ESTIMATION OF PSYCHOMETRIC QUALITIES OF WEST AFRICAN EXAMINATION COUNCIL CHRISTIAN RELIGIOUS STUDIES MULTIPLE CHOICE QUESTIONS

Clifford O. Ugorji
Department of Educational Foundation,
Nnandi Azikiwe University, Akwa
ugocliff3@yahoo.com

Abstract
This study estimated the psychometric qualities of WAEC Christian religious studies multiple choice questions. Descriptive Research Design was adopted for the study. A sample size of 1500 students participated in the study. This was obtained through a multistage sampling procedure. Christian Religious Studies Multiple Choice Questions (CRSMCQ) was the instrument used for data collection. The internal consistency reliability coefficient of the instrument was determined to be 0.75 using the Kuder-Richardson formula (K-K_{20}). The data collected was analyzed using the Normal Ogive Harmonic Analysis Robust Method (NOHARM) and Root Mean Square Reduction (RMSR). Findings revealed that the 2020 multiple choice items of the West African Senior School Christian Religious Studies Examination have 10 items that moderately measure the students' abilities, 26 items were too easy, and 14 items were difficult for the students. The findings also revealed that all the items with the exemption of items 10, 11, 18, 20, 23, 27, 28, 33, and 41 have the ability to differentiate between the higher-ability students and the lower-ability students, and they have good difficulty abilities.
Keywords: Asymptotic index, estimation, IRT, psychometric assessment, students' ability estimate, multiple-choice items

Introduction

A test is a series of questions or tasks systematically constructed to elicit certain information concerning the examinee with respect to their abilities and traits. In contrast, the process of administering such tests to examinees is referred to as examination. Testing is the most popular and widely used process of measuring and assessing students' cognitive learning outcomes (Okwilagwe & Ogunrinde, 2017). The outcome of such examinations provides essential parameters with which stakeholders judge the product of the educational system. The outcome of examinations is used for the purpose of decision-making, award-giving, curriculum modification, examinees placement purposes, modification instruction, and assessment of the quality of learning as well as the quality of learners. As such, the quality of these tests must be given keen attention, especially in test development. However, when an examination is not properly conducted, the expected feedback may be difficult or even impossible to achieve. Hence, the result of such evaluation leads to wrong decisions and judgments, which affect the teacher, the learner, the entire education industry, and society (Ojonemi et al., 2013). When test items are also poorly developed, the outcome also would provide information that is not reliable. Hence, a test development process needs to be properly scrutinized.

Proper scrutiny of Christian Religious Studies examination items is critical because apart from determining students' academic achievement, it also assesses students' level of acquisition of moral character needed to become effective citizens. Ocheoha (2015) explained that Christian Religious Studies campaigns for better citizenship through her curriculum contents. Despite the immense benefits of Christian Religious Studies in schools, poor academic performance is recorded yearly, and more students are running away from the subject. Adebiyi et al. (2016) reported that the attitude of the learners, teacher-related factors, and students' understanding of the items in the examination are some of the factors contributing to students' poor academic performance in Christian Religious Studies in Schools.

Scrutinizing test development processes cut across all test forms, whether a dichotomously scored test or a polytomously scored test. Dichotomously scored tests are those tests with two possible item scores (Thompsom, 2016). This implies that dichotomously score tests are tests whose item is made up of right answer and wrong answer(s). A very good example of dichotomously score test is the multiple-choice tests, which have 2, 3, 4 or 5 options, but are assigned only two possible scores (correct/incorrect). Polytomously scored tests, according to Thompsom (2016), are tests whose items alternate varying scores depending on the examinees or the correctness of the examinees' response that have more than two scoring steps. The most common example is the essay tests, which is subjective in nature. Polytomously scored tests are tests that items have several problem-solving steps. Irrespective of the type of test, scrutiny should be performed during the test development process.

Item difficulty in Item response theory is the parameter that determines the manner of which the item behaves along the ability scale. It measures the proportion of examinees who answered the item correctly (Maydeu-Olivares, 2015). According to Mozaffer and Farhan (2012), it is determined at the point of median probability, that is, the ability of 50% of respondents to endorse the correct answer. On an item characteristic curve, items that are difficult to endorse are shifted to the right of the scale, indicating the higher ability of the respondents who endorse it correctly, while those that are easier are more shifted to the left of the ability scale. It is also used to describe how difficult it is to achieve a 0.5 probability of correct response for a specific item given the respondent's latent variable level (Miller et al., 2009). The more difficult it is for a student to have a 50% chance of correctly answering an item, the higher the ability level needed to achieve this goal (Yang & Kao, 2014).

