

CHAPTER ONE

INTRODUCTION

Background to the Study

Test is an indispensable tool used for measuring learners' progress and achievement. It is a measuring tool used to appraise an individual's performance in a given task. It is used to evaluate and assess a student's proficiency in a specific subject or an area within the subject. It is used by different educational institutions to measure the proficiencies of students from different fields of study. It is used in a systematic procedure to compare the performance of an individual with a designated standard of performance. It is capable of determining an examinee's true ability so as to place learner on the ability scoring scale in order to prevent dropout. Test provides maximum information about an examinee's ability and maintaining measurement precision for the examinees, especially, those at the extreme of the ability continuum (Baker, 2001). Testing in schools is used to produce scores which are often used in making vital decisions such as selection, placement, promotion and certification.

The scores and grades obtained from tests are used by schools for these fundamental decisions. Higher institutions used test scores and grades for admission. External examining bodies use test grades for licensure and certification. At the senior secondary school level, test scores are used to appraise learners' abilities and skills in order to be certificated and get admission into higher institutions. Grades and test scores from standardized tests are used by higher institutions to determine an applicant that would be admitted into any of the programmes offered by the institution. The National Policy on Education (FRN 2013) stated that the goal of secondary education is to prepare an individual for useful living within the society and preparing them for higher education. Therefore, the scores obtained are expected to reflect the most accurate estimates

of abilities and skills of the learners. Doran, Moses and Eignor (2010) emphasized that examinees must be treated with fairness and equity on any ability test they took. A standardized test is administered and scored in an orderly manner to ensure legal defensibility (Thissen & Wainer, 2011).

Test can be dichotomously or polytomously scored. Dichotomous scoring is used for test like multiple choice items and alternative response items that requires an examinee to pick only one appropriate answer from a list of answers. Multiple choice items require less time for examinees to answer; easy to score and grade; provide greater coverage of material; allows for an easily assessment of an examinee's ability; and provide feedback about how distracters behaves and why correct answers are right. Multiple-choice items provide higher validity because of wider coverage of school syllabus. It paves way for uniform assessment of the examinees unlike essay items where optional items are allowed.

The purpose of any achievement test is to assess the true ability of examinees by answering each item based on the available information. This paves ways for the examinees to confirm if correct answer is chosen or not based on the specific knowledge possessed, otherwise, it will be answered incorrectly. Many a times, this assumption may not good enough especially when multiple-choice tests are involved with different responses. Liao, Ho, Yen, & Cheng (2012) affirmed that the responses of an examinee in multiple-choice test can be grouped into three types: (i) responses that give true ability of the examinee (whether correct or incorrect), (ii) right responses arising from lucky guesses, and (iii) false responses arising due to anxiety, heedlessness, or distraction. The last two divergent responses may result into inaccurate estimation of the actual ability of the examinee because they do not reflect the exact knowledge of the examinee. Thus,

guessing and heedlessness can result into biasness of item in estimation of computerized adaptive testing (CAT).

Liao, Ho, Yen, & Cheng (2012) submitted that CAT is designed to monitor the administration of items according to each examinee's ability. The selection of each item of the test to be administered is based on the estimation of examinee's ability. The divergent responses of the examinees may result into ability estimation errors, hence, inappropriate selection of items arises. Where guessing is likely to be a factor affecting an examinee's test responses, the three-parameter logistic (3PL) model provides an appropriate solution for correctly estimating the examinee's ability (Liao, Ho, Yen, & Cheng, 2012). The 3PL model adds a guessing parameter (the lower asymptote) to care the effect of guessing.

Heedlessness or error due to anxiety may result in to estimation biases than what guessing does on the test. Rulison and Loken, (2009) reported that an examinee's ability underestimated when examinee makes careless errors at the beginning of a test. Barton and Lord (1981) introduced a fourth parameter (the upper asymptote) in the four-parameter (4PL). The model cares for underestimation of the examinees performance due to errors as a result of heedlessness. Liao, Ho, Yen, & Cheng, (2012) reported that 4PL model helped the examinee to recover from mistakes without reducing measurement precision and efficiency. They added that the 4PL model is a more robust mechanism against divergent responses. Magis (2013) reported that 4PL model is useful model to detect fitted data especially there is careless or inattention patterns of the examinees. The justification for 4pl model was established by Ayanwale, Adeleke & Abayomi (2018) who affirmed that Four-parameter model or Upper asymptote is capable of caring for incorrect answer as a result carelessness.

Polytomously scored item allows examinees to express their views or opinions on a specific concept. This gives the examinees opportunity to earn a particular score on a given instruction. Polytomous IRT models allow for the scoring of items when multiple response categories are allowed. Instead of using a single item characteristic curve (ICC), multiple category characteristic curves (CCCs) would be used. An examinee with a given trait level will be most likely to respond in the category whose curve is highest for their trait level. Polytomous models use a single discrimination parameter for the item regardless of response option, such as the GRM, PCM, and GPCM.

Test scores are used to assess the abilities and skills of individuals competing for admissions or seeking professional certification. The evaluation of examinee's test scores may lead to a decision of excluding a candidate from certain academic program or to limit the ability of an examinee to practice the profession of his/her choice. The importance placed on the results of standardized testing make it essential that, the test scores should provide a fair and equitable evaluation of the skills and abilities, the test is purported to measure. The provision of a fair and equitable means of psychological or educational assessment is one of the major reasons for the existence of standardized testing. A test may become nonstandard when an unfair evaluation of an examinee's skills or abilities is done, which resulted into measurement error. The measurement error may occur due to physical testing environment, the administration of the test and the actual content of instruction and nature of psychometric characteristics of the test. Whether a test score is interpreted with regard to a norm, criterion or both, it must meet a specified or established standard. The norm may be established independently or by statistical analysis of a large number of participants.

There are two common frameworks used to analyze the score of standardized test, the Classical Test Theory (CTT) and Item Response Theory (IRT). CTT is about traditional testing and foundation through which IRT was built upon. CTT deals with the reliabilities of relative decisions, where an individual's score is compared with a reference group and used for ranking the individual as norm-referenced measurement. However, IRT measures the inherent ability that produces the test performance rather than measuring actual test performance. IRT deals with assessment of the items that form the tests. This makes the IRT sample free because items statistics are not depending on the test situation that generated them, so they can be used more varying irrespective of the respondents. Therefore, result obtained from IRT can be used for generalization.

Reckase (2009) suggested that in determine the number of dimensions underlying the tests, the responses of the examinees that answer the tests need to be subjected to parallel analysis. Metibemu (2017) viewed parallel analysis as a method based on the generation of random variables, for the determination of factors that underlies a test. The implementation of the parallel analysis procedure involves the generation of a large number of matrices (at least 500) of random data, with each matrix having the same number of variables and subjects as the real data set under assessment. So, each of the data randomly generated is factor analyzed and the eigenvalues mean result are obtained. This is used to compare the eigenvalues of the matrix of the real data in order to determine the factors underlying the real data (Metibemu, 2017)). A factor is retained if its eigenvalue is found greater than the mean of the eigenvalues from its randomly generated counterpart (Lesdema & Valero-Mora, 2007).

IRT Information function plays an important role in describing items, item selection, and comparing efficiency of different tests (Hambleton, Swaminathan& Rogers, 1991). Information function tells how well each ability level is being measured statistically. Item Information Function

(IIF) indicates the usefulness of an item in assessing ability. IIF indicates contribution an item can made to the assessment of ability. Item and test information is an advanced concept in IRT to replace reliability (Hambleton & Swaminathan, 1985). Information is determined through model parameters. The parameter model is used to determine the degree of information obtained. The item information function of a given ability level is the degree of the square root of the differentiation of item characteristics to its variance (Hambleton, 2000). A point with maximum slope provides more information value about the examinee's ability than when the slope is lesser. Items whose slopes are steeper provide more information than items whose slopes are less steep. When the slope is steep and the variance is low, the information function would be larger and vice versa (Baker, 2001).

IRT models regarded ability of the examinees and the difficulty of the items as independent parameters. Discriminating power of multiple choice items show the rate at which an item distinguishes between the candidates of varying ability Alonge (2003) as cited in Olutola (2015). Item with higher discriminating power contribute more to the precision of measurement than the item with lower discriminating power. An item provides less information when guessing is a factor in test performance. Items tend to make their best contribution to measurement precision around the discriminating value of the ability.

The item information function uses all items parameters in order to establish precision and measurement efficiency of the item on different examinees' ability. The quality of an item is determined by the item parameters which gives examinee's ability. So, test information function is also influenced by the quality and the number of the test items used. The amount of test information function is based on each item of the test used to measure the inherent ability of an examinee. Thus, test information function gives the amount of information on a particular ability

based on the test items using parameter estimation. An item measures ability precision than a single item. So, longer tests measure an examinee's ability with greater precision than the shorter tests. De Ayala (2009) submitted that item and ability parameter estimation are important in obtaining parameter value that produce curve that best fits the observed data.

So, the addition of the IIFs give the test information function (TIF). TIF tells how precisely the test measures the ability of an examinee on a given task. Baker (2017) affirmed that amount of information must be translated into a standard error of estimate when interpreting item or test information. DeMars (2010) and Baker (2017) added that the Standard Error of estimate (SE) is the inverse of the square root of the amount of item or test information. Mathematically, $SE = \frac{1}{\sqrt{I(\theta)}}$

Where $I(\theta)$ = item or test information at a given ability level. In this study the ability level of each of the examinees that took the test were used. According to rule of thumb, Standard error of Estimate < 0.5 for an item or a test to be considered informative. An informative item or test is one whose precision of measuring the underlying ability or proficiency is high. Standard error of estimation should be considered as the major index of dependability, as opposed to the reliability of a test.

Individuals and groups can be tested with different set of test items, at different time and their performance can, be directly compared (Anastasi and Urbina, 2002). The assumption, the robustness of sample and items influenced the stability of IRT parameters. The ability of an examinee can be compared irrespective of number of items with different sets of items administered or used (Osterlind & Everton, 2009). So, if there is a large sample of examinees, the estimated item parameters can be compared with another item even if they were taken by different set of examinees (Hambleton, 2000). De Ayala (2009) asserted that the comparison between two

people should be independent, then the two people and their observed reactions should be independent of the persons taken the test.

De Ayala (2009) explained relative efficiency of the test as the ratio of the test information for two test forms across the entire score scale. Efficiency of the two forms of test indicate the variability of the degree of information provided by the forms on the scoring scale. Relative efficiency concerns with the test form that behave well in providing greater amount of information at different point on the ability scoring scale. Relative efficiency varying as a function of the ability (De Ayala 2009). De Ayala (2009) represents the relative efficiency of one score (X) to another score (Y) as the level of their information. The scores X and Y may come from different instrument with certain dimension of idea in two different ways. The rate of the two curves is called relative efficiency, which varies as a function of the ability level. Lam (2012) submitted that TIF is used to build a test to meet desired statistical specification.

Examination is a frequently used assessment tool which provides indices of students' achievement. The role of test is very paramount in evaluating students' achievement in the school environment. Nigeria for instance, the examination bodies conducting Senior Secondary School Certificate to determine students' admission in to higher education. They were; West African Examinations Council (WAEC), National Examinations Council (NECO) and National Business and Technical Examination Board (NABTEB) mainly for certification and admission.

