UTILIZATION OF MULTIPLE MATRICES SAMPLING TECHNIQUE IN THE DEVELOPMENT OF CHRISTIAN RELIGIOUS STUDIES MULTIPLE CHOICE ITEMS

Kunuba, Agatha Chukwufumnanya (Ph. D.)
agathakunuba663@gmail.com 07065429664
Department of Science and Computer Education
Faculty of Education, Godfrey Okoye University.

Abstract

Multiple matrices sampling (MMS) technique is an approach that can be used in educational research, test measurement and evaluation. The effectiveness of MMS can be tested by sampling items and examinees. Multiple-choice items were developed in the subject area of Christian Religious Studies, using multiple matrices sampling technique. Five research questions guided the study, and two null hypotheses were formulated and tested at 0.5 alpha level. An instrumentation research design was used. A sample of 200 Upper Basic III students drawn from one of the public secondary schools in Enugu Education Zone formed the subjects for the study for the final administration of the instrument. A total of 160 Christian Religious Studies multiple-choice items (CRSMCI) were developed, validated and matrixed into four sets through genuine multiple matrix sampling technique for the initial administration (trial testing) to ascertain the reliability of the instrument. Kuder-Richardson 20 (KR-20) was used to establish the reliability of the four sets of CRSMCI at 0.66, 0.67, 0.75, and 0.71 respectively. Item analysis of the four sets were carried out to establish the item difficulty, discrimination and distracter indices. One hundred and four (104) items survived the item analysis. The 104 items were matrixed into four sets of 276 items, using partial multiple matrix sampling technique. The final administration of the test was done. Content validation was done using the Chi-square test of the Goodness-of-fit at an alpha level of 0.05. The data generated from the study were analyzed using Pearson Product Moment Correlation and Analysis of Variance (ANOVA). The result of the study showed that; the items that formed the CRS multiple-choice item packages did not deviate from the specifications of the test blueprint, there was an appreciable level of internal consistency of 0.93, 0.88, 0.84, 0.97 for the four sets of CRSMCI, and the validity coefficient of 0.98. There was a positive correlation among the sets of CRSMCI at 0.98, 0.94, 0.97, 0.97, and the mean performance of examinees who were tested with the sets of CRSMCI were relatively equal. The implications of the findings showed that, the use of multiple matrices sampling technique improved broad content coverage of curriculum and reduced the testing time given to examinees. The study recommended among others that, test blueprint should be used by CRS teachers and test developers to write CRS multiple-choice items before adopting it as an instrument. Also, multiple matrices sampling technique should be used by schools, CRS teachers and test developers to enhance content coverage and reduce testing time yet maintaining standard.

Keywords: Multiple Matrices Sampling, Genuine Matrix Sampling, Partial Matrix Sampling, Multiple-Choice Items, Christian Religious Studies, Test Development, Item Analysis

Introduction

In the field of test, measurement and evaluation, multiple matrices sampling (MMS) is an approach that can be applied in the development of items. Anigbo (2011) noted that it can be used to achieve a broad curriculum coverage and at the same time limiting the testing time for each individual student. It is a technique that gives a practical approach to test developers who may want to develop a pool of items that may cover the curriculum of the whole class or a term, and the number of items will be so many that it will not be practicable to administer all the items to the same candidates at a single seating. After items have been developed, these items are divided into subsets with MMS technique and then administered to different sets of the examinees. Therefore, MMS is a technique that combines respondent sampling and item sampling. Shoemaker in Anigbo (2011) says multiple matrices sampling is a technique used to sample both both the population of respondents and the totality of items, that is, respondent sampling and item sampling.

Multiple matrices sampling can be used for large scale assessment of students using tests. One of the most used assessment tools in education is to conduct tests. Beyond being considered as an instrument, tests can also be standard procedures used to systematically measure a sample of behaviour by posing a set of questions. Tests are designed to measure the quality, ability, skill or knowledge of a sample against a given standard, which usually could be deemed as acceptable or not. In educational practice, tests are methods used to determine the students' ability to complete certain tasks or demonstrate mastery of a skill or knowledge of content. Tests can take the form of multiple choices or a weekly spelling (Adom, Menser & Dake 2020). Multiple-choice items are mostly used for test because they can be more reliable in terms of scoring and analysis of examinees' scores. Though, multiple-choice items are not easy to develop.

Multiple-choice items can take different forms. It can be structured (objective type) or unstructured (essay type). According to Anigbo (2014), multiple choice items are tests where problems or questions are presented in the stem and the examinee is required to select the best answer or options. Using multiple-choice question format for test such as achievement test, has dominated large-scale testing. This is because, a test taker can answer many multiple-choice questions in a limited amount of testing time. The large number of questions makes it possible to test a broad range of content and provides a good sample of the test takers knowledge, and it helps in reducing the effect of the result of chance on the test taker's scores (Livingston, 2009). For this reason, multiple-choice test is often preferable to other forms of tests.

