat, Asif, Siraji & Maroof, 2012). A number of methods are available for testing reliability, but the most commonly used ones are:

1.4.4.1 Test-retest method

The test-retest method is a measure of stability and considers scores of students over a period of time. The same test is given to a group of students twice within an interval ranging from several minutes to years. The scores on the two administrations are correlated and the result is the estimate of the reliability of the test (Etsey, 2012).

1.4.4.2 Equivalent forms method

The equivalent-form is used to estimate reliability by giving two forms (with equal content, means, and variances) of a test to the same group either on the same day or a later day and correlating the results (Brennan, 2006). With this method, one determines how confident an examinee scores could be generalized to what the examinee would receive if the examinee took a test made up of similar but different items. In this case, it is the changes due to the specificity of knowledge that is measured and not changes from one time to another.

1.4.4.3 Measure of internal consistency

Measure of internal consistency has several types.

a. Split-half method: To start with the split-half method, a single test is given to the students. The test is then divided into two halves for scoring (Kulas & Stachowski, 2009). The two scores for each set of students are correlated to obtain the estimate of reliability. The test can be split into two halves in several ways. These include using odd-even numbered items, and also first-half and second-half (Nitko, 2012). The split-half method aids practicing teachers to understand the true performance of the student.

b. Cronbach alpha method: Cronbach alpha method is the average split-half correlation based on all possible divisions of a test into two parts (Nitko, 2012). This internal consistency is used when test items are scored pass-fail or when more than one point is awarded for a correct response (Salvia & Yesseldyke, 2001).

c. Kuder and Richardson (KR-20 and 21): Kuder and Richardson (KR-20 and 21) is more restricted method of estimating a test's reliability which is based on the average correlation between all possible split halves. Under KR-20 and 21, test items are scored dichotomously (that is, items that can be scored only right and wrong). It estimates the reliability of the scores from a single administration of a test (Nitko, 2001). According to Osterlind (2006), K-R20 is used to estimate the extent to which performance on an item relates to the overall test scores. It also used to determine if all items measure the same trait or students' performance on each item. Nitko (2001) indicated that internal consistency is founded on the idea that the consistency with which students respond from one assessment task to the next provide the basis for estimating the reliability coefficient for the total scores.

In this study, the Kuder and Richardson method was used to estimate the reliability of the teacher made test. A single test was used hence the reliability method rest in the internal consistency methods. The KR-20 and 21 gives a better result than the spilt-havles (Nitko, 2001).

1.5 Research questions

The purpose of the study was to validate teacher made mathematics test. Research the following research questions guided the study:

1. What is the content validity of Mathematics Multiple Choice Test (MMCT)?
2. What is the difficulty index of Mathematics Multiple Choice Test (MMCT)?
3. What is the discrimination index of Mathematics Multiple Choice Test (MMCT)?
4. What are the distracter indices of Mathematics Multiple Choice Test (MMCT)?
5. What is the reliability of Mathematics Multiple Choice Test (MMCT)?

2.0 Methodology

This study employed instrumentation research design. The design was used to collect information to validate the multiple-choice test developed by the teachers. In this study, teachers were made aware of the purpose of the study and were asked to develop a 30-items test for the study. The test for the study was produced by the HOD (mathematics). The test was accompanied with a scoring rubric as expected for assessment (Nitko, 2001)

The population for the study was SHS two students Baidoo Bonose SHS in the western region of Ghana, totalling 312. There were six form two classes on campus at the time of the study. The study made use of simple random and purposive sampling techniques. A purposive sample technique was used to select form two for the study. This was because, at the time of the study, it was the highest class on campus. It is believed that they have learned and completed much content than the first years. A simple random sample technique was to select one intact class for the study. With that the form 2 science with the class size of 35 made up of 12 girls and 23 boys was selected for the study

The instrument for the data collection of the study was a mathematics test developed by the mathematics teachers and called Mathematics Multiple Choice Test (MMCT). The test consisted of 30 multiple choice items with four options each. The instrument covered all the content learnt for the semester.