However, in recent time there is an inconsistency in performance of the students at West African Senior Secondary Certificate Examination (WASSCE) relative to their ability and using the examinee results score of different year for the same purpose at different time is of concerned. For instance, in May/June 2013, 1,025,703 candidate representing 66.94% obtaining credits and qualified for university admission while in 2014, 698,669 candidates represent 51.22% obtained

credits and qualified for university admission as well. But with the use of grade point system for admission, does it mean that candidates who sat for May/June 2013 have higher ability than the 2014 candidates and others? In 2013, candidates had more than half of her candidates with higher percentage in grade A1- B3 compare with 2014 candidates. 320,983 candidates representing 20.95% of 2013 had A1-B3 compare with 124,010 candidates representing 9.09% of 2014 obtained the same grade like candidates in 2013. This is shown in the table below

Table 1: Analysis of Students' Performance in WASSCE Economics Between 2007-2016

Year	Total Students Sat	A1	B2	B3	C4	C5	C6	Total Credit
2007	1,183,154	1,942	6,401	80,673	59,149	78,851	234,887	461,903
%		0.16	0.54	6.81	4.99	6.66	19.85	39.03
2008	1204515	1541	7331	91231	86566	121792	268884	577,345
%		0.12	0.58	7.18	6.81	9.59	21.16	45.44
2009	1,270,557	1541	7331	91231	85566	121792	269884	577345
%		0.12	0.58	7.81	6.81	15.63	14.49	45.44
2010	1,228,401	2,830	12,818	137,545	116,150	146,970	274,636	690,949
%		0.23	1.04	11.20	9.46	11.96	22.36	56.25
2011	1,413,886	3,897	13,866	175,431	109,880	352,967	185,217	841,258
%		0.28	0.98	12.41	7.77	24.96	13.1	59.50
2012	1,540,902	5125	16096	194141	116427	187855	344629	864273
%		0.33	1.04	12.60	7.56	12.19	22.37	56.09
2013	1,532,194	8,424	33422	279137	183652	201306	319562	1025703
%		0.55	2.18	18.22	11.99	13.15	20.86	66.94
2014	1,363,994	2867	9122	112021	97072	136565	341022	698669
%		0.21	0.67	8.21	7.12	10.01	25.00	51.22
2015	1176381	1352	5011	70760	63720	95036	275901	511,779
%		0.11	0.42	6.01	5.41	8.07	23.45	43.47
2016	1071540	3012	12703	122364	91871	117001	274798	621,755
%		0.28	1.18	11.41	8.57	10.91	25.64	57.99

Source: Office of the Branch Comptroller, WAEC, Ilorin. (December, 2017).

From the analysis of the results in table 1, the performance of Economics candidates who sat for May/June 2007-2016 examination is inconsistent. Alaka and Obadara (2013) reported that there is poor students' academic performance in all geo-political zones of Nigeria. They added that although, increase in enrollment may be attributed to expansion of secondary schools and increased access to schools due to universal basic education (UBE) and education for all

programmes in Nigeria. Adu and Ayeni (2004) reported that poor results from low performance of the students may arise due to teacher's interaction, student's failure to ask questions and use of lecturing method when given instruction.

The trend of students' performance in (WASSCE) Economics is inconsistency relative to the examinees ability, but a sudden rise in 2013 compared with other years is of concern and called for investigation if the test results or grades at different times can still serve the same purpose. Since item and test information functions can be used to recount and select items as well as construct good test items, So, the researcher aimed at analyzing the items and test information function of (WASSCE) Economics of 2013 and 2014 paper 1 (objectives test) in order to detect quality and effectiveness of item as well as the extent to which result of the test can be used interchangeably at different times considering students' ability.

Statement of the Problem

The performance of students in public examinations varies from time to time at different years. This variation may be as a result of quality of test items, composition of different items formats in the test and ability distribution of the examinees. Researchers, educationists and general public are interested in relative information and efficiency of items used for Senior School Certificate Examinations (SSCE) in Economics because of the emphasis on the use of grade points system for Universities admission.

Thissen (2000) compared the dichotomously scored items and Bock's nominal model and reported that dichotomously scored item worked better. Alagoz (2005) reported that constructed response items provides more information than the multiple choices items. and Lam (2012) worked on the impact of item parameters on IRT information functions in mixed format tests using English

Language in Canada. All these studies were carried out using essay and objective test items but, this study employed only objective test items.

Alagoz (2005) also investigated information function on two parameters logistic model, Moghadamzadah, Salehi and Khodaie (2011) compared the information functions of the item and test in one, two and three parametric model in Mathematics and Physics in Iran. Zhang (2012) investigated the impact of variability of item parameter estimators on test information function. Zieba (2013) investigated the comparison of information function of one and two parameters model in Poland. Ali, Chang, & Anderson (2015) investigated tests with polytomous items using the two parameters. Ibrahim (2017) worked on the item parameters using three parameter models. All these were carried out on one, two and three parameter logistic models but this study was carried out using four parameter model.

Thissen (2000) studies was in in America, Moghadamzadah, Salehi and Khodaie (2011) studies was in Mathematics and Physics in Iran. Lam (2012) studied was in English Language in Canada. Most of the above studies were carried out on Physics, Mathematics and English language. So, to the best of this researcher's knowledge, none of these studies was carried out in WASSCE Economics, especially in North-central, Nigeria using items and test information functions. Therefore, the current study is aimed at addressing part of these gaps.

Purpose of the Study

The main purpose of this study is to analyze the items and test information functions of West African Senior School Certificate Examination in Economics in North-central, Nigeria. Specifically, the study intends to find out the;

1. dimensionality of 2013 WASSCE objective Economics items;
2. dimensionality of 2014 WASSCE objective Economics items;

3. item information function of 2013 WASSCE objective Economics items;
4. item information function of 2014 WASSCE objective Economics items;
5. Test Information Function (TIF) of 2013 WASSCE objective Economics item;
6. TIF of 2014 WASSCE objective Economics items;
7. relative efficiency of 2013 and 2014 WASSCE objective Economics items.

Research Questions

In order to guide this study, the following research questions were raised.

1. What is the dimensionality of 2013 WASSCE objective Economics items?
2. What is the dimensionality of 2014 WASSCE objective Economics items?
3. What is the item information function of 2013 WASSCE objective Economics items?
4. What is the item information function of 2014 WASSCE objective Economics items?
5. What is the Test Information Function (TIF) of 2013 WASSCE objective Economics items?
6. What is the Test Information Function (TIF) of 2014 WASSCE objective Economics items?
7. What is the relative efficiency of 2013 and 2014 WASSCE objective Economics items?

Scope of the Study

This study aimed at analyzing the Item and Test Information Functions of 2013 and 2014 West African Senior School Certificate Examinations in Economics objective test items in the North-central zone, Nigeria. The study covered all 183,576 senior secondary schools (SS3) students in North-central geo-political zone of Nigeria. The target population were all SS3 students that offered Economics in 2017/2018 session at the public senior schools in the North-central, Nigeria.

Simple random sampling technique was used to select three states for the study. The states selected were Kogi, Kwara and Niger which constitute the sample of the study. Five percent of the total number of schools in each of the three states were used to select to 32 schools using proportionate sampling technique. Thirty-five (35) students that offered Economics were purposively selected from the selected sample. A total of 1,120 respondents were used as the sample of this study. 2013 and 2014 WASSCE Economics past questions paper 1(objective test) were adopted and used to obtain data for this study. Percentage, mean, Parallel analysis and Multidimensional item response theory (MIRT)information indices was employed to analyze the data obtained.

Operational Definition of Terms

The following terms and variables are defined according to the way they are used in this study.

Dichotomous Items:	Test scored as 1 or 0
Information Function:	Statistical tools used for measuring precision of items
Item information Function:	Amount of information an Economics items will give about the test.
Polytomous Item:	Essay question items scored in categories.
Relative Efficiency:	Comparison of 2013 and 2014 WASSCE objective Economics items.
Test Information Function:	Sum of the information obtained from all Economics objective items in 2013 and 2014.

Significance of the Study

The finding of this study would be useful to the Economics teachers, parents, test developers, university administration, examination bodies, item writers and researchers.

The outcome of this study would be useful to teachers of Economics to take cognizance of the curriculum and content of the test to determine ways of placing learner on ability scale. It would help Economics teachers to write good test items and provide information on how to assist the weak students after diagnostic test has been carried out. The result from the findings of the study could serve as a corrective feedback to the learners as well as making parents aware of the performance of their children. This is because it could have precise information about the progress and level of knowledge attained by the learners. It gives information on the nature of the items used by the examining bodies so as to detect if the items are fair and free of bias.

It would make test developers and users to take cognizant of item and test used for assessment. It would be important for all professionals involved in evaluation of learners to be informed of current trend in assessment and evaluation. So, the outcome of the study would provide accurate test score information that could inform students, parents, schools, examining bodies and item writers about what students could attain as well as expose area where adjustment or remediation is required.

The finding of study would be useful to tertiary institutions especially universities that screen University Matriculation Examinations results and SSCE grades to determine admission grades. The result of the findings would enlighten and guide the management of higher institutions in using the score or grade of students at different year interchangeably.

The finding of this study would keep the public informed of the quality and the integrity of the examination conducted by WAEC and establish if the questions used are fair to all examinees or not. This would go a long way in establishing societal confidence in the acceptance

of the result conducted by WAEC. The result of this study would make examining bodies, item writers to take note of the type of items developed for public consumption. The psychometric properties of the test used would be established. It would provide a framework in which inferences about the location of examinees and items can be achieved.

It would make evaluators to improve on assessment and at the same time increase the existing body of knowledge as well as broaden the horizon of researchers on relative information and efficiency of test items.

CHAPTER TWO

REVIEW OF THE RELATED LITERATURE

This chapter is concerned with the review of literature adjudged to be relevant under the following sub-headings;

- (a) Item Response Theory (IRT)
- (b) Item Characteristic Curve and Item Parameter Logistic Model
- (c) Items and Test Information Functions (TIF) of IRT Model
- (d) Information Function and Relative Efficiency
- (e) Ability and Parameter Estimations
- (f) Group Invariance of Item and Ability Parameters
- (g) Information Function and Standard Error of Estimate
- (h) Theoretical Framework
- (i) Appraisal of Related Reviewed Literature

Item Response Theory

It is a psychometric paradigm for the construction, scoring, and analysis of items and test forms. With IRT, examinees are no longer scored by the number of items they answer correctly. Instead, IRT provides a precise estimate of their location on the underlying distribution of ability, analogous to the standard z scale. It provides a statistical model that relate ability to the probability of responding to an item. It is a scale on which examinees and items are placed. So, it is pertinent to have a measurement scale because examinees do not have the exact image of the ability. It is generally established that the IRT models can handle the measurement of multiple ability or dimensions because ability scale has a midpoint zero (Reckase, 2003). The majority of currently used IRT models make the assumption that a single trait or ability is being measured.

Test construction using IRT concern with estimation of item parameters or selection of items based on their information functions to attain a target function. IRT provides an avenue to assess level of measurement that will be equivalent at various points of ability considering different sets of items. The item analysis statistics used by IRT and CTT are fairly comparable, but that of IRT gives an additional item characteristic and a more sophisticated mechanism for conceptualizing measurement error. As a statistical modeling approach, IRT has other advantages over CTT, for instance all CTT concepts are designed for a specific sample, while the parameters of an IRT model is invariant of sample. Besides, the item parameter estimation model makes IRT to be more flexible, for example, different sets of items could be administered to individual test-takers and yet the result obtained could be compared from different tests.

For IRT model to be properly implemented certain assumptions need to be fulfilled. They are assumptions of dimensionality, which can be unidimensionality or multidimensionality. Unidimensionality assumption holds when only one trait is measured by the items in the test. Unidimensionality assumption requires that there is only one underlying trait for the data. The unidimensional models varying with the number of item parameters that are used. This implies that the covariance among the items can be established by a single inherent dimension. However, this assumption cannot be met strictly in reality because other factors like level of motivation or test anxiety may account for the test performance, So, if the data exhibit a dominant factor, it is considered that the unidimensionality assumption is met adequately (Hambleton et al., 1991).