Item discrimination is represented with (a). Item discrimination is used to identify items that can differentiate between examinees with lower and those with higher ability of the proficiency or attribute being measured by the test items (Ayala, 2022). When the discrimination value is high, the item discriminates between examinees with different levels of the attribute measured. Thus, items with high discrimination are better. The purpose of using the test is to differentiate between examinees who know the attribute being tested and those who do not or on a scale, between those who have positive attitudes and those who have negative attitudes. High values of (a) indicate greater discrimination. Apart from difficulty and discrimination parameters, guessing parameters also influence the reliability of a test.

The guessing parameter/index is a c-parameter because examinees with very low ability would be expected to correct the item only by guessing. Ayala (2022) stated that in well-developed standardized tests, the c-parameter tends to be somewhat lower than chance because good distracters draw low-ability examinees away from the correct answer. However, if an item contains poor distracters that even a low-ability examinee can eliminate as possibilities, the c-parameter may be higher than chance. When the item parameters are known, the parameters could be estimated using a marginal maximum likelihood, maximum likelihood criterion, and Bayesian estimation procedures (De Ayala, 2009). These are the most widely used estimation procedures. Marginal maximum likelihood procedures apply to the one-, two-, and three-parameter models (Bock & Aitkin, 1981). The ability parameters are integrated, and the item parameters are estimated. With the item parameter estimates determined, the ability parameters are estimated. The sampling distributions of maximum likelihood estimates are known in large samples. In addition, when item parameters are known, estimation of ability can be carried out by obtaining the mode of the likelihood function or, in the case of Bayesian procedures, either the mean or the mode of the posterior density function of ability ($\boldsymbol{\theta}$).

These parameters are used to determine the examinee's ability. To locate the examinee's ability on the ability scale, estimating the examinee's ability becomes pertinent in IRT. In estimating the examinee's ability, the examinee can be evaluated in terms of how much underlying ability he or she possesses, and comparisons among examinees can be made to assign grades and award scholarships. In students' ability estimation, the test is used as a measure to determine unknown latent traits. The test comprises a number of items, and each item measures some aspects of the trait. When the test is taken, an examinee will respond to each of the items in the test, and the responses will be dichotomously scored. The answer given by the examinee will be a score of 1 or 0 for each item in the test. Thus, the list of 1 and 0 for the number of items is called the examinee's item response vector. This item response vector and the known item parameters estimate the examinee's unknown ability parameter. Every measurement has possibilities of errors; the limit of errors that could be allowed in a measurement is called the standard error of measurement.

Standard error of measurement is directly related to the reliability of the test. Standard error indicates the precision with which a parameter is estimated. The smaller the standard errors are, the more precise the parameter may be estimated. The standard error of estimate (SEE) helps one understand that the scores obtained in one educational measurement are only estimates and may differ considerably from individuals' presumed true scores. Standard error of measurement is a statistical estimate of the amount of random error in the assessment of results or scores (Chatterji, 2013). Meredith et al. (2007) maintained that standard measurement error allows one to determine the probable range within which the individual's true score falls. Generally, a measure of the error in a statistic's parameter estimate can be obtained. This measure is referred to as a standard error. The standard error is an index of the variability of an estimator with respect to the parameter it is estimating (De Ayala, 2009). The larger the value of a standard error, the greater the error and the less certain we are about the parameter's value. Similarly, the standard error estimate in IRT is the uncertainty about a person's location. The SEE specifies the accuracy of a person's ability with respect to the person's location parameter. When there is a small degree of uncertainty about a person's location, its SEE is comparatively smaller than when there is a greater degree of uncertainty. Interpreting students' scores and making judgements on their performance with respect to an examination such as Christian Religion Studies cannot be effective if the psychometric properties of such examination are not established. Establishing the psychometric properties entails determining the level of difficulty of the items of the examination and the ability of the items to identify the low-ability and high-ability students.

Theoretically, this study is anchored on Item Response Theory. The IRT framework was pioneered by Fredrick Lord in 1953. The basic tenet of IRT is centered on the unobserved traits of the examinee. It is a group of mathematical models that attempt to

explain the relationship between latent traits (unobservable characteristics or attributes) and their manifestations (observed outcomes, responses, or performance). They establish a link between the properties of items on an instrument, individuals responding to these items, and the underlying trait being measured. For this study, the theory provides the opportunity to determine the psychometric properties of the 2020 West African Senior School Christian Religious Studies Examination. Using a group of mathematical models is most efficient in establishing a link between the properties of items in an instrument, individuals responding to the items, and the underlying construct being measured.