Multiple-choice items can be scored objectively with the use of modern technology without rater's bias. It also allows for the inclusion of a broad range of topics on a single exam thereby effectively testing the breath of student's knowledge (Weimer, 2018). However, when a broad range of topics are included in a single exam, presenting as much content as possible, it could increase the length of the instrument and affect the testing time. It will be difficult to cover the contents of subjects that have large content areas, such as Christian Religious Studies. Studies on MMS technique have been carried out by researchers who used it (1) to estimate group mean, (2) to establish a broad curriculum content and at the same time reducing the testing time per individual student, (3) to estimate aggregate performance from the partial knowledge of examinees ability, (4) for a comparative estimation of average students' score between sampling items, examinee sampling and multiple matrix sampling, and (5) to investigate the effects of reducing the length of the questionnaire. From the available literature, none of studies in MMS was carried out in CRS. None has equally developed multiple-choice items in CRS that has a broad content, using MMS technique to administer the items to examinees.

Christian Religious Studies (CRS) is that aspect of learning that deals with inculcating in the learners, certain elements of education, intellectual theory and practices of Christ as contained in the Holy Bible (Kesmen & Mellemut, 2022). Religious study is an academic field devoted to research into religious beliefs, behaviours, and institutions. It describes, compares, interprets, and explains religion, emphasizing systematic, historically based, and cross-cultural perspectives. The origin of religious studies could be dated back to the nineteenth century when scholarly and historical analysis of the Bible had flourished, and Hindu and Buddhist texts were first being translated into European languages. Early influential scholars included Friedrich Max Müller in England and Cornelius P. Tiele in the Netherlands. Today CRS is practiced by scholars worldwide (Wikipedia, 2022). CRS therefore, is a subject offered in school which is meant to prepare learners for useful living through inculcation of Christian attitudes and values, and to prepare learners for higher education and it has its unique value and importance to the whole education process (Ntama, Owulu and Monity 2016).

CRS is a subject that has a large content area which ranges from the Old Testament to the New Testament Bible and writings. Apart from the fact that test is used to measure students' learning, it can also be used to find out the extent to which the subject contents have been covered. There could be a way of presenting as much cognitive content as possible without overstraining the examinee and at the same time reducing the testing time. The essence of developing CRS multiple-choice items is (1) to ensure a technical standard in the process of developing the items, (2) to ensure appropriate psychometric properties including item analysis, (3) to prepare test manual so that teachers can use standardized test, (4) to develop valid and reliable instrument that can be used for diagnosis and placement for decision making, and finally, most studies carried out in this area were done in other disciplines and not in CRS.

Item analysis is an essential step in the development of multiple-choice items. Item analysis, according to Abonyi (2011) has to do with the assessment of the adequacy of each of the items that make up the test or instrument. During the process of item analysis, each of the items is assessed in terms of its difficulty, discrimination, and distractor index. Item analysis aims at improving tests by revising or discarding ineffective items, and in classroom achievement tests it provides diagnostic information on what examinees know and what they do not know (Enyi, 2009). In working out the measures of validity and reliability of test items, individual items are analyzed to see if they have the required characteristics. Item analysis helps in analyzing the responses of examinees. Item analysis is assessed in terms of its difficulty, discrimination and distracter index. Item analysis is done after scoring answer sheets to assess the quality and performance of the test items. Through item analysis reliability and validity of a test is ensured. Researchers have developed models for testing, assessing, and analyzing data, using psychometric techniques to determine the validity and reliability of tests or test items.

The essence of test is to make valid judgement about cognitive ability of students in content areas they are exposed to. Test items can be subjective or objective. Objective tests are tests that require the learner to select the correct option from two or more options or complete statements or compute simple calculations. Objective tests are categorized into multiple choice, true/false and matching items. one of the major advantages of multiple-choice items is its ability to make wide coverage of contents, provide prompt feedback and measure higher cognitive abilities. The aim of developing CRS multiple-choice items test is to help educators or educational institutions make valid inference of students' knowledge in each subject content area. This requires that the test developers must look at the domains of educational goals to see how test items can cover the contents of subject areas. In test measurement, content coverage is one of the major considerations in achievement test. In a situation whereby an item developer is faced with the pressure of curriculum coverage and testing time there is already an unsatisfied need. If the item developer writes fewer items, a lot of content areas might be omitted in the curriculum. By so doing, the test items may fail to test the students' depth of understanding and skill in applying knowledge. On the other hand, if the item developer considers too much testing time there will be inadequate content coverage and lower quality assessment.

Content coverage and testing time are two important aspects that are considered in the development of multiple-choice test items. An attempt to cover the subject content can lead to increase in the length of the instrument. Content area shows where the students are expected to exhibit their performance and it also shows the curriculum standard of such test. It can affect data quality as a result of lengthy items or questions and long testing time can overburden the examinees.