Different statistical procedures such as factor analysis, multidimensional scaling, and DIM test, nonparametric tests for essential unidimensionality, and the use of multidimensional IRT models. Unidimensionality assumption must be carefully considered before employing a unidimensional IRT model. Factors like speededness or multiple content areas can give additional

dimensionality into a test. If unidimensionality assumption is violated, it implies that performances of examinees on certain set of items depend on their individual standing on two or more abilities. In other words, the ordering of examinee on these traits was not identical, so, it would not be possible to represent the interaction of examinee and task with a single person parameter for ability.

Unlike UIRT, Multidimensional IRT (MIRT) considers items as a grouped or clustered in different domains. It is designed to measure multiple constructs. Multidimensional models are capable of explaining the multiple ability and dimensions inherent in persons and tests. Researchers such as Reckase (2009) and DeMars (2010) have studied multidimensionality. The multidimensional assumptions are monotonicity and local independence. In monotonicity, there is positive relationship between the examinee's ability and probability of answering the items correctly. That is, the higher the examinee ability level, the higher the probability of the examinee answering any particular test item correctly (Smith, 2009). Local independence concerns with the probability of solving any item independent of the outcome of other item, that is it controls for both person and item parameters (Embretson and Reise, 2000).

Multidimensional IRT has two types of models. The compensatory and non-compensatory MIRT models. In compensatory MIRT models, deficiency in an ability can be offset by increasing the interaction of others. In non-compensatory MIRT models, deficiency in one ability cannot be completely offset by increasing the ability of others. Embretson and Reise (2000) reported that multidimensional IRT models are used to assess dimensionality of tests in which items reflect different skills, knowledge, or cognitive processes. Multiple content areas are commonly used for assessment of subjects like (reading, math, science) as stated by the No Child Left Behind Act of 2001 (NCLB, 2002). Assessment with multiple content areas may be more prone to

multidimensionality when compared to reading and mathematics. For this reason, there is need to test the dimensionality of economics test items when using IRT model.

Local independence is another assumption of IRT which requires that, the probability of examinees scored a particular item correctly is independent of another items. It states that the examinees' responses to any pair of items in the test are independent when examinees' abilities are held constant. So, for a given trait level, item responses should be uncorrelated. It is only when this criterion is met that the probability of a response would be seen as the product of the independent item.

Adedoyin, Nenty and Chilisa, (2008) submitted that invariance is the when the ability parameter of any two individuals is independent of the difficulty parameter of the items the persons attempted. The difficulty parameter of any two items is free of the ability parameter of the persons answering the items. So, the concept of parameter invariance established that item parameters are free of the group of examinees that attempted the items. This property is important in a CAT, where examinees take different sets of items of varying difficulties, but their ability levels can still be expressed on the same metric. Any time that a set of items refers back to the same stimuli, this assumption is threatened. When this assumption is violated, trait estimates may be inflated due to over estimation of item information Adedoyin, Nenty and Chilisa, (2008). So, an assumption is made when a particular IRT model is chosen that the model will fit the data. In other words, it must be assumed that the mathematical function (ICC) which represents the relationship between trait level and item response will be an accurate reflection of that relationship for the data.

Item Characteristic Curve and Item Parameter Logistic Models

Item characteristic curve shows the relationship between the likelihood of the correct response to an item and the ability. This relationship is modeled by a mathematical function called

item characteristic function and the graph of this function is called item characteristic curve (ICC). It is a linear or nonlinear function for the regression of item score on the trait or ability measured by the test (Hambleton, 1989). The assumption can be justified by how well the chosen IRT model accounts for the test data. The probability that an examinee correctly answers the item depends on the form of ICC. Therefore, the likelihood of correct response by an examinee does not based on number of examinees in the population who have the equal ability.

The number of parameters required to describe an ICC depends on the chosen IRT model. Usually, an ICC has one, two, or three parameters that are called item parameters. Generally, an ICC can have linear or nonlinear form even though nonlinear forms are more useful. The most popular mathematical form of ICC is the logistic form whose graph is an S-shaped curve. In fact, many people have proposed different forms of ICC such as step function or cubic function. This will bring significant insights on the relationship between ability and performance of the learners. The horizontal axis represents level of ability denoted by symbol θ , the vertical axis indicates the probability of correct response to the item denoted by $P(\theta)$. There are several ways to interpret $P(\theta)$. It can be considered as probability that a particular examinee at the ability level θ answers the item correctly. An alternative way is to interpret $P(\theta)$ as probability of an examinee's giving a correct answer to a randomly chosen item from a population of item supposed to measure the same ability. In other words, it is that $P(\theta)$ probability associated with a randomly selected examinee at ability level θ answers the item correctly.

ICC is S-shape curve in which the likelihood of correct response is near zero at the lowest ability and approaches 1 at highest ability level. The difficulty of the item and discrimination are important in item analysis. Adewuyi and Oluotun (2001) viewed difficulty of an item as the rate at which an item is been answered by the examinee correctly. Alonge (2003) expressed the

difficulty index as the percentages of the examinees that select the right option. The higher the difficulty index of an item the easier the question and the lower the difficulty index the more difficult the question items. Abiri (2006) indicated that multiple choice tests with fewer numbers of options have better difficulty index than those with larger number of options.

Item discrimination describes how well an item can make difference between examinees with low and high abilities. This deals with steepness of the item characteristic curve in its middle part. Discriminating power of multiple choice items is the ability to differentiate between the brilliant and poor students (Alonge, 2003). Discriminating power of a test ranges from zero to one (0-1). The closer this value is to one (1) the better the item is (Oyedeki, 1991). Baker (2001) expressed that most test items discriminate in a positive way while some items have negative discrimination.

Hambelton (1989) investigated the accuracy of estimation in the light of the guessing factor and reported that the guessing increases measurement error. De Ayala (1989) reported that a nominal response model provides higher ability estimation than a dichotomous model-based CAT. Zieba (2013) compared two-parameters Logistic model with one parameter and reported that the estimation of precision depends on the amount of information of parameter estimated. Ibrahim (2017) reported a significant difference in IRT methods of differential item function DIF in dichotomous and ordinal test items due to DIF.

IRT are the building blocks of testing that based on the scoring of response options into dichotomous and polytomous items. Item parameters depend on the IRT model that fits such items. For dichotomous items, the two parameter models are characterized by item discrimination power and difficulty. Examinees' responses to this type of items are discretely scored so that they will receive a score of 1 when their answer is correct and a score of 0 when incorrect answer is provided.

These models are based on the number of item parameters incorporated in each model and the mathematical expressions of the three models are also similar. These three models are primarily used to describe data scored in a binary, right/wrong fashion.

The purpose of any achievement test is to assess the true ability of examinees by answering each item based on the available information. This paves ways for the examinees to confirm if correct answer is chosen or not based on the specific knowledge possessed, otherwise, it will be answered incorrectly. Many a times, this assumption may not good enough especially when multiple-choice tests are involved with different responses.

Liao, Ho, Yen, & Cheng (2012) affirmed that the responses of an examinee in multiple-choice test can be grouped into three types: (i) responses that give true ability of the examinee (whether correct or incorrect), (ii) right responses arising from lucky guesses, and (iii) false responses arising due to anxiety, heedlessness, or distraction. The last two divergent responses may result into inaccurate estimation of the actual ability of the examinee because they do not reflect the exact knowledge of the examinee. Thus, guessing and heedlessness can result into biasness of item in estimation of computerized adaptive testing (CAT). Liao, Ho, Yen, & Cheng (2012) submitted that CAT is designed to monitor the administration of items according to each examinee's ability. The selection of each item of the test to be administered is based on the estimation of examinee's ability. The divergent responses of the examinees may result into ability estimation errors, hence, inappropriate selection of items arises. Where guessing is likely to be a factor affecting an examinee's test responses, the three-parameter logistic (3PL) model provides an appropriate solution for correctly estimating the examinee's ability (Liao, Ho, Yen, & Cheng, 2012). The 3PL model adds a guessing parameter (the lower asymptote) to care the effect of guessing.

Heedlessness or error due to anxiety may result in to estimation biases than what guessing does on the test. Rulison and Loken, (2009) reported that an examinee's ability underestimated when examinee makes careless errors at the beginning of a test. Barton and Lord (1981) introduced a fourth parameter (the upper asymptote) in the four-parameter (4PL). The model cares for underestimation of the examinees performance due to errors as a result of heedlessness. Liao, Ho, Yen, & Cheng, (2012) reported that 4PL model helped the examinee to recover from mistakes without reducing measurement precision and efficiency. They added that the 4PL model is a more robust mechanism against divergent responses. Magis (2013) reported that 4PL model is useful model to detect fitted data especially there is careless or inattention patterns of the examinees. The justification for 4pl model was established by Ayanwale, Adeleke & Abayomi (2018) who affirmed that Four-parameter model or Upper asymptote is capable of caring for incorrect answer as a result carelessness.

Baker (2004) submitted that two parameter model better calculates item information value than the three-parameter model. Each model has properties and assumptions which make it more or less attractive. The Rasch model has the advantage of requiring the fewest parameters, so it is easier to work with and provides fewer estimation problems than the other models. Dibu, Popoola, Ojo and Ifewulu (2014) reported that IRT has three different models parameter models. The item parameters, b , a , and c , are synonymously known as the item difficulty, item discrimination, and pseudo-guessing parameters (Zumbo, 2001). Though, four parameters model of IRT (4MPL) which proved effective has not gotten recognition among researchers because of its complexity. The model assumes that there are some items so difficult that even with extreme level of trait, not every examinee will respond to the item correctly (Ojerinde, 2013). It allows a high ability student to miss easy item without his or her ability being underestimated. The 4PM allows each item in

upper asymptote to be less than 1 (Linacre, 2004), to account for the possibility error that may occurred from a very high ability respondent.

Researchers such as Osgood, McMorris, & Potenza, (2002) analyzed a self-report delinquency scale using IRT with multinomial responses using 2PML. The steeper the slope of the ICC, the greater the discrimination, and the better the information provided by the item. In the 3PL model, the smaller the pseudo-guessing parameter, the greater information provided (Mutasem & Hassan 2013). However, in 4PM Linacre, (2004) and Waller & Reise, (2009) suggested that d as an item-specific upper asymptote and introduced an upper asymptote parameter, expressed by the lowercase d , into the 3PL model, resulting in the 4PL model: $P_{4PL}(\theta) = c + (d - c) \frac{1}{1 + \exp[-1.702a(\theta - b)]}$

Loken & Rulison, (2010) concluded that 4PL model can not only be used to correct guessing but imprecision and inefficiency caused by careless mistakes can be reduced.

Test can be dichotomously or polytomously scored. Dichotomous scoring is used for test like multiple choice items and alternative response items that requires an examinee to pick only one appropriate answer from a list of answers. Multiple choice items require less time for examinees to answer; easy to score and grade; provide greater coverage of material; allows for an easily assessment of an examinee's ability; and provide feedback about how distracters behaves and why correct answers are right. Multiple-choice items provide higher validity because of wider coverage of school syllabus. It paves way for uniform assessment of the examinees unlike essay items where optional items are allowed.

Polytomously scored item allows examinees to express their views or opinions on a specific concept. This gives the examinees opportunity to earn a particular score on a given instruction. Polytomous IRT models allow for the scoring of items when multiple response categories are

allowed. Instead of using a single item characteristic curve (ICC), multiple category characteristic curves (CCCs) would be used. An examinee with a given trait level will be most likely to respond in the category whose curve is highest for their trait level. Polytomous models use a single discrimination parameter for the item regardless of response option, such as the GRM, PCM, and GPCM.