Most students with lower ability may also respond correctly to an item to which they should have responded wrongly, given their ability levels, which now yields an asymptote parameter. Since Christian Religious Studies Multiple Choice Items are used to assess students' knowledge, morals, and level at which learning has taken place, and the outcome of the assessment will be used to make decisions and recommendations; when the quality of the examination is not ascertained, the outcome and decisions would be faulty. On this premise, this study tends to determine the psychometric properties of the CRS Multiple Choice Examination using the Item Response Theory and establish the item parameters since the item response theory framework is more efficient in doing it. This study determined the psychometric qualities of the 2020 multiple-choice test items of the West African Senior School Certificate Examination in Christian Religious Studies.

Research Questions

1. What is the item facility index of 2020 multiple choice items of West African Senior School Christian Religious Studies Examination?
2. What is the item discrimination index of the 2020 multiple choice items of the West African Senior School Christian Religious Studies Examination?

Method

This study adopted a descriptive survey research design. A descriptive research design aims to collect data and systematically describe it: the characteristics, features, or facts about a given population or phenomenon (Nworgu, 2015). The design is suitable for the study because the researcher is interested in collecting data and describing the item facility, item discrimination, asymptotic index, and students' ability estimate of the 2020 West African Senior School Christian Religious Studies Examination.

The study participants comprised 1,500 senior secondary school students (S 3) offering CRS, which is 5 percent of the entire population of the students. The decision for the sample size was based on the findings of De Ayala (2009), who opined that based on research, it appears that for Marginal Maximum Likelihood Estimation (MMLE), a sample

of 1,000 or 2,000 persons will lead to reasonably accurate item parameter estimates with the 3PL model under normal conditions. The researcher adopted a multi-stage sampling procedure. Twenty-four intact SS3 classes of the selected schools formed the sample used for the study.

The instrument for data collection was Christian Religious Studies Multiple Choice Questions (CRSMCQ), adopted from the West African Senior Schools Certificate Examination for 2020. The instrument was made up of 50 items. Each item has four response options: A, B, C, and D, in which only one among the options is the key for each item. All the responses of the students were coded in Microsoft Excel; each item that was correctly responded to was scored as one (1), while the items that were not correctly responded to were scored as zero (0). In order to determine the internal consistency of the instrument, thirty instruments were administered to SS 3 students in another school that are not part of the study, and the students shared similar characteristics in terms of the same curriculum content. The instrument was collected and scored. The instrument was further subjected to an internal consistency reliability estimate, and a reliability coefficient of 0.81 was obtained using Kudder Richardson 20 ($K\text{-}R_{20}$).

The data collected using the instrument was analyzed with respect to the research questions that guided the study. The multidimensional four-parameter logistic model of item response theory using R programming software was used to answer research questions 1, 2, 3, and 4. The acceptable values for students' abilities (θ) theoretically, range from -∞ to +∞; practically, values between -3 and +3 were considered acceptable. Item facility index (b) theoretically ranges from -∞ to +∞but is generally between -3 and +3. Item discrimination index (a) theoretically ranges from -∞ to +∞, but practically, it ranges from 0 to perhaps 2 or 3. This study used values of between +0 to +3, because of the range of tests under study. Items with negative locations were screened out. Theoretically, the asymptotic index (c) ranges from 0 to 1 but would more realistically range from 0 to 3 (De Mars, 2010 & De Ayala, 2009).

Results

Table 1 shows the result of the Item facility index of 2020 multiple choice items of West African Senior School Christian Religious Studies Examination. The result shows that only 10 items fall within the acceptable item facility index range of -3 to +3. Fourteen items fall above the acceptable range while 26 items fall below the acceptable range. This implies that the 2020 multiple choice items of the West African Senior School Christian Religious Studies Examination have 10 items that moderately measure the abilities of the students; 26 items were too easy for the students, while 14 items were difficult for the students.

Table 1. Item facility index of the Multiple-choice items

Item Facility	Above the	Within the	Below the

(-3 to +3)	acceptable range	acceptable range	acceptable range
Number of Items	14	10	26

Table 2 shows the result of the Item discrimination index of the 2020 multiple choice items of the West African Senior School Christian Religious Studies Examination. The 2020 multiple-choice items of the West African Senior School Christian Religious Studies Examination have three discrimination indices because the examination is a three-dimensional test. The discrimination indices for the first dimension show that 12 items show a very high discrimination index, 12 items show a moderate discrimination index, and 26 items have a low discrimination index. The discrimination indices for the second dimension show that 17 items show a very high discrimination index, 12 items show a moderate discrimination index, and 21 have a low discrimination index. The discrimination indices for the third dimension show that 22 items show a very high discrimination index, 14 items show a moderate discrimination index, and 14 items have a low discrimination index.