In a situation where the item developer is faced with pressure of curriculum coverage, testing time tends to create a need for more technical approach like multiple matrices sampling technique. A cursory look at BECE past questions compiled by Examination Development Centre, Ministry of Education, Enugu State, the researcher noticed that most of the CRS BECE multiple-choice are wanting in this regard thereby, underpinning on the above advantages of multiple matrices sampling.

Developed items in CRS is expected to go through the validation process to measure what it purports to measure. Most often CRS instruments are developed in schools, even sometimes at state levels without ascertaining that the items are valid and relevant to the subject contents and objectives. The questions that bother on this study are these: How can a broad content coverage and testing time work out simultaneously and effectively without reducing each, and yet achieve the goal of testing? How can test developers like that of BECE develop qualitative CRS items with regards to curriculum coverage and testing time? How can the contents and objectives of the developed CRS items be validly measured?

MMS is a sampling technique that can be used to address the concerns relating to measures for increasing the cognitive content and at the same time reducing the testing time. MMS as a technique that can be used to sandwich the developed CRS multiple-choice items into samples known as subsets, and they are administered to the sample of examinees drawn from the population of studies. This can be done using Genuine Matrix Sampling and Partial Matrix Sampling. Anigbo (2011) defines genuine matrix sampling as the situation where the students or groups of students are sampled so that only some of the students or groups take any test at all. In this case, both items and students or groups are sampled. While in partial matrix sampling, a subset of the package was selected to be common to all the students or groups, and the remaining subsets are then matrix sampled. It means that the common subset serves as an anchor and helps to improve comparability of students or group results, the matrix-sampled items increasing content coverage per testing time.

The purpose of the study was to utilize multiple matrices sampling technique in the development of Christian Religious Studies multiple-choice items (CRSMCI).

The study specifically sought to find out

1. the items that formed the CRS multiple-choice item packages.
2. the internal consistency of the CRS multiple-choice items, using Kuder-Richardson formula 20 (K-R 20) after the administration of the test
3. the reliability estimates for the subsets in the CRS multiple-choice items
4. the test-retest reliability coefficient of the subsets in the CRS multiple-choice items
5. the psychometric qualities of item difficulty, item discrimination and distracter indices of the items in the CRS multiple-choice items

The study was guided by the following research questions:

1. What are the items that formed the CRS multiple-choice item packages?
2. What is the internal consistency of the CRS multiple-choice items?
3. What are the reliability estimates for the sets in the CRS multiple-choice item?
4. What is the test-retest reliability coefficient for the sets in the CRS multiple-choice items?
5. What are the psychometric qualities (item difficulty, item discrimination and distracter indices) of the items of Christian Religious Studies?

The following null hypothesis was tested at 0.05 level of significance:

$H0_{1:}$ There is no significant difference in the items that formed the CRS multiple-choice item packages from the specifications of the test blueprint

$H0_{2:}$ There is no significant difference between the mean performance scores of each group of examinees.

Method

This study was carried out in Enugu Education Zone in Enugu State. The choice of this study was delimited to Enugu Education Zone which comprised three local government areas and 31 secondary schools managed by Enugu State Ministry of Education.

The population for the study was made up of 442 students of the Upper Basic III (UBS3) students from one public secondary school in Enugu Education Zone of the State. The school is managed by Post Primary School Management Board (PPSMB) of Enugu State (Source: PPSMB 2023). The reason for adopting UBS3 is because the Christian Religious Multiple-Choice Items (CRSMCI) was developed with the UBS3 Syllabus. Also, one secondary school was selected because the study was concerned about testing CRSMCI by using multiple matrices sampling technique. The study adopted instrumentation research deign because it developed, validated and produced tests on CRSMCI.

A sample of 200 students drawn from 442 UBS3 were used for the final administration of the instrument. A Multistage sampling procedure involving three stages was adopted in the study. First, the sampling of Enugu East to represent Enugu Education Zone. This was done through simple random sampling technique. Names of the Local Government Areas (Enugu North, Enugu East and Isi-Uzo) were written in small pieces of paper and one was picked to represent the zone. Next was the sampling of a secondary school to represent the government secondary schools. This was done through convenient sampling. The reason for sampling one secondary school is because the study was interested in testing an instrument. The final stage was the sampling of students done during the school hour, and those not sampled were moved to another class. There were eight classes of the Upper Basic III and disproportionate stratified random sampling was used. The eight classes represented eight subsets. Then the researcher randomly chose 25 students from each of the classes, given a total of 200. Hence, every subset stood the same chance of being selected for the exercise.