Test scores are used to assess the abilities and skills of individuals competing for admissions or seeking professional certification. The evaluation of examinee's test scores may lead to a decision of excluding a candidate from certain academic program or to limit the ability of an examinee to practice the profession of his/her choice. The importance placed on the results of standardized testing make it essential that, the test scores should provide a fair and equitable evaluation of the skills and abilities, the test is purported to measure. The provision of a fair and equitable means of psychological or educational assessment is one of the major reasons for the existence of standardized testing. A test may become nonstandard when an unfair evaluation of an examinee's skills or abilities is done, which resulted into measurement error. The measurement error may occur due to physical testing environment, the administration of the test and the actual content of instruction and nature of psychometric characteristics of the test. Whether a test score is interpreted with regard to a norm, criterion or both, it must meet a specified or established standard. The norm may be established independently or by statistical analysis of a large number of participants.

Polytomous scoring item concerns with the use of open-ended items enables test designers to measure skills that cannot be measured by objective items. Ali, Chang. & Anderson (2015) reported that when polytomous items are involved, the two parameters model are more appropriate in selecting items to satisfy a particular test specification. Many testing programs used mixed

format tests in selected response items and constructed-response items. Given that the use of open-ended items is getting increased demand by the test constructors, and the tests combining these items with multiple choice items are being administered more often, problems and issues regarding this testing situation are studied from different aspects by researchers. Both dichotomously and polytomously-scored items are ubiquitous in educational and psychological test. Due to the different scoring scheme, polytomous IRT models can be used in item parameter calibration and ability estimation. In the dichotomous item, the test administrator classifies the observed responses into two categories: correct or incorrect. This dichotomization treats all incorrect answers as equivalent to one another.

If an examinee correctly performs all the operations and their response is categorized as a correct response, they receive a value of 1. Otherwise, the response is categorized in the incorrect category with assigned value of 0. The value assigned to responses reflects only whether the examinee has correctly performed a task (De Ayala, 2009). Polytomous IRT models are used in scoring of items when multiple response categories are given. Polytomous scoring include Likert type scaling for attitudes, essay scoring in which different values are awarded for different essay qualities, or any a situation when partial credit might be awarded to signify different levels of item performance. When a gradient of varying amounts of the trait measured applied, polytomous models would be appropriate. Polytomous IRT models are extended from the dichotomous models, but differ because multiple parameters are used to represent the probability of responding in each category rather than a single item difficulty parameter. Questions with three or more response categories, Samejima in 1969 proposed a model for graded response or ordered response. These parameters are called step difficulties or category boundaries, depending on the particular model chosen.

The ICC for polytomously scored responses can be Graded Response Model (GRM), Nominal Model, Partial Credit Model (PCM) and Rating Scale Model (RSM). Polytomous models have multiple category characteristic curves (CCCs) which show the relationship between trait level and the probability of responding to an item in a given category. An examinee with a given trait level will be most likely to respond in the category whose curve is highest for their trait level. Polytomous models use a single discrimination parameter for the item regardless of response options.

Various works has been carried out on dichotomous and polytomous scoring items and test information function. Lin (2007) studied polytomous items by fitting the GRM to the data and reported that deleting items with low discriminating values have a better impact on the information function. Thissen (2000) compared dichotomously scored items and Bock's nominal model and found that the nominal response model yielded more information than the dichotomously scored items, especially when incorrect responses contained certain useful information.

Mutasem and Hassan (2013) investigated on stability of IRT item-parameters estimation and reported that longer tests contain more information irrespective of sample size and ability level. They added that item difficulty parameter appeared to be the most sensitive to fluctuations in sample size and test length; while item guessing parameter appeared to be the least informative. Mutasem and Hassan (2013) reported that different samples yielded different level of accuracy of estimation and recommended 500 sample size and 30 test length as the minimum requirements for accurate parameters estimation but 60 items for 1000 respondents and 30 items for 2000 respondents using two Parametric Models.

Items and Test Information Functions

Information function of a test is used to evaluate how precisely individual items used in the test measure the level of a given latent trait. Fischer (as cited in Baker, 2001) defined information as the inverse of the precision with respect to a parameter estimated. Parameter estimated with precision would provide more information than the parameter not estimated precision. Ibrahim (2017) added that the precision and accuracy of test score can be determined with item parameter estimation

Precision is measured by the variability of the estimates around the value of the parameter. Hence, a measure of precision is the divergence of the estimators, which is denoted by σ^2 . The amount of information, denoted by I , is given by the formula: $I = \frac{1}{\sigma^2}$

The amount of information function is influenced by the level of ability of the examinees and the items of the test. Each item of the test measures the inherent ability. So, the amount of information based on a single item, can be computed at any ability level. Information function makes test expert to aware the contribution of each item to the total test. The implication of choosing a particular item is free of other items in the test. When a single item is used, the amount of information at that point on the ability point level will be smaller. Each item measures ability with greatest precision at the ability point similar to the item's difficulty parameter. The amount of item information reduces as the ability level move away from the item difficulty and closer zero at the extremes part of the ability level.

However, test information at a given ability level is the sum of the item information at that level. Thus, the test information function is defined as:

$$I(\theta) = \sum_{i=1}^N I_i(\theta)$$

Where: $I(\theta)$ is the amount of test information at an ability level of θ ,

$I_i(\theta)$ is the amount of information for item i at ability level θ ,

N is the number of items in the test.

The information function provides meaningful information about the precision of measurement at a particular ability level. Reckase (2009) submitted that test information is estimated using: $T_S(\theta) = \sum_{n=1}^N \text{item info}_N$

On the test there are 50 items, therefore, $T_S(\theta)$ becomes

$$T_S(\theta) = \text{item1info} + \text{item2info} + \text{item3info} + \dots + \text{item50info}$$

The test information function will be given higher value than a single item information function because as item increases, the better precision. Thus, longer tests will measure an examinee's ability with greater precision than shorter tests.

The highest value of the test information function will be achieved when ability estimation is at center point of the ability scale. The test information function tells how well a test behaves when estimating ability over the whole range of ability scores. Baker (2017) recommends while estimating the item or test information function that the amount of information must be transform into standard error of estimate. DeMars (2010) and Baker (2017) defined standard error of estimate (SE) as the inverse of the square root of the amount of item or test information. Baumgartner (2002) explained that in the response analysis, every estimate of item difficulty measure comes with its standard error, that the allowed limit of standard error for test is 0.5 and the smaller the standard error, the better the test item and the higher the reliability.

The information functions are used for test construction. Test construction under the IRT framework uses IIF and TIF to either construct or evaluate tests for any desired set of test specifications. IIF is used to build good test and SEM is used to interpret, report score and finding difference between tests.

Information Function and Relative Efficiency

Since item and test information function concerned with how well an item or test behave at a particular ability level considering the invariance of test items, so, it makes relative efficiency of the test becomes paramount issue. Relative efficiency index is related to the reliability of a test. It can be used to determine differences between two tests in order to know which of test item is better off (Crocker & Algina, 2003). It can also be used to evaluate the accuracy of the item parameters estimation. So, in IRT, every estimate of item difficulty measure comes with its

standard error of estimate. Standard error is inversely related with information function. i.e the smaller the standard error, the better the test item information and the higher the reliability.

De Ayala (2009) established the relative efficiency of the test as the ratio of the test information for two test form across the entire scoring scale. Efficiency of the two forms of test is to indicate the variability in the degree of the information provided by the types across the scoring scale. Relative efficiency determines the test types that provides higher amount of information at different point on the scoring scale. Relative efficiency varying as a function of the ability (De Ayala 2009).

De Ayala (2009) represents the relative efficiency of one score (X) to another score (Y) as the ratio of their information. The score X and Y may come from two different instrument that measure the construct or arise from scoring the same instrument in two different ways. Thus, the relative efficiency of the 2013 and 2014 WASSCE objective Economics test was obtained by dividing the test information of the 2014 WASSCE objective Economics test by the information of the 2013 WASSCE objective Economics test.

Moghadamzadah, Salehi and Khodaie (2011) reported that information function on two parameters logistic model was higher than that of one parameter and the three parameters. But no much difference in the 2PL and 1PL models Lam (2012) repoted that TIF is used to build a test to meet desired statistical specification. Lam (2012) studied the impact of item parameters on IRT information functions in mixed format tests and reported that the information increases when the number of items in the test increases.

Ability and Parameter Estimations

Ability parameters estimation is used as a determinant of ability of examinees on a given test. Ability parameter assumptions includes unidimensionality of the items, Local Independence,

and Item Characteristic Curves (ICC). The unidimensionality assumed that item of test measures only one trait and when there is only one dominant factor of ability. Local Independence of item implies that examinees responses to the items is statistically independent of the examinees ability at a given time. Ibrahim (2017) stressed that important decisions like conducting examinations for admissions and certification depend on test scores obtained then those scores are not only precise but also accurate. The latent trait means the ability of the learners may as a result of any of the following factors like the level of student's knowledge, intelligence, satisfaction, stress, and so on. The most important step in applying IRT model to test data is that of estimating parameters that specified selected model. All item parameters (a,b,c) and ability parameter (θ) must be estimated using test data (e.g. examinees' response to items).

The purpose of ability estimation is to determine ability level for each examinee and item parameters for each item from item responses. If the item parameters are established, then the estimation of the ability can be determined. Baker (2001) submitted that, the former ability value and predetermined values of item parameters must be estimated when establishing ability estimation. An adjustment of ability value is made to improve agreement between computed probabilities with the examinee's actual item scored. The process continues until the adjustment of ability becomes so small that the change in estimated ability is negligible. The process of item parameters estimation is similar to the process of ability estimation. Firstly, the initial value for each item parameter is given. These values along with known ability are used to compute the probability $P(\theta)$ at each ability level. The agreement between observed value of probability and computed probability is determined over the ability scale. So, adjustment to item parameters is made to improve the agreement between ICC estimated values of item parameter and the observed degree of correct response. This process of adjustment continues until the adjustments is minimal.

However, in general, parameter estimators are subject to estimation errors (bias, variance, and covariance). Adonu (2014) highlighted the procedures for estimating item parameters in IRT, they are; (i) correlation method (ii) regression method (iii) approximation (PROX) method (iv) maximum likelihood procedure. Theoretically a test maker may decide to create a test with a test information function that is very high over a large range of ability levels. This can be achieved by using a number of questions that have high information functions over a wide range of ability levels. Once a quality test has been formed, approximating ability levels based on how an examinee responded to the final step will follow. This is achieved by holding the item parameters stable and using methods to estimate the ability parameter. In IRT, estimation of the examinee ability parameter is straight forward when item parameters from the model are established. So estimation of examinee ability occurs after item parameters have been estimated

The examinee ability parameter is commonly calculated by Maximum Likelihood Estimation (MLE) because of its desirable asymptotic properties (i.e., the estimates are unbiased, consistent, and efficient as the number of items increases). The MLE approach to test scoring combines the probability expressions for all items of a test into a likelihood function. The term likelihood function is used because item responses are observed, which makes a probabilistic expression inappropriate. Note that in IRT, two examinees that answered the same number of responses correctly could have two different scores using item response theory depending on which items they responded correctly to. It is believed that IRT is a superior method for scoring an examinees ability, especially when considering high stakes exams.