The test was administered to the selected SHS Two class during the end of semester examination period. This was to obtain on information to validate the test that has been developed by the mathematics teachers. Students were asked not to write their names on the scripts but their student number. This was to ensure confidentiality of students' score

The test was scored using the scoring rubric submitted by the HOD. Scripts were coded for easy detection of error in data entry. Table of specification was used to analyse research question one. Classical test theory was used to judge the quality level of the test. The Kuder and Richardson (KR-20 and 21) was used to estimate the reliability of the test.

3.0 Results and Discussion

Research Question 1: What is the content validity of Mathematics Multiple Choice Test (MMCT)?

A table of specification was prepared for the test items to find out the degree of content validity. The results of the table of specification presented in Table 1.

Table 1: table of specification for the MMCT

Topic	Knowl.	Comp	Appl.	Anal.	Syn.	Eva.	Total
Surd		1	1				2
Indices		1	1				2
Sets		1	1	1			3
Vectors and bearing	1	1	3				5
Log and Number bases	1	1	2				4
Linear equation and inequalities		1	3				4
Relation and functions	1		2	1			4
Numbers and numerals	2		1				3
Algebraic expression		2		1			3

Totals	5	8	14	3	30

Table 1 presents the table of specification for the MMCT. The table showed many of the items were on application of knowledge as given for mathematics. However, only three items measured high level.

Research Question 2: What is the difficulty index of Mathematics Multiple Choice Test (MMCT)?

The table below shows the P-index of each of the 30 items.

Table 2: item difficulty

No.	R	P	No.	R	P
1	23	0.66	16	10	0.29
2	30	0.86	17	28	0.80
3	14	0.40	18	22	0.63
4	20	0.57	19	21	0.60
5	19	0.54	20	31	0.89
6	29	0.83	21	29	0.83
7	26	0.74	22	8	0.23
8	21	0.60	23	27	0.77
9	27	0.77	24	31	0.89
10	29	0.83	25	11	0.31
11	11	0.31	26	13	0.37
12	5	0.14	27	17	0.49
13	21	0.60	28	13	0.37
14	25	0.71	29	30	0.86
15	23	0.66	30	35	1.00

Key R = Total number of students who correctly answered the item
P = Proportion Correct (P = R/T)

Table 2 presents the item difficulty. The table reveals that 16 of the items comprised of 3 extremely difficult items (12, 16 and 22) and 13 easy items (2, 6, 7, 9, 10, 14, 17, 20, 21, 23, 24, 29, and 30) on the tests need to be either modified or taken out of the test. However, 14 items (1, 3, 4, 5, 8, 11, 13, 15, 18, 19, 25, 26, 27 and 28) were observed to be effective and should be maintained in the item bank.

Research Question 3: What is the discrimination index of Mathematics Multiple Choice Test (MMCT)?

Table 3: distribution of students' scores according to ability levels

Student ID	Scores	Level	Student ID	Scores	Level	Student ID	Scores	Level
011	28	U	015	19	M	025	16	L
020	25	U	026	19	M	016	16	L
003	24	U	017	19	M	027	16	L
014	24	U	018	19	M	028	15	L
032	24	U	029	18	M	019	15	L
006	21	U	002	18	M	030	14	L
030	21	U	021	18	M	031	14	L
008	21	U	022	18	M	005	13	L
004	20	U	013	18	M	033	12	L
010	20	U	024	17	M	012	9	L

001	20	U	007	7	L
034	20	U			
023	20	U			
009	20	U			

Table 3 presents distribution of students' scores according to ability levels. The table reveals that 14 students fell in the upper group, 10 in the middle group and 11 in the lower group. This was done by calculating 0.33(33%) of the 35 students for each group after the scores have arranged from largest to smallest as suggested by Tamakloe, Atta and Amedahe (1996). The upper group was 11.55 approximately the first 12 students but the 12^{th} student has a score of 20 therefore all students who scored 20 fell in the upper group making it 14 in number. The upper and middle groups sum up to approximately 24 students therefore 14 are in the upper group leaves 10 students in the middle group with the last 11 in the lower group.