The instrument used for this study was CRS Multiple-Choice Items (CRSMCI) constructed by the researcher. It was a 160-item drawn from the contents of the Upper Basic CRS syllabus. It contained multiple-choice items with four options lettered A, B, C, and D for each item. The constructed-response questions have one correct answer and three distracters. The CRSMCI was used to assess the extent of the curriculum content coverage and testing time. The 160 CRSMCI developed, were matrixed into four sets of 40 items using Genuine Matrix Sampling. Initial administration of each of the four sets was administered to each individual examinee. Based on the scores obtained from the trial testing, the difficulty index, discrimination index was established. These were ascertained by categorizing the research subjects into upper and lower criterion groups. Percentage passes for each item among these two groups were used in the exercise.

For the item difficulty, the items were normally distributed. 12 items were of high difficulty, more items were of average difficulty and few were easy. In the discrimination index, a very good number of items discriminated positively while 56 (out of 160) of the items discriminated negatively. Item selection was done, 104 items were selected while 56 items were discarded. The analysis of distracters was limited only to those of the selected items. The distracters that did not attract any patronage, or have positive indices were reviewed. Only distracters with negative indices were retained. After the trial testing and item analysis, scores obtained from the trial testing were used to determine the internal consistency of the instruments. The level of internal consistency was measured using Kuder-Richardson 20 (K-R 20), with reliability estimates of 0.66, 0.67, 0.75 and 0.71 for the four sets of trial test packages.

The 104 CRSMCI items that survived were matrixed into 176 items using Partial Matrix Sampling and were adopted for the final administration.

Results

Research Question 1:

What are the items that formed the CRS multiple-choice item packages?

The 176 items were composed into four sets of 44 items each with the use of partial multiple sampling technique, 24 items are common to each of the sets. The CRSMCI package was developed with the syllabus of the Upper Basic III

Table 1: *Summary of partial multiple matrix sampling sets of items*

Students or groups	Common	Set 1	Set 2	Set 3	Set 4	Total
1	24	20				44
2	24		20			44
3	24			20		44
4	24				20	44

Research Question 2:

What is the internal consistency of the CRS multiple-choice items?

Responses of the 200 research participants for the final administration of the instrument were used to test the internal consistency of the CRSMCI. The responses of the research subjects to the items of the test on a single administration were subjected to KR-20 approach. Summary of the result is presented on Table 2, 3, 4, and 5.

Table 2: *Summary of test of internal consistency of the CRSMCI Set 1 using KR-20 approach*

Number of items	SDt	SDt^2	∑Pq	K-R 20
44	5.21	27.14	2.44	0.93

Table 2 shows the extent of internal consistency of Set 1 of CRSMCI. As shown on Table 2 the internal consistency index is 0.93. This indicates that the test has an appreciable level of internal consistency.

Table 3: *Summary of test of internal consistency of the CRSMCI Set 2 using KR-20 approach*

Number of items	SDt	SDt^2	∑Pq	K-R 20
44	4.05	16.42	2.35	0.88

Table 3 shows the extent of internal consistency of Set 2 of CRSMCI. As shown on Table 3 the internal consistency index is 0.88. This indicates that the test has an appreciable level of internal consistency.

Table 4: *Summary of test of internal consistency of the CRSMCI Set 3 using KR-20 approach*

Number of items	SDt	SDt^2	∑Pq	K-R 20
44	3.99	15.9	2.85	0.84

Table 4 shows the extent of internal consistency of Set 3 of CRSMCI. As shown on Table 4 the internal consistency index is 0.84. This indicates that the test has an appreciable level of internal consistency.

Table 5: *Summary of test of internal consistency of the CRSMCI Set 4 using KR-20 approach*

Number of items	SDt	SDt^2	∑Pq	K-R 20
44	3.64	13.26	2.96	0.79

Table 5 shows the extent of internal consistency of Set 4 of CRSMCI. As shown on Table 5 the internal consistency index is 0.79. This indicates that the test has an appreciable level of internal consistency.

Research Question 3

What are the reliability estimates for the sets in the CRS multiple-choice item?

The reliability of the sets of CRSMCI were determined using test retest approach. Data obtained from the first and second administration of the test to the research subjects were subjected to Pearson Product Moment Correlation approach to determine the reliability of the instrument. Summary of result is presented on Table 6:

Table 6: *Summary of Reliability Coefficient for the sets of CRSMCI*

No of items	Sets	*r*
50	1	0.98
	2	0.94
	3	0.97
	4	0.97

The sets of CRSMCI have correlation index of 0.98, 0.94, 0.97 and 0.97 respectively. This indicates that the test is reliable.

Research Question 4

What is the test-retest reliability coefficient for the sets in CRS multiple-choice items?