There are two main techniques of estimating parameters for binary response IRT models: the maximum likelihood estimation which is mostly popular method of estimation. However, Bayesian estimation method can be more effective if prior information for item parameters is

available (Embreston & Reise, 2000). Bayesian model parameter estimation for IRT models is similar to marginal maximum likelihood estimation, though, Bayesian method requires background information of item parameters (Johnson, 2007). Maximum Likelihood estimation is a popular statistical method used to estimate the model's parameters through a joint probabilistic function of observed data. The likelihood function is equal to the product of probabilities associated with each item response if the local independence is true. De Ayala (2009) submitted that item and ability parameter estimation are important step in obtaining parameter value which produce curve that best fits the observed data. The IRT approaches were adopted in the analysis of mixed-item format tests. The number of items that is correct is used for the ability estimates and item statistics are used for item estimates. Zhang, Xie, Song, and Lu (2011) used a measurement-error approach and demonstrated how the uncertainty of item parameters can cause a large bias in ability estimation. So, maximum likelihood procedure performs better when estimating the item parameters and ability for all examinees (Baker, 2001).

Group Invariance of Item and Ability Parameters

Item parameters are group invariant because the values of each item estimated from another groups of examinees are the same. So, item parameters are independent of one another. The basic feature of IRT models is that the examinee's ability is not varying with respect to the items used (Baker, 2001). In other words, if other tests are used to estimate examinee's ability, it will give the same value. Thus, item parameters obtained from one group of examinees remain stable across other groups of examinees.

Mutasem and Hassan (2013) as cited in Hambleton (2000) stated that the length of a test and the number of examinees is required for proper estimation of item parameters. Larger samples and longer tests are required to provide accurate estimates, especially when both item and ability

parameters are being estimated and when complex IRT models are fitted to the data (Hambleton, 2000). So, the group invariance of item parameters and the item invariance of examinee's ability are what is called the principle of IRT invariance. This principle has given IRT more advantage over CTT.

Duong (2004) observes that ability concerns with the characteristics of examinees that the test intended to measure. It includes factual knowledge and skills of an examinees. MacDonald and Sampo (2002) observed that measurement based on IRT overcome the weaknesses of measurement based on classical test theory. So, item parameter estimates and the person ability estimates are not depending on the particular sample of examinees and test items administered. Therefore, meaningful comparisons of ability can be done when the parameters of the items used to make the different measurement are not the same.

Information Function and Standard Error of Estimate

Information functions provide a basis for selecting items from the item bank. It assists in test construction exercises. The item information function is used in determine the efficiency of the test items at different test takers' ability levels. Lord (1980) explained it is used to determine the reliability and error of measurement of the test. The standard error of estimate is the reverse of item information function which indicated non-precision of items or test measurement. Baker (2017) recommended that when interpreting item or test information function, the amount of information must be transformed into a standard error of estimate. Baker (2017) stated that standard error of estimate (SE) is obtained as the reverse of the square root of the amount of item or test information. The issue of standard error and information function are related with the reliability and validity.

One of the major contributions of IRT is that it serves as an extension of the concept of reliability. Traditionally, reliability refers to the precision of measurement (i.e., the degree to which measurement is independent of error). But IRT precision is not uniform throughout the entire range of test scores. Item and test information is an advanced concept in IRT to replace reliability (Hambleton & Swaminathan, 1985).

Validity measures how well the test measures that characteristic. Validity gives meaning to the test scores. It can tell you what you may conclude or predict about someone from his or her score on the test. If a test has been demonstrated to be a valid predictor of performance on a specific job, you can conclude that persons scoring high on the test are more likely to perform well on the job than persons who score low on the test, all else being equal.

Validity measures the degree to which specific conclusions or predictions about individual based on their test scores is. Validity tells how good a test performs at a given situation. Validity for instance indicates the linkage between test performance and job performance. Reliability tells how trustworthy a test score will behave. Valid conclusions could not be drawn from a test score expect test is reliability is obtained. It is noted that when a test is reliable, it may not be valid. Thus, it is paramount while selecting any test that it must be reliable and valid for prescribed purpose.

Theoretical Frame Work

This study is guided by Item Response Theory (IRT). Classical Test Theory (CTT) is traditional method of testing on which IRT is built upon. CTT models do not allow test developers to predict how well an examinee might do when dealing with a specific item. The most important limitation is that the examinee characteristics and item characteristics cannot be separated. They are dependent of each other. Test item characteristics change when the examinee context changes and examinee characteristics change when the test context changes. Therefore, it is difficult to

compare examinees who take different tests and to compare test /items whose characteristics are estimated from different groups of examinees.

Item Response Theory

It aims at measuring the inherent ability which produce the test performance rather than measuring test performance alone. This makes the IRT sample free because items statistics are independent of the test situation which generated the score. Item Response Theory (IRT) referred as modern psychometrics because it is used in large scale education assessment exercise, testing programs and professional testing firms.

Assumption of IRT Unidimensionality assumption assumed that item of test measures only one trait and when there is only one dominant factor of ability. In the IRT models differences of an individual is characterized by a single parameter affecting an examinee response to an item. It can be used with specialized Factor Analytic models when having dichotomous items.

Local independence of item implies that examinees responses to the items is statistically independent of the examinees ability at a given time. Local dependence occurs when distortions occurs in item parameter estimates. When local dependence exists, it means a large relationship between two or more items can essentially dominate the ability, thus making the scale not to have construct validity. Wainer and Lewis (1990) explained that, when local dependence is identified, the next step is to form testlets by combining locally dependent items. This can be resolved by deleting one or more of the local dependence items from the test items in order to achieve local independence.

For the purpose of any IRT study, there should be model fit assessment and dimensionality assessment.

Dimensionality assessment: In assessing dimensionality the concerned is to determine the number of dimensions that the test items will measure i.e to determine whether is unidimensional or multidimensional. The traditional approach for assessing dimensionality involves the use of factor analysis of a matrix. When one factor analyzes a correlation matrix derived from binary data, there is a possibility of obtaining artifactual factor (s) that are related to the nonlinearity between the items and the common factors. So, in this study multidimensionality IRT is appropriate for the data generated and parallel analysis implemented in FACTOR package version 10.5.03 was used.

Model –Data Fit Assessment

The fit assessment allows the researcher to know the exact model to choose for the data collected. BIGSTEPS and WINSTEPS are used for model-data fit assessment and produces two fit statistics. INFIT and OUTFIT, for examining model-data fit. INFIT and OUTFIT provide information on differences in responses, depending on whether the discrepancies occur close to or farther away from the estimated parameter. The fit statistics are calculated for both persons and items at the individual person and item levels. So, each individual and item have both the INFIT and OUTFIT statistic. Essen, Idaka & Metibemu (2017) stressed that fit of model to the data must accurately portray the true relationship between ability and performance on the item. They held that model misfit has negative consequences leading to violation of invariance property. The invariance property of item and ability is not dependent on the examinees distribution and characteristics of set of test items. So, Bolt (2002) believed that it is imperative for test developers to establish a particular model fits the data before and an item is valid

The assessment of item parameter model that fit the data is very paramount in determine item and test information function. This is to guide against difference between estimated value and

test information function. Item selection in IRT models is based on the intended purpose of the test. More so, the appropriate software to be selected depends on the nature data obtained. For dichotomous and polytomous BILOG and MULTILOG software can be used depending on the data fit statistic. However, the selection of items depends on the amount of information that contributed to the overall information given by the test (Hambleton, 2000).

In carrying out study using IRT, the following steps need to be taken:

1. Check if the assumption of IRT hold, if it does check the model chosen fit the data collected as well as the dimensionality assessment of the test.
2. Determine the nature of test used. Is the test essay or multiple- choice item, if it is multiple – choice, any of the four dichotomous models (1mpl, 2mpl,3mpl or 4mpl) and if it is essay polytomous models will be used.
3. Proceed to analysis using any of IRT software like BILOG, MULTILOG or *mirt* package. MULTILOG software is used for polytomous tests, BILOG and *mirt* package is used for dichotomous based on the dimension of the test items. When this is done the items and tests information can be determined.

Therefore, the study carried out the dimensionality assessment and fit statistics. Metibemu (2017) viewed Parallel analysis as a method based on the generation of random variables, for the determination of factors that underlies a test. The implementation of the parallel analysis procedure involves the generation of a large number of matrices (at least 500) of random data, with each matrix having the same number of variables and subjects as the real data set under assessment. Then each of the randomly generated data is factor analyzed and the mean of the resulting eigenvalues are obtained and thereafter compared with the eigenvalues of the matrix from the real

data to determine the factors underlying the real data (Horn, 1965). A factor is retained when its eigenvalue of real data is greater than the mean of the eigenvalues from its randomly generated counterpart (Lesdema & Valero-Mora, 2007). The extracted factor(s) provide evidence for evaluating the number of dimensions underlying the test under study. In this study parallel analysis implemented in FACTOR package version 10.5.03 was used. Unlike the traditional Horn's parallel analysis that reports the eigenvalues of the real and simulated data set, FACTOR reports the percentage of variance accounted for by the factors.

Thus, the relative estimate of factors is done with respect to percentage of variance noted for by the factors of the real and simulated data sets. Finch & French, (2015) and Oguoma, Metibemu and Okoye (2016) prominent among the measures to achieve fitness include Chi-square difference test and use of information indices. Information indices are simply measures of variance not explained by a model, with an added penalty for model complexity. The most popular of these indices are the **Akaike information criterion** (AIC; Akaike, 1973), the **Bayesian information criterion** (BIC; Schwarz, 1978), and the **sample-size-adjusted BIC** (SBIC; Enders & Tofih, 2008). These information indices are computed using the -2loglikelihood chi-square value and is interpreted such that the model with the lower value exhibits a better fit to the data. This study is in line with IRT procedure though, because the test items used were whole, thus, Parallel analysis and Multidimensional Item Response Theory (*MIRT*) information indices were employed to analysis the data gathered for the study.

Appraisal of Related Literature Reviewed

The literature reviewed in this study was carried out on (IRT), Item Characteristic Curve and Item Parameter Logistic Model, Items and Test Information Functions (TIF) of IRT Model, Information Function and Relative Efficiency, Ability and Parameter Estimations, Group

Invariance of Item and Ability Parameters, and Information Function and Standard Error of Estimate.

The theory is important in explaining the basic concepts in IRT. Studies in both dichotomous and polytomous scoring items information function are reviewed. The studies reviewed include, Samejima (1996) who examined the relationship between the trait level and item category parameters for the polytomously scored items and its effect on item information function. He reported that polytomous items are better with graded response model (GRM) because it produces higher information than dichotomous items.

Hambelton (1989) investigated the accuracy of estimation with guessing factor and reported that the guessing factor increases measurement error than other parameter models. In two-parameter model, the guessing factors are removed. Therefore, there is a negative relationship between the information function and the standard error of measurement. Zhang (2012) reported that the difference between the estimated and true test information function is due to uncertainty about item parameters in IRT models such as the biases, variances, and covariance of item parameter estimators. Though, Zieba (2013) reported that one-parameter model provides higher information for the value of the ability at a point where precision corresponds with item's difficulty.

Thissen (2000) compared the dichotomously scored items and Bock's nominal model and found that the nominal response model yielded more information than the dichotomously scored items, especially at the lower levels of the ability scale. Donoghue (1993) reported that polytomous scoring items provide much more information compare with short response items and multiple choice items. Moghadamzadah, Salehi and Khodaie (2011) reported that information function on two parameters logistic model was higher than that of one parameter and the three parameters. But

no much difference in the 2PL and 1PL models. Alagoz (2005) reported that both objective and constructed items provided only one set of scores on the items answered. Alagoz (200) added that increasing the length of test by combining two item types will add to overall measurement accuracy and the reliability of the test. Zieba (2013), Moghadamzadah Salehi & Khodaie (2011) and Alagoz (2005) who reported that information function on two parameters logistic model was higher than that of one parameter and the three parameters. But this is in contrary with the findings of Ojerinde and Onyeneho (2012) that 3PLM was more suitable for multiple choice ability tests. Loken & Rulison (2010) reported that the four parameter model is capable of reducing the estimated error made by high ability students who made a careless mistake. Waller & Reise, 2009) reported that the overall model fit may not necessarily improved with different parameter estimation but the test information function will change. Loken and Rulison (2010) reported that the 4PL model can not only be used to correct guessing but imprecision and inefficiency caused due to careless mistakes can be controlled.