Table 4: item discrimination index of each item

Item	R_U	$P_U = R_U/N_U$	R_L	$P_L = R_L/N_L$	$D= P_U - P_L$
1	12	0.86	4	0.36	0.50
2	13	0.93	8	0.73	0.20
3	10	0.71	2	0.18	0.53
4	10	0.71	5	0.45	0.26
5	12	0.86	0	0.00	0.86
6	14	1.00	8	0.73	0.27
7	11	0.79	7	0.64	0.15
8	11	0.79	5	0.45	0.34
9	13	0.93	6	0.54	0.39
10	13	0.93	5	0.45	0.48
11	8	0.57	1	0.09	0.48
12	5	0.36	0	0.00	0.36
13	11	0.79	3	0.27	0.52
14	10	0.71	6	0.54	0.17
15	11	0.79	4	0.36	0.43
16	5	0.36	2	0.18	0.18
17	12	0.86	5	0.45	0.41
18	11	0.79	3	0.27	0.52
19	10	0.71	6	0.54	0.17
20	14	1.00	6	0.54	0.46
21	12	0.86	7	0.64	0.24
22	6	0.43	0	0.00	0.43
23	12	0.86	5	0.45	0.41
24	14	1.00	8	0.73	0.27
25	5	0.36	3	0.27	0.09
26	5	0.36	4	0.36	0.00
27	10	0.71	2	0.18	0.53
28	5	0.36	1	0.09	0.27
29	13	0.93	7	0.64	0.29
30	14	1.00	10	1.00	0.00

R_U = number of students in the upper group answering the item correctly,
R_L = number of students in the lower group answering the item correctly
N_U = number of students in upper group. In this case 14
N_L = number of students in lower group. In this case 11
P_U = proportion of students in upper group answering the item correctly
P_L = proportion of students in lower group answering the item correctly
D = the item discrimination index

Table 4 presents item discrimination index of each item. It reveals that, with the exception of items 26 and 30, all the items discriminated between the upper group and lower group. However, following the 0.30 suggestion by Etsey (2012), it could be seen that 11 items which are 2, 4, 6, 7, 14, 16, 19, 21, 24, 25 and 28 did not discriminate well enough between the upper group and the lower group. This is because they all have an index less than the suggested 0.30. All the other 19 items discriminate well and are considered good items. Item 5 with an index of 0.86 discriminated very well and meets the standard for standardized tests.

Research Question 4: What are the distracter indices of Mathematics Multiple Choice Test (MMCT)?

Table 5: Distracter analysis

			A			B			C			D			NR	
	Key	U	M	L	U	M	L	U	M	L	U	M	L	U	M	L
1	B	1	0	1	12	7	4	1	3	5	0	0	1	0	0	0
3	D	4	5	5	0	0	3	1	2	0	9	3	2	0	0	1
4	A	10	5	5	2	2	1	2	1	2	0	2	3	0	0	0
5	C	0	0	0	2	3	1	12	6	1	0	1	9	0	0	0
6	D	0	0	0	0	3	0	1	0	1	13	7	9	0	0	1
12	C	5	0	0	3	3	3	5	7	6	0	0	2	1	0	0
16	D	9	6	6	0	1	2	0	0	1	5	3	2	0	0	0
25	A	5	3	3	5	6	6	1	0	2	1	1	0	2	0	0
29	C/D	0	0	2	1	1	1	10	6	5	3	3	3	0	0	0
30	A/B/C/D	0	0	0	2	2	0	12	8	10	0	0	1	0	0	0

Table 5 presents the distractor analysis of 10 selected items. Results from the table shows how students in the upper, medium, and lower ability chose the distracter options which were A, C, and D. The table indicated that all the distracters attracted at least 1 examinee from question 1. Also, the distracter attracted more or equal of the low achievers than high achievers. This indicates that all the distracters have functioned well. Distracter A has function well by attracting more of the lower group than the upper group. Though all the distracted attracted some examinee, distracter C attracted more of the upper group than the lower group hence must be modified. For Question 4, all the distracters attracted at least one person and attracted the same number of examinees. This means that all the distracters are functioning well. Distracter B attracted more in the upper group than in the lower group hence needs modification. Question 5 indicated that Distracter A did not attract any examinee. This means that it is not functioning well and therefore must be changed or modified. Distracter B attracted more in the upper group than in the lower group hence needs modification. Question 6 indicated that Distracter A did not attract any examinee. This means that it is not functioning well and therefore must be changed or modified.