Test-retest reliability coefficient was carried out to find out the reliability correlation between the first and second administration of the sets of CRSMCI, using Pearson Coefficient Correlation. Summary of the test is presented below in table 7:

Table 7: *summary of test-retest reliability coefficient of the sets of CRSMCI*

Sets	***r***	**no**	***p*-value**
Set 1			
1st admin	.947**	50	.000**
2nd admin	.947**	50	.000**
Set 2			
1st admin	.927**	50	.000**
2nd admin	.927**	50	.000**
Set 3			
1st admin	.967**	50	.000**
2nd admin	.967**	50	.000**
Set 4			
1st admin	.955**	50	.000**
2nd admin	.955**	50	.000**

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

The Pearson Correlation analysis was carried out to find out if there exist a positive correlation between first and second administration of the sets of CRSMCI. Firstly, the normality test was carried out using Shapiro-Wilk and Kolmogorov Smimov test which both indicates that the data was normally distributed. The result of the Pearson Correlation test of set 1 shows there is a significant correlation as follows: r = .947, n = 50, p = .000. The test of set 2 shows a positive correlation of r = .927, n = 50, p = .000. The test of set 3 shows a positive correlation of r = .967, n = 50, p = .000. The test of set 4 shows a positive correlation of r = .955, n = 50, with a corresponding p-value of .000. There is a strong positive correlation between the sets in the CRS multiple-choice items.

Research Question 5

What are the psychometric qualities of the items of Christian Religious Studies?

Table 8: *Item Characteristic Table for Set 1*

Items	P	d	Items	P	d	Items	P	d
1	0.59	0.44	16	0.75	0.34	31	0.80	0.25
2	0.73	0.44	17	0.69	0.41	32	0.76	0.32
3	0.67	0.46	18	0.70	0.25	33	0.75	0.30
4	0.77	0.30	19	0.70	0.34	34	0.69	0.32
5	0.72	0.37	20	0.52	0.20	35	0.59	0.39
6	0.79	0.32	21	0.52	0.34	36	0.77	0.30
7	0.61	0.48	22	0.77	0.25	37	0.76	0.23
8	0.76	0.23	23	0.77	0.30	38	0.74	0.23
9	0.72	0.23	24	0.72	0.37	39	0.68	0.39
10	0.56	0.62	25	0.55	0.60	40	0.74	0.27
11	0.79	0.32	26	0.79	0.32	41	0.61	0.20
12	0.54	0.62	27	0.54	0.62	42	0.54	0.25
13	0.72	0.41	28	0.72	0.41	43	0.52	0.34
14	0.69	0.37	29	0.69	0.37	44	0.43	0.39
15	0.74	0.27	30	0.73	0.20			

Key: p…Item difficulty d…discrimination index *…negative index

Table 9: *Item Characteristic Table for Set 2*

Items	P	d	Items	P	d	Items	P	d
1	0.59	0.44	16	0.75	0.34	31	0.73	0.30
2	0.73	0.44	17	0.69	0.41	32	0.61	0.34
3	0.67	0.46	18	0.70	0.25	33	0.59	0.48
4	0.77	0.30	19	0.70	0.34	34	0.65	0.37
5	0.72	0.37	20	0.52	0.20	35	0.66	0.48
6	0.79	0.32	21	0.52	0.34	36	0.74	0.27
7	0.61	0.48	22	0.77	0.25	37	0.66	0.39
8	0.76	0.23	23	0.77	0.30	38	0.46	0.37
9	0.72	0.23	24	0.72	0.37	39	0.61	0.39
10	0.56	0.62	25	0.62	0.55	40	0.48	0.37
11	0.79	0.32	26	0.67	0.41	41	0.63	0.34
12	0.54	0.62	27	0.70	0.34	42	0.61	0.48
13	0.72	0.41	28	0.75	0.20	43	0.59	0.30
14	0.69	0.37	29	0.60	0.37	44	0.55	0.27
15	0.74	0.27	30	0.62	0.51			

Key: p…Item difficulty d…discrimination index *…negative index

Table 10: *Item Characteristic Table for Set 3*

Items	P	d	Items	P	d	Items	P	d
1	0.59	0.44	16	0.75	0.34	31	0.66	0.30
2	0.73	0.44	17	0.69	0.41	32	0.61	0.34
3	0.67	0.46	18	0.70	0.25	33	0.75	0.25
4	0.77	0.30	19	0.70	0.34	34	0.63	0.30
5	0.72	0.37	20	0.52	0.20	35	0.52	0.34
6	0.79	0.32	21	0.52	0.34	36	0.60	0.27
7	0.61	0.48	22	0.77	0.25	37	0.53	0.37
8	0.76	0.23	23	0.77	0.30	38	0.56	0.25
9	0.72	0.23	24	0.72	0.37	39	0.59	0.30
10	0.56	0.62	25	0.61	0.25	40	0.66	0.20
11	0.79	0.32	26	0.68	0.25	41	0.62	0.41
12	0.54	0.62	27	0.53	0.27	42	0.70	0.25
13	0.72	0.41	28	0.62	0.27	43	0.70	0.20
14	0.69	0.37	29	0.60	0.37	44	0.58	0.41
15	0.74	0.27	30	0.61	0.25			