Therefore, this study focused on the analysis of item and test information function in order to establish accuracy of the test independent of sample for generalization.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter is concerned with the methods and procedures employed to carrying out this study under the following sub-headings:

- a. Research Design;
- b. Population, Sample and Sampling Techniques; `
- c. Instrumentation;

- d. Procedure for Data Collection; and
- e. Data Analysis Techniques.

Research Design

The research design adopted for this study was the descriptive survey design. This is the most appropriate research design for this study. This is because the purpose of the study was to examine the item and test information function of 2013 and 2014 Senior School Certificate Examination Economics in North-central, Nigeria. Also, the researcher was not after manipulation of any variable but the respondents would give answers to test items and these responses were analyzed to make generalization. This makes the work descriptive in nature. Check and Schutt (2010) explained survey research as an efficient method for systematically collecting data from a broad spectrum of individuals and educational settings. The use of this design helped to provide answers to range of research questions generated in order to draw a conclusion.

Population, Sample and Sampling Techniques

The population for this study comprises all 183, 576 public senior secondary schools three (SS3) students in the North-central Nigeria. The target population for the study was all Senior Secondary (SS3) students of Economics in all senior secondary schools in North-central, Nigeria in 2017/2018 academic session. The states were Kwara, Kogi, Plateau, Niger, Benue, Nasarawa and Federal Capital Territory Abuja.

Multistage sampling procedure was adopted to select the sample of the study. At the first stage, simple random sampling technique was employed to select three states. The states selected were Kogi, Kwara and Niger which constituted the sample of the study. At the second stage, five

percent of the total number of the schools offering Economics in each of the three states were used to select to 32 schools.

At the third stage, 35 students that offered Economics were purposively chosen from each of the thirty-two 32 schools selected as sample. Purposive sampling was adopted because Economics as a subject is meant only for Humanities and Business students but optional for science students. So, the total sample used was 1,120 students. The respondents (SS3 students) selected for this study are in best position to response to the tests because they are preparing for WASSCE which is the focus of this study.

Instrumentation

The 2013 and 2014 WASSCE Economics papers 1 were adopted as instrument for this study. A total 100 multiple –choice items were used for the study. Economics Paper 1 of 2013 and 2014 contains 50 objective items each. The instrument was tagged Economics Achievement Test (EAT) and was categorized into (EAT1) and (EAT2) for 2013 and 2014 respectively. The same marks were allocated to both forms and the same number of questions were asked to answer. The objective items (paper 1) were scored dichotomously using 1 or 0. The scores obtained was subjected to IRT Information Function indices. The instrument was assumed to be valid since it was prepared and administered by an examination body then, it was assumed reliable because it has been used for the certification of students in the 2013 and 2014.

Procedure for Data Collection

A letter of introduction was collected by the researcher from the Head of Department, Department of Social Sciences Education, Faculty of Education, University of Ilorin, Ilorin,

Nigeria to seek for the permission of the selected school authorities (the principals and teachers) of the 32 schools chosen as sample for this study. This was to allow the selected students to respond to the instrument appropriately. After the approval was given, the researcher, with assistance of research assistant administered the test to the selected schools. The instrument was administered within two days in each of the chosen school, Economics Achievement Test (EAT I) which is 2013 Economics paper 1 lasted for one hour in the first day and Economics Achievement Test II (EAT II) which is 2014 Economics paper 1 also took one hour in the second day. The scripts were collected from the respondents after the required time for the tests and were marked with use of marking rubrics. The items were scored dichotomously.

Ethical Consideration

The researcher collected a letter of introduction from the Head of Department, Department of Social Sciences Education, Faculty of Education, University of Ilorin, Ilorin, Nigeria to obtain information from the students of the schools in the selected sampled states: Kwara, Kogi, and Niger. The researcher took the letter to the Ministry of Education of the selected states to collect the number of schools.

In administration of the tests, the researcher collected a letter from the Head of Department, Department of Social Science Education, Faculty of Education, University of Ilorin, Ilorin, Nigeria to seek for the permission of the selected school authorities (the principals and teachers) of the 32 schools chosen as sample for this study. This allowed the students selected to respond to the instrument Economics Achievement Test (EAT). These question items were administered with the help of the Economics teachers in the schools. The respondents (students of Economics) were duly informed to seek for the consent before the administration of the tests. All students were informed that information obtained would be handled with topmost confidentiality.

Anonymity was maintained as the identity of the students who responded to the test items was not exposed. This was done by not using names in the study. The issue of plagiarism was addressed as all works cited were properly referenced. At the end of the study, the work would be published for all stakeholders to benefit from.

Data Analysis Techniques

The data collected were scored dichotomously. A tick for correctly scored and zero for wrong answer. Percentage, Mean, Parallel analysis and Multidimensional IRT (*MIRT*) information indices were employed to analysis the data gathered for the study.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

This chapter is concerned with the analysis and results of the data collected for the study. The results were presented in tables for easy interpretation. Percentage was used to describe demographic data of the students (respondents) and this was followed by the results.

Demographic Characteristics of the Selected Sample

Table 2: Distribution of Students on the Basis of Gender and Mean Score

Gender	Frequency	Percentage (%)	Year	N	Mean
Male	512	45.7	2013	1120	0.7278
Female	608	54.3	2014	1120	0.5589
Total	1120	100			

Table 2 shows that out 1,120 students sampled for this study 512 (45.7 %) were males and 608 (54.3 %) were females. The mean score of 2013 items was 0.73 while that of 2014 items was 0.56.

Model-Data Fit Assessment for 2013 WASSCE Economics Objective Items

The goodness fit statistic was obtained through Chi-square difference test. The fitness of the four parameter models to the data set were compared. The model that produced the best fit to the data is adjudged the model that fit the data.

Table 3: Model-Data Fit Assessment of 2013 WASSCE Economics Objective Items

Factor	AIC	SABIC	BIC	-LogLikelihood	X2	Df	P
Comparing M1PL and M2PL models							
M1PL	64546.36	64917.17	65555.6	32072.18	461.997	43	0.000
M2PL	64170.37	64620.5	65395.51	31841.18			
Comparing M2PL and M3PL models							
M2PL	64170.37	64620.5	65395.51	31841.18	1795.158	50	0.000
M3PL	62475.21	63017.58	63951.41	30943.6			
Comparing M3PL and M4PL models							
M3PL	62475.21	63017.58	63951.41	30943.6	715.198	50	0.000

M4PL	61860.01	62494.63	63587.26	30586.01
------	----------	----------	----------	----------

Table 3 indicates that when the fitness of M1PL and M2PL models to the Economics data set were compared, the result showed that the M2PL had AIC = 64170.37, SABIC = 64620.5, BIC = 65395.51 values that were lesser than the AIC = 64546.36, SABIC = 64917.17, BIC = 65555.6 values of the M1PL. In addition, the Likelihood ratio test that M2PL fitted the data better than 1PL with statistically significant ($\chi^2(43) = 461.997$, $p < 0.05$). The showed that the 2PL model fitted the data well than the 1PL model. The fitness of 2PL model was compared to the fitness of 3PL model to the test dimension of the data.

The 3PL model and 4PL model were compared and the results showed that 4PL model fitted the data better than the 3PL model (M4PL model's AIC = 61860.01, SABIC = 62494.63, BIC = 63587.26 values were lesser than the M3PL model's AIC = 62475.21, SABIC = 63017.58, BIC = 63951.41 values. Likelihood ratio test that 4PL model fitted the data better than 3PL was statistically significant, $\chi^2(50) = 715.198$, $p < 0.05$). These results showed that 4PLmodel fitted the data better than 1PL, 2PL and 3PL models. These results showed that the Multidimensional 4-parameter logistic model fitted the 2013 WASSCE Economics test with 4 dimension.

IRT Model fitted the 2014 WASSCE Economics Objective Items

Table 4: Model-Data Fit Assessment of 2014 WASSCE Economics Objective Items

Model	AIC	SABIC	BIC	-2logLikelihood	X2	Df	P
Comparing M1PL and M2PL model-data fitness							
M1PL	62514.13	62700.45	63021.26	-31156.06	3029.01	48	0.000
2PL	59581.12	59855.99	60329.26	-29641.56			
Comparing M2PL and M3PL model-data fitness							
M2PL	59581.12	59855.99	60329.26	-29641.56	1467.621	50	0.000
M3PL	58213.5	58580.61	59212.69	-28907.75			
Comparing M3PL and M4PL model-data fitness							
M3PL	58213.5	58580.61	59212.69	-28907.75	488.245	50	0.000
M4PL	57825.25	58284.61	59075.5	-28663.62			

Table 4 indicates that when the fitness of M1PL and M2PL models to the Economics data set were compared, the result showed that the M2PL had AIC = 59581.12, SABIC = 59855.99 and BIC = 63021.26 values that were lesser than the AIC = 62514.13, SABIC = 62700.45 and BIC = 60329.26 values of the M1PL model. In addition, the Likelihood ratio test indicated that 2PL model fitted the data better than 1PL which was statistically significant at ($\chi^2(48) = 3029.01$, $p < 0.005$). These results indicated that the 2PL model fitted the data better than the 1PL model. In an attempt to look for another better model for the test data, the fitness of 2PL model to the Economics data was in turn compared to the fitness of 3PL model to the test dimension of the items. The result also showed that the 3PL model fitted the data well than the 2PL model with (3PL model's AIC = 58213.5, SABIC = 58580.61, BIC = 59212.69 values respectively. This is lesser than the 2PL model's AIC = 59581.12, SABIC = 59855.99, BIC = 60329.26; the Likelihood ratio test that 3PL model fitted the data better than 2PL model was statistically significant, ($\chi^2(50) = 1467.621$, $p > 0.005$)).

The model-data fit for M3PL model and M4PL model were also compared. The results showed that 4PL model fitted the data better than the 3PL model with (M4PL model's AIC = 57825.25, SABIC = 58284.61, BIC = 59075.5 values lesser than the M3PL model's AIC = 58213.5, SABIC = 58580.61, BIC = 59212.69 values respectively; Likelihood ratio test that M4PL model fitted the data better than 3PL was statistically significant, $\chi^2(50) = 488.245$, $p > 0.005$). These results showed that M3PL model clearly fitted the data better than M1PL and M2PL models. More so, results of this study showed that the Multidimensional 4-parameter logistic model fitted the 2014 WASSCE Economics test.

Answering Research Questions

The scores obtained from 2013 and 2014 Economics Achievement Test (EAT I and II) in the three states of the North–central geopolitical zone of Nigeria were analyzed.

Research question 1: What is the dimensionality of 2013 WASSCE objective Economics items?

To answer this research question, the responses of the examinees to 2013 WASSCE objective Economics items was calibrated with four parameter model (M4PL) with 4 dimensions. The research questions were answered through parallel analysis using FACTOR package version 10.5.03.