Further, For Question 12, each option attracted at least 1 examinee. However, option A attracted the same number of high achievers as the key C with 1 high achiever not making any response. This suggests that either the item was ambiguous or it was mis-keyed hence the key needs re-examination or the item needs clarity. For Question 16, each option attracted at

least 1 person. However, option A attracted more of high achievers than the key, D. This suggests that either the item was ambiguous or it was mis-keyed hence the key needs re-examination or the item needs clarity. In Question 25, each option attracted at least 2 examinees. However, option B attracted the same number of high achievers as the key with 2 high achievers not making any response. This suggests that either the item was ambiguous or it was mis-keyed hence the key needs re-examination or the item needs clarity. For Question 29, each option attracted at least 2 examinees. Many students in each group selected option C than the other option that was also accepted. This means that many examinees especially the high achievers used the approach that resulted in option C. It can be concluded that the approach to option C should be considered not the other approach. This suggests that the item needs a minor clarity. Further, in Question 30, option A did not attract any examinee hence needs to be replaced. Many students in each group selected option C than the other options that were also accepted. This means that many examinees especially the high achievers saw option C as the best option though the negative sign was absent. It can be concluded that option C should be considered not the others. However, the best key needs a minor correction to avoid ambiguity.

From the 10 Item analysed for distracters, only one item, Item 1 was without issue of distracters. All the others had issues of either mis-keyed, other acceptable keys, and option not attracting any examinee.

Research Question 5: What is the reliability coefficient of Mathematics Multiple Choice Test (MMCT)?

Table 6: item variance using K-R_{20}

Item	R	p	q= 1-p	pq	Item	R	p	q= 1-p	pq
1	23	0.66	0.34	0.22	16	10	0.29	0.71	0.21
2	30	0.86	0.14	0.12	17	28	0.80	0.20	0.16
3	14	0.40	0.60	0.24	18	22	0.63	0.37	0.23
4	20	0.57	0.43	0.25	19	21	0.60	0.40	0.24
5	19	0.54	0.46	0.25	20	31	0.89	0.11	0.10
6	29	0.83	0.17	0.14	21	29	0.83	0.17	0.14
7	26	0.74	0.26	0.19	22	8	0.23	0.77	0.18
8	21	0.60	0.40	0.24	23	27	0.77	0.23	0.18
9	27	0.77	0.23	0.18	24	31	0.89	0.11	0.10
10	29	0.83	0.17	0.14	25	11	0.31	0.69	0.21
11	11	0.31	0.69	0.21	26	13	0.37	0.63	0.23
12	5	0.14	0.86	0.12	27	17	0.49	0.51	0.25
13	21	0.60	0.40	0.24	28	13	0.37	0.63	0.23
14	25	0.71	0.29	0.21	29	30	0.86	0.14	0.12
15	23	0.66	0.34	0.22	30	35	1.00	0.00	0.00
					Total		18.55	11.45	5.55

R = Total number of students who correctly answered the item

T = Proportion Correct (P = R/T)

Pq = Item variance

$$\frac{n}{n-1}\left(1-\frac{\sum pq}{s2}\right)$$

$$\frac{30}{30-1}\left(1-\frac{5.55}{4.326^2}\right)$$

$$1.034(1-0.297)$$

K-R_{20} : rxx = 0.73

Table 6 presents that analysis of item variance using K-R_{20}. The K-R_{20} formula for the reliability was used because the items were not of equal difficulty levels. It is also effective for multiple choice items. K-R_{20} is used to measure the internal consistency. The reliability coefficient from the K-R_{20} shows that the degree of consistency or dependency on the scores of the test is 0.73 or 73%. This means that the error level in the scores is 0.23 or 23%. The results of the reliability indicate that the test scores are reliable with a reliability degree of 0.73.