Key: p…Item difficulty d…discrimination index *…negative index

Table 11: *Item Characteristic Table for Set 4*

Items	P	d	Items	P	d	Items	P	d
1	0.59	0.44	16	0.75	0.34	31	0.70	0.20
2	0.73	0.44	17	0.69	0.41	32	0.48	0.37
3	0.67	0.46	18	0.70	0.25	33	0.70	0.34
4	0.77	0.30	19	0.70	0.34	34	0.67	0.23
5	0.72	0.37	20	0.52	0.20	35	0.65	0.32
6	0.79	0.32	21	0.52	0.34	36	0.63	0.25
7	0.61	0.48	22	0.77	0.25	37	0.59	0.34
8	0.76	0.23	23	0.77	0.30	38	0.58	0.51
9	0.72	0.23	24	0.72	0.37	39	0.62	0.27
10	0.56	0.62	25	0.75	0.25	40	0.62	0.37
11	0.79	0.32	26	0.55	0.27	41	0.58	0.46
12	0.54	0.62	27	0.73	0.34	42	0.73	0.30
13	0.72	0.41	28	0.61	0.20	43	0.58	0.32
14	0.69	0.37	29	0.54	0.44	44	0.70	0.25
15	0.74	0.27	30	0.65	0.23			

Key: p…Item difficulty d…discrimination index *…negative index

A close look at Tables 8 to 11 reveal that difficulty indices for individual items ranged from 0.50 to 0.79. This means that the items were normally distributed. Ideally, difficulty indices according to Rayos (2018) should range from 0.20 to 0.80. Items whose difficulty indices were less than 0.20 and above 0.80 are not retained. For the discrimination indices, the items in the tables ranged from 0.23 to 0.62. This indicates that the items discriminated positively. This is in line with Abonyi (2011) who stated that item discrimination index lies between + 1.00 and – 1.00. When the index is positive, it means that the items discriminates in the right direction and when the index is negative it means the items discriminates in the wrong and so need to be revised. The distracter indices of the options ranged from – 0.10 to – 0.64. Options with positive indices were reviewed.

Hypotheses

H0₁: There is no significant difference in the items that formed the CRS multiple-choice item packages from the specifications of the test blueprint.

Table 12: *Test of the goodness-of-fit of CRSMCI packages based on the test blueprint.*

Content	Remembering (30%)	Understanding (35%)	Applying (20%)	Analyzing (5%)	Evaluating (5%)	Creating (5%)	X^2cal	Alpha	X^2crit	Decision
God and His creation (15%)	7 (7)	9 (8)	5 (3)	1 (1)	1 (1)	1 (1)				
Gods call (10%)	5(5)	6(6)	2 (2)	1(1)	1 (0)	1 (0)				
Keeping God in our relationship (10%)	5 (5)	6 (6)	2 (2)	1 (0)	1 (0)	1 (1)				
The early life of Jesus (15%)	7 (7)	9 (9)	5 (4)	1 (1)	1 (1)	1 (1)				
The ministry of Jesus (15%)	7 (7)	9 (9)	5(4)	1(1)	1(1)	1 (0)	1.13	0.05	15.50	Null hypothesis not rejected
The Sacrifice of Jesus (10%)	5 (5)	6 (6)	2(2)	1 (0)	1 (1)	1 (1)				
The beginning of the church (10%)	5 (5)	6 (5)	2 (1)	1 (1)	1 (1)	1 (1)				
The Ministry of the Apostles (10%)	5 (5)	6 (6)	2 (2)	1 (0)	1 (1)	1 (1)				
The Christian church today (5%)	2 (2)	3 (2)	1 (1)	1 (1)	1 (1)	1 (1)				

Table 12 shows that the calculated value (1.13) was lesser than the critical value (15.50). It was tested for significance at an alpha level of 0.05 using test of the goodness-of-fit at an alpha level 0.05, to test its significance. The null hypothesis was not rejected. Therefore, the items that formed the CRS multiple-choice item packages did not deviate from the specifications of the test blueprint.

H0₂: There is no significant difference between the mean performance of each group of examinees.