Table 5: Parallel Analysis of the Responses of Examinees in 2013 Economics Objective Test

Factors	Real-data % of variance	Mean of random % of variance	Factors	Real-data % of variance	Mean of random % of variance
1	11.5*	4.1	26	1.6	2
2	7.4*	3.9	27	1.6	1.9
3	3.9*	3.8	28	1.5	1.8
4	3.8*	3.7	29	1.5	1.8
5	3.5	3.6	30	1.4	1.7
6	3.4	3.5	31	1.4	1.6
7	3.1	3.4	32	1.3	1.5
8	3.1	3.3	33	1.2	1.5
9	3	3.2	34	1.2	1.4
10	2.8	3.1	35	1.2	1.3
11	2.7	3	36	1	1.3
12	2.7	2.9	37	1	1.2
13	2.5	2.9	38	0.9	1.1
14	2.5	2.8	39	0.9	1
15	2.4	2.7	40	0.8	1
16	2.3	2.6	41	0.7	0.9
17	2.2	2.6	42	0.7	0.8
18	2.2	2.5	43	0.6	0.7
19	2.1	2.4	44	0.4	0.6
20	2.1	2.4	45	0.4	0.6
21	2	2.3	46	0.2	0.5
22	1.9	2.2	47	0.2	0.4
23	1.8	2.2	48	0.1	0.3
24	1.7	2.1	49	0	0.2
25	1.7	2	50	0	0

* Advised number of dimensions: 4

Table 5 indicated that the analysis of the number of factors that underlying the WASSCE 2013 Economics objective using parallel analysis. On the table, column 1 represents the factors of the 2013 WASSCE Economics test and the randomly generated set of data. Column 2 and 3 represent the percentage of variance responsible each factor of the real and randomly generated set of data. For Factor 1, the variance of the real data is 11.5% while the random variance was 4.1%. Thus, the variance of real data for factor 1 is greater than the random data of factor. ($11.5\% > 4.1\%$). In Factor 2, 3, and 4 the variance of the real data and random data were ($7.4\% > 3.9\%$, $3.9\% > 3.8\%$ and $3.8\% > 3.7\%$) respectively. So, the percent of variance of real data of four factors are more than the percent of random data in this analysis. The result shows that there were four factors underlying the performance of examinees in 2013 WASSCE Economics test. Thus, the 2013 Economics Achievement Test was calibrated using MIRT model with 4 dimensions.

Research Question 2: What is the dimensionality of 2014 WASSCE Economics Objective Test?

To answer this research question, the responses of the examinees to 2014 WASSCE objective Economics items was calibrated with M4PL model with 2 dimensions. The research question was answered through parallel analysis using FACTOR package version 10.5.03.

Table 6: Parallel Analysis of the Responses of Examinees to 2014 Economics Objective Test

Factors	Real-Data %of variance	Mean of random % of Variance	Factors	Real-Data %of variance	Mean of random % of variance
1	24.4*	4.1	26	1.3	2
2	12.1*	3.9	27	1.2	1.9
3	3.2	3.7	28	1.2	1.8
4	3	3.6	29	1.1	1.8
5	2.8	3.5	30	1	1.7
6	2.7	3.4	31	0.9	1.6
7	2.6	3.3	32	0.9	1.6
8	2.5	3.3	33	0.8	1.5
9	2.5	3.2	34	0.8	1.4
10	2.4	3.1	35	0.7	1.3
11	2.3	3	36	0.7	1.3
12	2.2	2.9	37	0.6	1.2
13	2.2	2.9	38	0.6	1.1
14	2.1	2.8	39	0.5	1
15	2.1	2.7	40	0.5	1
16	2	2.6	41	0.4	0.9
17	1.9	2.6	42	0.3	0.8
18	1.9	2.5	43	0.3	0.7
19	1.7	2.4	44	0.3	0.6
20	1.6	2.4	45	0.2	0.6
21	1.6	2.3	46	0.2	0.5
22	1.5	2.2	47	0.1	0.4
23	1.4	2.2	48	0	0.3
24	1.4	2.1	49	0	0.2
25	1.3	2	50	0	0

* Advised number of dimensions: 2

Table 6 indicates that there were two factors of the real set of data with which the percentage of variance greater than the percentage of variance of the generated data set. From table 6, column 1 represents the factors of the 2014 WASSCE Economics test and the randomly generated set of data. The result shows that there were two factors that underlie the performance of examinees in 2014 WASSCE Economics test in the North-central. Column 2 and 3 represent the percentage variance accounted for by the factors of the real and randomly generated data sets

factors. The table 3 shows that there were two factors of the real data set with percent of variance greater than the percent of variance of the generated data set. The percent of variance of real data and random data for Factor 1 are $24.4\% > 4.1\%$ and for factor 2, it was $12.1\% > 3.9\%$. So, the variance of real data two factors out fifty was greater than the random data in this analysis. The result shows that there were two factors that underlie the performance of examinees in 2014 WASSCE Economics test. Thus, the 2014 Economics Achievement Test was calibrated using MIRT model with 2 dimensions.

Research question 3: What is the item information function of 2013 WASSCE Economics test?

Table 7: Summary of the Standard Error of Estimate of the 2013 WASSCE Economics Objective Test

Item	percent of examinees whose ability can be precisely measured	No of examinees the items measured precisely	percent of examinees whose ability cannot be precisely measured	Item	percent of examinees whose ability can be precisely measured	No of examinees the items measured precisely	percent of examinees whose ability cannot be precisely measured
1	1.5	17	98.5	26	0.4	4	99.6
2	0.2	2	99.8	27	1.2	13	98.8
3	0.4	4	99.6	28	0	0	100
4	0.4	4	99.46	29	0	0	100
5	0.4	4	99.6	30	1	11	99
6	0	0	100	31	1.3	14	98.8
7	2.9	32	97.1	32	2.7	30	97.3
8	0	0	100	33	0.6	6	99.4
9	1.3	14	98.8	34	0	0	100
10	0	0	100	35	0	0	100
11	0	0	100	36	2.5	28	97.5
12	2.9	32	97.1	37	11.5	128	88.5
13	2.1	23	97.9	38	0.4	4	99.6
14	2.9	32	97.1	39	0	0	100
15	0.3	3	99.7	40	0	0	100
16	1.7	19	98.3	41	0.4	4	99.6
17	0	0	100	42	1.8	20	98.2
18	1.7	19	98.3	43	0	0	100
19	2.4	26	97.6	44	0	0	100
20	0	0	100	45	0.7	7	99.3
21	0	0	100	46	0.2	2	99.8
22	2.1	23	97.9	47	0	0	100
23	0	0	100	48	0	0	100
24	0	0	100	49	0	0	100
25	1.7	19	98.3	50	0	0	100

Table 7 indicates that item1 can be precisely measure the ability of 1.5% (17 examinees) out of the 1,120 examinees that took 2013 EAT 1(the proficiency of examinees in 2013 WASSCE Economics test) and that 98.5% of examinees ability cannot be precisely measured. Item 37 who has highest examinees of (128 examinees) shows that 11.5%. of the examinees ability can precisely

measure and that 88.5% examinees ability cannot be precisely measured. However, a total of 21 items cannot measure the ability of any examinee precisely.

Research Question 4: What is the item information function of 2014 WASSCE objective Economics items?

Table 8: Summary of the Standard Error of Estimate of the 2014 WASSCE Economics Multiple-Choice Test

Item	Percent examinees whose ability can be precisely measured	No of examinees the items measured precisely	Percent examinees whose ability cannot be precisely measured	Item	percent examinees whose ability can be precisely measured	No of examinees the items measured precisely	percent examinees whose ability cannot be precisely measured
1	0	0	100	26	2.6	29	97.4
2	0.9	10	99.1	27	7.7	86	92.3
3	0.1	1	99.9	28	0	0	100
4	0.2	2	99.8	29	0	0	100
5	0	0	100	30	0.1	1	99.9
6	0	0	100	31	6.5	72	93.5
7	0	0	100	32	0.4	4	99.6
8	3.4	38	96.6	33	0	0	100
9	0	0	100	34	2.1	23	97.9
10	9.5	106	90.5	35	6	67	94
11	0	0	100	36	0	0	100
12	11.1	124	88.9	37	6.4	71	93.6
13	0.4	4	99.6	38	0	0	100
14	5.6	62	94.4	39	6.9	77	93.1
15	0	0	100	40	0	0	100
16	0	0	100	41	1.2	13	98.8
17	0	0	100	42	0	0	100
18	0	0	100	43	1.3	14	98.7
19	9	100	91	44	0	0	100
20	3.1	34	96.9	45	0	0	100
21	8.2	91	91.8	46	0	0	100
22	0	0	100	47	7.1	79	92.9
23	0	0	100	48	0	0	100
24	0	0	100	49	0	0	100
25	6.9	77	93.1	50	0	0	100

Table 8 indicates that item 1 cannot be precisely measured the ability of any examinee who took the test. Item 2 can be precisely measure the ability of 0.9% (10 examinees) out of 1,120

examinees that took 2014 WASSCE Economics test and that 99.1% of examinees ability cannot be precisely measured. Item 3 can be precisely measure the ability-y of 0.1% (1 examinee) out of 1,120 examinees and 99.9% of examinees ability cannot be precisely measured. Item 12 who has highest examinees of (124 examinees) shows that 11.1% of the examinees ability can precisely measure and that 89.9% examinees ability cannot be precisely measured. Thus, 26 Items cannot measure the ability of any examinee precisely measure. The implication of this study was that most of these items provide less information.

Research question 5: What is the Test Information Function (TIF) of 2013 WASSCE Objective Economics Items?

To answer this research question, the Test information of the 2013 WASSCE Economics test was assessed. To achieve this, the information of the 50 items of the test for each of the examinees that took the test was examined.

Table 9: Summary of the Standard Error Estimate of 2013 WASSCE Economics Test

	Percent
Percent examinees whose ability can be precisely measured by the test	17.5
percent examinees whose ability cannot be precisely measured by the test	82.5

Table 9 presents the summary of the estimated SE of the 2013 WASSCE Economics. This estimate was obtained using the amount of information underlying the test at the level of ability of the examinees who took the test (see Appendix III for the information of 2013 WASSCE Economics test). The table shows the percentage of examinees whose proficiency in Economics test was able to precisely. The table showed that the 2013 WASSCE Economics test was able to measure precisely the Economics proficiency of 17.5% of the examinees that took the test. The result showed that the information of the test was low. This implies that the extent to which the

2013 WASSCE Economics objective test precisely measures student's proficiency in Economics was low.

Research question 6: What is the Test Information Function (TIF) of 2014 WASSCE Objective Economics Items?

To answer this research question, the test information of the 2014 WASSCE Economics test was estimated. To achieve this, the information of the 50 items at the ability level of the examinees on the test were also summed.

Table 10: Summary of the Standard Error Estimate of 2014 WASSCE Economics Test

	Percent
percent examinees whose ability can be precisely measured	5.4
percent examinees whose ability cannot be precisely measured	94.6

Table 10 presents the summary of the estimated SE of the 2014 WASSCE Economics. This estimate was obtained using the amount of information underlying the test at the level of ability of the examinees who took the test (see Appendix IV for the information of 2014 WASSCE Economics test). The table shows the percentage of examinees whose proficiency in WASSCE Economics test was able to precisely measure. The table 9 showed that the 2014 WASSCE Economics test was able to measure precisely the Economics proficiency of 5.4% of the examinees that took the test. The result showed that the information of the test was also low. This implies that, the extent to which the 2014 WASSCE Economics objective test precisely measures student's proficiency in Economics was low.

Research question 7: What is the relative efficiency of 2013 and 2014 WASSCE Objective Economics items?

The result of RE of the tests for the scores of the examinees who took the tests are presented in Table 11.