Content validity of Mathematics Multiple Choice Test (MMCT)

This study found that most of the items constructed by the teachers measured up to application level of the cognitive domain. The profile dimension of mathematics education in Ghana suggested 70% for application of knowledge (Ministry of Education, 2012 & Gyamfi, 2022). However, only 56.67% of the items were on application of level. This means that teachers need to improve of their skills of test construction. Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, Alavi-Majd and Nikanfar (2015) found in their content validity study that the instrument enjoys an appropriate level of content validity S-CVI with the average approach, which was equal to 0.93. The study of Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, Alavi-Majd and Nikanfar (2015) used the Lawshe (1927) method to estimate content validity ratio whiles this used the table of specification. However, both reported a good level of content validity ratio.

Difficulty index of Mathematics Multiple Choice Test

The study found that only 14 out of the 30 items observed had acceptable of difficulty. Each item of a test requires to have acceptable level of difficulty, 0.30-0.70 (Nitko, 2012). However, only few items in this study had the acceptable level of difficulty. Yeboah, Gyamfi, Wintson and Prempeh (2022) in a similar found that Six (6) out of the 40 items were found to be difficult with P-indices less than 0.3 and twelve (12) of the items were easy with P-indices greater than 0.7. The study of Shete, Kausar, Lakhkar and Khan (2015) showed that out of total 40 items, difficulty indices of 10 MCQ items were easy ($P \geq 70\%$) while about 12 MCQ were difficult ($P \leq 30\%$), and the remaining 18 of the items were within acceptable range (P = 30-70%). The studies of Yeboah, Gyamfi, Wintson and Prempeh (2022) and Shete, Kausar, Lakhkar and Khan (2015) indicates that items by classroom teachers have a deficiency with item difficulty.

Discrimination index of Mathematics Multiple Choice Test

The validation of the teacher made MCQ revealed that quite a number of the items constructed by the teachers did meet the acceptable discrimination indices. Asamoah-Gyimah and Amedahe (2012) stated that a good item should discriminate well the upper group and lower group and that the distracters of the option should function as intended. This study revealed that quite a number of the items constructed by the teachers discriminated well, which is good for the tests. Yeboah, Gyamfi, Wintson and Prempeh (2022) found that eighteen (18) out of the 40 items discriminated well. That is their D-indices were 0.4 or greater. Eleven (11) items had indices between 0.2 and 0.39 indicating that items discriminated satisfactorily and six (6) items had low discriminating indices ($D < 0.2$). Three (3) items, had negative indices and two (2) items also had 0.0 indices meaning these items did not discriminate between high and low achievers. Even though this study and that of Yeboah, Gyamfi, Wintson and Prempeh (2022) indicated that quite a number of the teacher made items though in different subject areas had good discrimination indices, the percentage is still low.

Distracter indices of Mathematics Multiple Choice Test

The study found that only one out of ten randomly sampled items had good distracters. There is much to be desired. A similar study by Yeboah, Gyamfi, Wintson and Prempeh (2022) using the similar method indicated that 4 out of the 10 items selected at random had issues with distracters. Date, Borkar, Badwaik, Siddiqui, Shende and Dashputra (2019) also found out that distractor efficiency related to presence of zero or 1 non-functional distrator (NFD) is 80%. This indicates that selecting a plausible distracter in MCQ is a bite challenging. Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (2010) stated defective options affects the validity of the results of the test for decision making.

Reliability of Mathematics Multiple Choice Test

Reliability of the scores of the test using KR 20 and 21 was found to be 0.73, indicating a dependable result. The result of the study confirms to the acceptable level of reliable test of 0.70 (Nitko, 2012). The reliability estimate obtained from the Cronbach alpha formula is 0.78 or 78%. This clearly means that the error margin of examinees scores for the two halves of test is 0.22 or 22% (Yeboah, Gyamfi, Wintson & Prempeh, 2022). It can therefore be seen that classroom teachers are able to construct items with acceptable level of reliability. Different methods used in estimating the reliability coefficients in this study and that of Yeboah, Gyamfi, Wintson and Prempeh (2022) attest to this.