Table 2. Item discrimination index of the Multiple-choice items

Item Discrimination (-3 to +3)	Above the acceptable range	Within the acceptable range	Below the acceptable range
a1	12	12	26
a2	17	12	21
a3	22	14	14

Table 3 shows the result of the asymptotic index of the 2020 multiple-choice items of the West African Senior School Christian Religious Studies Examination. The results show that all the asymptotic index of 2020 multiple choice items of the West African Senior School Christian Religious Studies Examination fall within the acceptable range of the asymptotic index. This implies that only some of the items provide the opportunity for students with lower ability to correctly respond to the answer of each item.

Discussion

Item facility indices of the test items

The study revealed that only 10 items fall within the acceptable range of item difficulty, 14 items fall above the acceptable range, and 26 items fall below the acceptable range. This implies that all the items are partially suitable for the level of the students for which the test was designed. This implies that the items of the Test were relatively easy for the students for whom they were developed. The difficulty levels of the items are within the

ability level of the examinees. The implication is that the instrument is good at classifying students based on their cognitive levels since the feedback provided by the examination places students on the same continuum. The range in item difficulty was in line with the findings of De Mars (2010) that item difficulty should generally range from -3 to +3 so as not to be too easy or too hard for the intended test population. The finding agrees with Chong (2013) that the difficulty parameter, also known as the threshold parameter value, tells us how easy or difficult an item is. The finding also corresponds with the findings of Obinne (2008) that negative difficulty estimates indicate that the items are easy while positive estimates indicate that the items are difficult.

Item discrimination indices of the test items

The Multiple-Choice Items of the examination assess three underlying constructs; hence, it is a multidimensional test. As such, they have discrimination indices a1, a2, and *a3*. With regard to *a1, a2, and a3, 12, 12, and 14 items of the Multiple-Choice Items of 2020 West African Senior School Certificate Examination in Christian Religious Studies fall within the acceptable range of item discrimination,* which is -3 to +3. Whereas 12 17and 22 fall above the acceptable range, and 26, 21, and 14 items fall below the acceptable range of item discrimination. This implies that many items could discriminate between high-ability and lower-ability students. The Multiple-Choice Items of the 2020 West African Senior School Certificate Examination in Christian Religious Studies can differentiate between the students with more cognitive advantage from those who are average and those with low cognitive ability. Item discrimination as the quality of test items enables the item characteristic curve to display the location of each student. This provides reliable feedback to the teachers on the students so that they can give maximum attention to each other so that they can meet up with others in the same class. Regarding the range of the discriminating index, the findings agree with the findings of De Ayala (2009) that the reasonably good values of item discrimination range from approximately -3 to +3. The finding also aligns with the findings of Reeve and Fayers (2005), which state that the discriminating parameter indicates how well an item distinguishes between respondents below and above the item discriminating parameter. The finding agreed with Yang and Kao (2014) that the discrimination index contributes to increasing the validity and reliability of a test by revealing whether the items are working well. The findings also align with Maydeu-Olivares's (2015) finding that the ultimate purpose for designing a precise measure is to include items with a high discrimination index to map out individuals along the latent trait continuum.

Conclusion

This study estimated the psychometric properties of Multiple-Choice Test Items of the 2020 West African Senior School Certificate Examination in Christian Religious Studies.

The Multiple-Choice Test Items of the 2020 West African Senior School Certificate Examination in Christian Religious Studies could differentiate between high-ability students and low-ability students. The study revealed that most of the items were suitable for the level of students they were designed for, as well as concerning item discrimination. Making decisions about examinations with different or unknown underlying constructs is appropriate and reliable because the total score obtained in any examination is a representation of the aggregate of all the abilities measured by that examination.

References

Adebiyi, S.J, Adimula, I.A, Oladipo, O.A & Joshua, B.W. (2016). Assessment of IRI and IRI-

Plas models over the African equatorial and low-latitude region. *JGR: Space Physics, 121*(7), 7287-7300 https://doi.org/10.1002/2016JA022697

Akaranga, S. I., & Ongong, J. J. (2013). The phenomenon of Examination Malpractice: An Example of Nairobi and Kenyatta Universities. Journal of Education and Practice, 4(18): 87 – 96

Ayala, R.J. (2022). The theory and practice of item Response theory, London; The Guilford Press.