The One-Way ANOVA (analysis of variance) was used to compare the mean performance of each group of examinees in order to determine whether there is statistical evidence that the associated group means were significantly different. This was tested for significance at an alpha level of 0.05

Table 13: *Summary of ANOVA of the mean performance of each group of examinees*

Sets		Sum of squares	Df	Mean Square	F	Sig
Set 1	Set 2	435.55	19	22.924	1.618	.116
	Set 3	177.180	19	9.325	.394	.982
	Set 4	265.880	19	13.994	1.139	.366
Set 2	Set 1	480.618	13	36.971	1.570	.141
	Set 3	112.261	13	8.635	.401	.960
	Set 4	311.399	13	23.954	2.669	.010
Set 3	Set 1	443.580	16	27.724	1.034	.450
	Set 2	260.976	16	16.311	.897	.578
	Set 4	170.513	16	10.657	.758	.717
Set 4	Set 1	569.888	13	43.838	2.080	.042
	Set 2	202.167	13	15.551	.850	.608
	Set 3	199.749	13	15.365	.804	.652

Table 13 represents the test of the significant difference between the mean performance of examinees in Sets 1 with (2,3,4); Sets 2 with (1,3,4); Sets 3 with (1,2,4) and Set 4 with (1,2,3) respectively.

For Set 1 (2,3,4) the F-calculated of 1.618 has an associated probability score of 0.116, F-calculated of 0.394 to 0.982, and F-calculated of 1.139 has an associated probability score of 0.366. The conclusion is that 0.116>0.05 0.982>0.05, and 0.366>0.05. This means that there is no significant difference between the mean performance of examinees for sets 1, 2, 3 and 4.

For Set 2 (1,3,4), the F-calculated of 1.570 has an associated probability score of 0.450, F-calculated of 0.401 has an associated probability score of 0.960 and F-calculated of 2.669 has an associated probability score of 0.01. While 0.401>0.05, 0.960>0.05, 0.01<0.05. This means that while there is no significant difference between the mean performance of examinees in sets 2,1 and 3, there is significant difference between sets 2 and 4.

For Set 3 (1,2,4), the F-calculated of 1.034 has an associated probability score of 0.450, F-calculated of 0.897 has an associated probability score of 0.578 and F-calculated of 0.758 has 0.717. This implies that 0.450>0.05, 0.578>0.05, and 0.717>0.05. Therefore, the null hypothesis of no significant difference between the mean performance of examinees is not rejected for sets 3, 1, 2, and 4.

For Set 4 (1,2,3), the F-calculated of 2.080 has an associated probability score of 0.042, the F-calculated of 0.850 has an associated probability score of 0.608, and F-calculated of 0.804 has an associated probability score of 0.652. this means that 0.042<0.05, 0.608>0.05, and 0.652>0.05. This means that while there is no significant difference between the mean performance of examinees in sets 4, 2 and 3, there is significant difference between sets 4 and 1.

Decision: the null hypothesis of no significant difference between the mean performance of examinees in Sets 1 with 2,3,4; Sets 2 with 1,3; Sets 3 with 1,2,4 and Set 4 with 2,3 is not rejected, whereas the null hypothesis is rejected between sets 2 and 4, sets 4 and 1.

Discussion:

The result of the data in table 1 revealed that partial multiple sampling technique can be used to multiply the number of CRSMCI into subsets, thereby reducing both the number of items administered to each examinee and testing time required. This is in line with Anigbo (2011) who demonstrated the possibility of using multiple matrices sampling technique to achieve a broad curriculum coverage and minimizing the testing time for each student.

In the same vein, the result in table 12 showed that the items that formed the CRS multiple-choice item packages did not deviate from the specifications of the test blueprint. This implies that the test blueprint is a necessary requirement for the development of CRSMCI. This supports the observation of Ebuoh (2004) that test blueprint is needed in determining the accuracy of the content validity of test instrument.

The results of Tables 2, 3, 4, and 5 revealed that the CRSMCI has an appreciable level of internal consistency of 0.93, 0.88, 0.84 and 0.79 respectively. This implies that there is consistency of the examinees' responses to the items of the CRSMCI test for the four sets. The test measured what the CRSMCI intended to measure. English and Keeley (2015) observed that tests with higher internal consistency accurately measures the intended construct of the test developer.

The results in Table 6, showed that the sets of CRSMCI have correlation index of 0.98, 0.94, 0.97 and 0.97 respectively. The results of the correlation indicate that the test is reliable. Similarly, the result of Table 7 on test-retest reliability coefficient of the first and second administration of the sets in CRSMCI showed a strong positive correlation between the sets in the first and second administration of the instrument. This suggests that the results of the sets in CRSMCI are consistent. This agrees with Akeem (2015) who stated that correlation coefficient (r) values are considered good if $r \geq 0.70$.

The results in Tables 8 to 11 showed the item analysis of each of the items that make up the CRSMCI, assessed in terms of its difficulty, discrimination, and distracter index. Difficulty indices for individual items ranged from 0.50 to 0.79, meaning that the items were normally distributed. For the discrimination indices, the items in the tables ranged from 0.23 to 0.62, indicating that the items discriminated positively. The distracter indices of the options ranged from – 0.10 to – 0.64 options. Item analysis of CRSMCI aided the selection of valid and reliable items for testing. In accordance to this claim, Kumar, Jaipurkar, Shekhar, Sikri and Srinivas (2021) noted that item analysis is an essential tool that is used to provide input on validity and reliability of multiple-choice items. It can be used to identify items which can be revised or discarded.