5.0 Conclusion and Recommendations

It has been seen that much attention should be taken when constructing mathematics items because any mistake can change the key of the item. For example, item 30 was given out as bonus because the omission of the negative sign in option C rendered the item with no key hence has to be given as bonus to all students. When proper directions are not given or the stem of the item does not provide adequate information on the approach to use, different answers could be obtained though, all the procedures are correct especially in computational subjects like mathematics. For example, options C and D of Item 29 have to be accepted as keys because the item did not specify how the rectangle should be labelled. Different labelling resulted in the acceptance of the two answers as correct. The mere fact an item has a low P-value does not mean the item is not effective or needs revision. The low P-value may be as a result of the item discriminating well. Items 11 and 28 are examples in this situation. It is therefore recommended that mathematics teachers should be given in-service training in test construction. This will help them construct items with less defect. Also, the assessment course taken by teachers in the training institution should intensify the test construction.

References

Annan-Brew, R (2020). Differential item functioning of West African senior secondary certificate examination in core subjects in southern Ghana. UCC, Ghana: PhD thesis

Asamoah-Gyimah, K. & Amedehe, F. (2003). *Introduction to measurement and evaluation*. Cape Coast: Unpublished

Asamoah-Gyimah, K. & Amedehe, F. (2015). *Introduction to measurement and evaluation* (Revised). Cape Coast: CODE

Asamoah-Gyimah, K., & Anane, E. (2018). *Assessment in schools*. University of Cape Coast: Unpublished Mimeograph.

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies 2*(9), 27-33.

Brennan, R. L. (Ed). (2006). Educational measurement (4th ed). USA: American Council on Education, Praeger Series on Education.

Crocker, L &Algina, J. (2008). Introduction to Classical and Modern Test Theory. USA: Lengage learning.

Date, A. P., Borkar, A.S., Badwaik, R. T., Siddiqui, R. A., Shende, T. R., & Dashputra, A. V. (2019). Item analysis as tool to validate multiple choice question bank in pharmacology. *International of Journal Basic and Clinical Pharmacology, 8* (9),1999-2003.

Etsey, Y. K. A., (2012). Assessment in education. Cape Coast: Unpublished

Gyamfi, A. (2022). Controlling examination malpractice in Senior High Schools in Ghana through performance-based assessment. *Journal of Advances in Education and Philosophy, 6*(3), 203-211

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidenced Based Nursing, 18* (3), 66-67.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111–115.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. Journal of Research in Personality, 43, 489-493.

Liaquat, H, Asif, J. M., Siraji, J., & Maroof, K. (2012). Development and standardization of intelligence test for children. International Journal of Learning & Development, 2(5), 190-202

Ministry of Education (2012). *Core Mathematics syllabus for senior secondary schools*. Accra: CRDD

Nitko, A. J. (2001). *Educational Tests and Measurements for students* (3rd ed.). USA: Prentice-Hall, Inc

Nitko, A. J. (2004). *Educational measurement* (4rd Ed.). USA: American Council on Education & Praeger.

Nitko, A. J. (2012). *Educational Tests and Measurements* (7th ed.). USA: Prentice-Hall, Inc.

Osterlind, S. J (2006). *Modern measurement: Theory, principles and application of mental appraisal*. Upper Saddle River, N J: Pearson Merill.

Pande, S. S., Santosh, R., Pande, S. R., Parate, V. R., Nikam, A. P. & Agrekar, S. H. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. *South-East Asian Journal of Medical Education, 7* (1), 45 - 5

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment in special and remedial education*. Boston: Houghton Mifflin.

Shete, A. N., Kausar, A., Lakhkar, K., & Khan, S. T. (2015). Item analysis: An evaluation of multiple-choice questions in physiology examination. *Journal of Contemporary Medical Education, 3*(3), 106-109.

Tamakloe, E. K., Atta, E. T., & Amedehe, F, K., (1996). *Principles and methods of teaching*. Black Mask, Cantoments, Accra.

Yeboah, A., Gyamfi, A., Wintson, D. K., & Prempeh, A. D. (2022). Item Analysis, Reliability Estimates, Percentile Ranks and Stanine Scores Among University Students in Ghana. *Advances in Social Sciences Research Journal, 9*(8), 42-60.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: Development of an instrument for measuring Patient-Centered Communication. *Journal of caring sciences, 4*(2), 165–178.