Babatunde, K. O. & Benson, A. A. (2020). Using test theory models to assess senior students' ability in constructed-Response Mathematics test. Retrieve from; https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=babatunde+and+Benson+%282020%29&btn=$dgs_qabs&u=%23p%3D3Ors4ppWvLUJ

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443 459. https://doi.org/10.1007/BF02293801

Chatterji, M. (2003). Designing and using tools for educational assessment. *Journals of psychoeducational assessment, 22(1), 169 – 174.* Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/073428290402200207

Chong, A.Y.L. (2013) Predicting m-Commerce Adoption Determinants: A Neural Network

Approach. Expert Systems with Applications, 40, 523-530. https://doi.org/10.1016/j.eswa.2012.07.068

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*; London: The Guilford press.

De Mars, C. (2010). *Item Response Theory: Understanding statistics measurement*. New York: Oxford University press

Emaikwu, S. O. (2012). Assessing the Impact of Examination Malpractices on the Measurement of Ability in Nigeria. International Journal of Social Sciences & Education, 2(4): 748 – 757

George, I. N., & Ukpong, D. E. (2013). Contemporary Social Problems in Nigeria and its Impact on National Development: Implication for Guidance and Counselling Services. Journal of Educational and Social Research, 3(2): 167 – 173

Harris, A. (2004) *Distributed Leadership: Leading or Misleading Educational Management and Administration*, *32(*1), 11-24 ISBN 0263-211X

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. *Handbook of item response theory modeling: Applications to typical performance assessment*, 111-127.

Meredith, D. G., Joyce, P. G., and Walter, R B., (2007). *Educational research: an introduction* (8th ed.). United State of America: Pearson Press.

Miller, M. D., Linn, R., & Gronlund, N. (2009). *Measurement and evaluation in teaching (10th ed.)*. Upper Saddle River, NJ: Merrill, Prentice Hall

Mozaffer, H. R. & Farhan, J. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142.

Nnam, M. U., & Inah, A. F. (2015). Empirical Investigation into the Causes, Forms and Consequences of Examination Malpractice in Nigerian Institutions of Higher Learning. International Journal of Novel Research in Humanity and Social Sciences, 2(1): 52 – 62

Nworgu, B.G. (2015). *Educational measurement and evaluation: theory and practice.* Nsukka*:* University Trust Publishers.

Obinne, A.D.E (2008). Comparison of the psychometric Properties of senior Certificate Biology examinations conducted by West African Examinations Council and National Examinations Council (Unpublished doctoral Thesis) University of Nigeria, Nsukka, Nigeria.

Ocheoha KC (2015). Economics of Fadama Resource Management Practices in Federal Capital Territory, Abuja. An unpublished masters dissertation, submitted to the Department of Agricultural Economics, University of Nigeria, Nsukka

Ojonemi, P. S., Enejoh, W., Enejoh, A., & Olatunmibi, O. (2013). Examination Malpractice: Challenges to Human Resource Development in Nigeria. International Journal of Capacity Building in Education and Management, 2(1): 91 – 101

Okwilagwe, E. A. & Ogunrinde, M. A. (2017). Assessment of unidimensionality and local independence of WAEC and NECO 2013 Geography achievement tests. African Journal of Theory and Practice of Educational Assessment, 5(3), 31 – 45.

Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In *In: Assessing Quality of Life in Clinical Trials: Methods and Practice 2nd edn (ed Fayers,P. M.;Hays,R. D.), Oxford University Press, Oxford* (pp. 55-73).

Thompsom, P. O. (2016). Comment on 3PL IRT Adjustment for guessing. *Applied Psychological Measurement. Retrieved from: https://journals.sagepub.com/doi/full/10.1177/0146621612459369*

Tommy, S. U. & Udo. T. O. (2019). Using multiple-choice tests to evaluate students' understanding of accounting. Accounting Education: *An International Journal, 17 (Supplement)*, 55-68.DOI:10.1080/09639280802009249

Yang, W. T. & Kao, R. H. (2014). The optimal achievement model and underachievement in Hong Kong, *An application of the Rasch Model, Psychology Science Quarterly*. 50(2), 147-172

Yoella, B.M, Joachim, M., & Oded, M.F. (2002). prospect theory analysis of guessing in multiple choice test. *Journal of behavioral decision marking,* 5(14), 313 – 327 https://doi.org/10.1002/bdm.417

Zahrakar, K. (2013). Study of effectiveness of rational, emotive, behavior therapy (REBT) with group method on decrease of stress among diabetic patients. *Psychology, Medicine*