The result in hypothesis two showed the test of no significant difference between the mean performance of each group of examinees for the four sets of CRSMCI; Sets 1,2,3,4; Sets 2,1,3,4; Sets 3,1,2,4 and Set 4,1,2,3 respectively. The result revealed that the examinees performed equally in Sets 1 with 2,3,4; Sets 2 with 1,3; Sets 3 with 1,2,4 and Set 4 with 2,3. But the examinees did not perform equally in sets 2 and 4, 4 and 1.

Conclusions

Based on the findings of this study, the following conclusions were drawn:

1. The items that formed the CRS multiple-choice item packages did not deviate from the specifications of the test blueprint.
2. The CRSMCI has an appreciable level of internal consistency.
3. In terms of its difficulty, the CRSMCI were normally distributed. Also, the items discriminated in the right direction, and its distracter index was negative indicating that the items were good.
4. There is significant positive correlation among the sets of CRSMCI.
5. The mean performance of examinees who were tested with the sets of CRSMCI were relatively equal.

Recommendations

Resulting from the findings of this study, the following recommendations have been made by the researcher:

1. Test blueprint or table of specification should be used by CRS teachers and test developers to write CRS multiple-choice items before adopting it as an instrument.
2. Multiple matrices sampling technique should be used by schools, CRS teachers and test developers to enhance content coverage and reduce testing time yet maintaining standard.
3. Item analysis should be used by test developers to ascertain the psychometric qualities of items.
4. Government through the ministry of education should adopt on-going training of CRS teachers through seminars, workshops, and in-service training on how to develop multiple-choice items.
5. CRS teachers should be trained and encouraged by the school management to always make use of the test blueprint for their teacher-made tests.

References

Abonyi, O. S. (2011). Instrumentation in Behavioural Research: A Practical Approach. Fulladu Publishing Company.

Adom, D., Menser, J.A., & Dake, D.A. (2020). Test, measurement and Evaluation: Understanding and use of the Concepts in Education. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(1) 109-119. https//doi.org/10.11591/ijere.v911.20457

Akeem, B.O. (2015), Principles and Methods of Validity and Reliability Testing of Questionnaires Used in Social and Health Science Researches. *Nigerian Postgraduate Medical Journal* 22(4), 195-201. http://doi.org/10.4103/117-1936.173959

Anigbo, L.C. (2011) Demonstration of the Multiple Matrices Sampling Technique in Establishing the Psychometric Characteristics of Large Samples. *Journal of Education and Practice 2*(3) 19-25.

Ebuoh, C.N. (2004). *Educational Measurement and Evaluation for Effective Teaching and Learning*. Sky Printing Press.

English, T. & Keeley, J. (2015). Internal Consistency Approach to Test Construction. http://doi.org/10.1002/9781118625392.wbecp156

Enyi, G. S. (2009). Introduction to Measurement and Evaluation in Universities, Polytechnics and Colleges of Education: A professional Approach. Linco Press.
https://core.ac.uk/download/pdf/234633146.pdf
https://www.enugustate.gov.ng
https://www.iprjb.org/journals/index.php/AJEP/article/download/1448/1561

Kesmen, Y.S., & Mellemut, P.A. (2022). The Effects of Christian Religious Studies Literacy on Academic Performance and Achievement of Secondary Students on Their Academic Achievement: A Case Study of Secondary School 1 Students in Quaan-Pan Local Government in Plateau State of Nigeria. *African Journal of Education and Practice, 8*(1) 1-19.

Kumar, D.; Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item Analysis of Multiple-Choice Questions: A Quality Assurance Test for an Assessment Tool. http://doi/10.1016/j.mjafi.2020.11.007

Livingston, S. (2009). Constructed-Response Test Questions: Why we use them; How we score them. *R&D Connections*, *11*, 1-8. https://eric.ed.gov/?id=ED507802

Ntama G.U, Owulu E.E. and Monity F.M (2016) Students' Variables and Academic Performance in Christian Religious Studies in Calabar Municipality, Nigeria in *International Journal of Social Science and Humanities Invention*, http://doi.org/ 10.18535/ijsshi/v3i11.02

PPSMB Enugu School Directory. https://ppsmbenugu.com.ng/school-directory

Rayos, F. (2018). Improving the Test items. https://www.slideshare.net

Religious Studies (2022). In Wikipedia. https://en.wikipedia.org/wiki/Religious_studies

Weimer, M. (2018, February 21). Multiple-Choice Tests: Revisiting the Pros and Cons. https://www.facultyfocus.com/articles/educationalassessment/multiple-choice-tests-pros-cons/