Table 11: Relative Efficiency of 2013 and 2014 WASSCE Economics Test Items

ID	RE	ID	RE	ID	RE	ID	RE	ID	RE	ID	RE	I D	R E	I D	R E
1	490.8	2	6.1	3	0.4	4	38.1	5	0.6	6	10. 0				
9	2.0	10	0.6	11	2.4	12	1.7	13	1.2	14	0.6 14.				
17	0.0	18	0.9	19	4.7	20	0.1	21	41.7	22	3 26.				
25	8.1	26	9	27	3	28	12.2	29	1.3	30	9				
33	81.2	34	7.4	35	0.0	36	1.4	37	0.4	38	1.2				
41	0.2	42	8.7	43	0.0	44	0.0	45	0.1	46	0.0				
49	1597. 6	50	1.1	51	1.1	52	0.1	53	0.4	54	2.1				
57	9.8	58	0.0	59	19.2	60	5.4	61	2.3	62	0.3				
65	0.4	66	0.3	67	1.3	68	6.3	69	2.0	70	3.6				
+	+	+	+	+	+	+	+	+	+	+	+				
+	+	+	+	+	+	+	+	+	+	+	+				
+	+	+	+	+	+	+	+	+	+	+	+				
+	+	+	+	+	+	+	+	+	+	+	+				
+	+	+	+	+	+	+	+	+	+	+	+				
+	+	+	+	+	+	+	+	+	+	+	+				
109		109	30.	109		110		110	443.	110					
7	37.9	8	5	9	31.5	0	9.2	1	7	2	1.2				
110		110	81.	110		110		110		111					
5	67.9	6	6	7	0.5	8	0.4	9	0.0	0	0.5				
111		111		111		111	541.	111	614.	111	13.				
3	0.0	4	0.0	5	0.1	6	8	7	5	8	6				

The table 11 shows the abridged of the relative efficiency of 2013 and 2014 WASSCE Economics test of all the 1,120 examinees in Economics Achievement Test (EAT) 2013 and 2014 that participated in the study. The detailed of (relative efficiency of 2013 and 2014 WASSCE Economics test of all the 1,120 examinees is showed in appendix V). For examinee 1, there was

80% difference between 2013 and 2014 WASSCE Economics items. That 2014 WASSCE Economics items were 80% more informative than 2013 for examinee 1. So, no matter the level of ability of the examinee 1, there was 80% difference between 2013 and 2014 WASSCE Economics items. For examinee 1103, 1105 there was 90% difference between 2013 and 2014 WASSCE Economics and that 2013 WASSCE Economics was 90% more informative than that of 2014 WASSCE Economics items. But, examinees 6, 9,17,35 among others have 0% difference between 2013 and 2014 WASSCE Economics. This implies that there was no difference in 2013 and 2014 WASSCE Economics items for examinee 6,9,17,35 among others. The implication of the study was that both 2013 and 2014 WASSCE Economics test items were less informative. But, 2013 WASSCE Economics test items provide relatively more information than the 2014 WASSCE Economics test items.

Summary of Findings

Based on the data collected, collated, analyzed and interpreted, the following findings were obtained that:

1. 2013 WASSCE Economics objective test measured multidimensional traits with 4 dimensions;
2. 2014 WASSCE Economics objective test measured multidimensional traits with 2 dimensions;
3. 2013 WASSCE Economics objective items provide less information with mean of 0.73;
4. 2014 WASSCE Economics objective items provide least information with the mean of 0.56;
5. 2013 WASSCE Economics test measured the precision of 17.5% of the examinees that took the test;

6. 2014 WASSCE Economics test measured the precision of 5.4% of the examinees that took the test; and
7. 2013 and 2014 WASSCE Economics test items were less informative with 17.5% and 5.4% respectively, but 2013 WASSCE Economics test items was relatively more informative than that of 2014 items.

CHAPTER FIVE

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

This chapter contains the discussion of the results presented in the chapter four, conclusions of the study based on the findings, recommendations to the relevant stakeholders and suggestions for further study.

Discussion of Findings

It was observed from the results obtained from the 2013 Economics Achievement Test (EAT I) shows that there were four factors of the real data in which the percent of variance greater than the percent of variance of the generated data set. The result shows that there were four factors underlying the performance of examinees in 2013 WASSCE Economics test. That is the dimension of 2013 WASSCE Economics test was multidimensional, which implies that the 2013 WASSCE Economics test measured 4 dimensions of examinees' ability. It showed that there were two factors of the real data set which percent of variance greater than the percent of variance of the generated data set. The percent of variance of real data and random data for Factor 1 are $24.4\% > 4.1\%$ and for factor 2, it was $12.1\% > 3.9\%$. So, the variance of real data of two factors out fifty was greater than the random data in this analysis. The result shows that there were two factors underlie the performance of examinees in 2014 WASSCE Economics test. Thus, the 2014 Economics Achievement Test was multidimensional, which implies that the 2014 WASSCE Economics test items measured 2 dimensions of examinees' ability.

The results obtained from the data of the study indicated that the dimensions of the 2013 and 2014 WASSCE Economics test items were multidimensional and measured more than one examinees' trait. The results also showed that M4PL model fit the data better than M1PL, M2PL and M3PL models. This result is in agreement with the study of Osgood, McMorris & Potenza

(2002) reported that a 4PL model better fit to the data than the others. Loken & Rulison (2010) reported that the overall fitness is improved when using the 4PM than other parameters. Loken & Rulison (2010) added that 4PM allows each item in upper asymptote error occurred due careless mistake of even a very high ability respondent

However, the study was contrary to the study of Donoghue (1993), Samejima (1996), Alagoz (2005), Moghadamzadah, Salehi & Khodaie (2011), Zieba (2013) and Ibrahim (2017). Donoghue (1993) reported that polytomous scoring items provide much more information than those from the short response items and multiple choice items. Samejima (1996) revealed that polytomous scoring items yielded more IRT information than the optimal dichotomization of the same items. Alagoz (2005) reported information function on two parameters logistic model was higher than that of one parameter and the three parameters. Moghadamzadah, Salehi and Khodaie (2011) and Zieba (2013) reported that information function on two parameters model was higher than one parameter and the three parameters models. But there is no much difference in the 2PL and 1PL models.

The results revealed that the standard error of estimate for the 2013 WASSCE Economics test items was higher which resulted into less the item information function. Thus, the smaller the standard error, the better the test item information and the higher the precision of the test items. Since, the standard error of estimate of 2013 WASSCE Economics objective is higher, the information function of the items will be lesser and the precision of the test items were also less. The results showed that the 2013 WASSCE Economics test items estimated 17.5% of examinees ability with precision. This indicated that the information function of these items were less with mean of 0.73. In the same vein, results also revealed that the standard error of estimate for the 2014 WASSCE Economics test items is higher and this makes the items to be least informative.

The results of the study revealed that the 2014 WASSCE Economics test measure precisely the Economics proficiency of 5.4% of the examinees that took the test. The results showed that the percentage of examinees that the test measured their proficiency was lower than that of 2013 with mean of 0.56.

The implication of the study was that the performance of examinees in EAT 1 and EAT 2 were low but that of EAT 2 was relatively lower compare with that of EAT 1 among senior secondary school students in the North-central geo-political zone of Nigeria. These findings were concord with the study of Adeyegbe, (2002), Onipede, (2003), Adu and Ayeni, (2004), Adeyemi, (2010), Alaka and Obadara (2013), Adewuni, (2016), and Abdulsalam-Nuhu, (2018). Adeyegbe (2002) reported that there was a fall in the performance of students in Senior Secondary Certificate Examination. Onipede (2003) reported that performance of the students in Senior Secondary Certificate Examination (SSCE) was below expectations across the nation in many subjects. Adu and Ayeni (2004) reported that pitiful results from miserable performance of the students may arise as a result of teacher's interaction, student's failure to ask questions and use of lecturing method when given instruction. In the same vein, Adeyemi (2010) reported that the performance of students in (SSCE) was low. Alaka and Obadara (2013) reported that there is poor students' academic performance in all geo-political zones of Nigeria. Adewuni (2016) also reported that the performance of students in NECO Senior School Certificate Examination objective test in Government for 2013 and 2014 were low. More so, Abdulsalam-Nuhu (2018) reported that the performance of examinees in practical Geography achievement test was on average.

The results of the study showed that both 2013 and 2014 WASSCE Economics test items were less informative. However, 2013 WASSCE Economics test items provide relatively more information than that of 2014 WASSCE Economics test items. These indicated that the

information function of 2013 and 2014 WASSCE Economics test items measure the same dimensionality and can be used interchangeably. Though, the 2013 WASSCE Economics test items were relatively better than the 2014. This may be, because the 2013 and 2014 WASSCE Economics objective items was a part of the score used to grade students and placed them for admission at higher institutions. The test items that combined the objective and essay may provide greater information.

Finding from the results also revealed that there is significant difference between the test information function of 2013 and 2014 WASSCE Economics objective test items. The precision with which 2013 WASSCE Economics test estimates the proficiency of examinees was better than the precision with which 2014 WASSCE test measures students' proficiency in Economics in North –central, Nigeria. This is line with the study of Lam (2012) who investigated the relative efficiency of the constructed response items, short answer items and objective revealed that polytomous items provided more information when compare with the dichotomously scored items. Lam (2012) added that short answer items also provided more information than Objective-items. Thus, the ratio of the 2013 WASSCE Economics test items was greater than that of 2014 test items.

Conclusion

The study concluded that items and tests of 2013 and 2014 WASSCE Economics can be used interchangeably. Although, the 2013 test items provided relatively higher information than that of 2014.

Implications of the Study

The implication was that Economics items used for the assessment possessed the same multidimensional traits.

Other implications of the findings are to expose:

1. examination bodies like WAEC to the use information function of each item and the need to research on dichotomous and polytomous items;
2. examination bodies like WAEC, educational researchers and evaluators to the use of information function of each items for item selection and banking;
3. examination bodies like WAEC, educational researchers and evaluators on the need to research in another subject area;
4. classroom teachers and educational researcher to use IRT model and estimate its items parameter;
5. teachers and educational researchers to the use of multidimensional IRT model in order to determine the dimension of the test items; and
6. teachers and educational researchers to the use of 4PL model so as to save high ability students who because of anxiety or wrong interpretation of item would have gotten the item right

Recommendations

The following were recommended based on the findings and conclusion of this study:

1. Evaluators, test developers, teachers and examination bodies should be trained on the use of IRT framework. This will ensure its usage and popularity in our environment.
2. Examinations bodies like WAEC, NECO, NABTEB, JAMB and others should adopt IRT information function in order to reduce measurement error faced in comparison of scores and grades.
3. Different IRT software should be made available to the researchers by the educational managers and school authorities.

4. Seminar, training and re-training on psychometric properties of test should be organized to the teachers and the lecturers in order to improve the quality of testing in schools.
5. The government, ministries of education and other stakeholders should procure different IRT software and as well as sponsor the training of individual to learn the statistical analysis using IRT software package. This is to ensure its usage and popularity when necessary.

Limitations of the Study

The study was limited to 1,120 students in the North-central, Nigeria because of finance and time. The resources available limit the study to multiple choice items alone which is an aspect of Economics examination used to grade the students and placed them for admission at the higher institutions.

Suggestions for Future Study

This study was carried out to analysis the items and test information functions of senior secondary school certificate examination Economics in the North-central, Nigeria. The sample size could be increase by other researchers in future. Subjects like Mathematics, English language, Government and others could be used instead of Economics by other researchers in the future study. In this study, dichotomous items were used. Future study may combine both the dichotomous and polytomous scoring items. This study used (MIRT) information indices which is different from others used when both dichotomous and polytomous scoring models are involved.