Comparative Analysis of Item Parameter Estimates of 2016 Physics WAEC and NECO Senior School Certificate Examination Items Using the 3-Parameter Model of Item Response Theory

Dr. Temitope Babatimehin*

Department of Educational Foundations, Obafemi Awolowo University, Ile-Ife, Nigeria.

***Corresponding Author**

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110400150

Received: 18 April 2026; Accepted: 23 April 2026; Published: 16 May 2026

ABSTRACT

This study analysed 2016 WAEC and NECO Senior School Certificate Physics Examination Items using Item Response Theory. The study determined the difficulty levels of 2016 WAEC and NECO Physics objective tests. It also investigated the discriminating power of items on WAEC and NECO Physics objective tests and finally ascertained the difference in guessing parameter of WAEC and NECO 2016 physics objective test items. These were done with a view to providing information on the comparability of the psychometrics qualities (discrimination and difficulty) of the examination items. Descriptive survey research design was adopted for the study. The population of the study consisted all public Senior Secondary school Physics student in Osun state. A sample size of 1020, SS3 Physics Students was selected using multi-stage sampling technique. From each of the three Senatorial Districts in Osun State, two Local Government Areas (LGAs) were selected using simple random sampling technique. Three schools were selected from each Local Government Area using simple random sampling technique, making a total of 18 schools. The research instruments for the study were WAEC and NECO Physics objective tests. These instruments were the adopted versions of 2016 WAEC and NECO Physics objective tests. These instruments were administered on the SS3 Physics Students who enrolled for 2019 WAEC. The data were analysed using chi square, mean and standard deviation.

The results showed that the average difficulty levels of NECO and WAEC Physics objective items were 2.11 and 1.25 respectively. Also, the results showed that the average discrimination levels of NECO and WAEC Physics tests items were 3.43 and 2.37 respectively. The results equally showed that the average vulnerability to guessing of 2016 NECO and WAEC Physics test items were 0.13 and 0.16 respectively.

The study concluded that the item parameters of WAEC and NECO Physics objective tests were statistically comparable.

Keywords: WAEC, NECO, Item Response Theory (IRT) and Item Parameters

INTRODUCTION

Psychometrics is the field of study that is concerned with theories and methods of measurement of knowledge, abilities, attitudes, personal traits, academic achievement and educational attainment. There is need for psychometrics as a way to solve measurement problems by practitioners in the process of test development and identification of biased items.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

Test theories enable the prediction of outcomes of tests by identifying parameters of item difficulty, item discrimination and the ability of the test takers. The main concern of test developers when constructing a test is the nature and quality of test items and how examinees respond to these items (Awopeju and Afolabi, 2016). Item difficulty and item discrimination are characteristics that determine the validity and reliability of a test.

Most popular two statistical procedures currently available to psychometrics are Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory (CTT) has served measurement practitioners for several decades as the foundation of measurement theory whereby the conceptual groundwork, assumptions, and extension of the basic principles of CTT have given room for the development of some excellent psychometrically sound scales (Dibu, 2013). Despite the benefits of CTT, recent findings by measurement experts show that classical test theory is no longer giving the services expected due to some shortcomings in the theory for assessment. Among the shortcomings in CTT include: item statistics or indices (that is, item difficulty and discrimination) are examinee (sample) dependent because the values of the indices in question depend on the particular examinee samples in which the test was administered. This implies that item parameters are not invariant characteristics of the item, but take on values that depend on who tried the items). Therefore, the estimates of CTT are not generalizable across populations. As a result of the shortcomings of CTT aforementioned, IRT had gained ground in educational assessment (Steyer, 2014).

Baker (2001) describes IRT as a modern test theory designed to estimate item characteristic parameters and examinees' latent abilities. Ani (2014) states that IRT is a modeling procedure that describes the interaction between an examinee's test performance and the latent trait underlying the performance. Ugwoke (2015) maintains that IRT is essentially designed to measure single, specific latent traits, abilities, attributes that are not observable. Therefore, the theory assumes an interaction between a person's possession or level of a particular attribute, trait or ability and his or her response to a test item. The combination of methodological advances and increasingly powerful software has increased applicability and interest in the use of IRT.

Studies like Steyer (2014) showed that the primary interest in IRT is on the item-level information as opposed to test-level information in CTT. Therefore, through IRT ability estimates (θ), invariant estimates of item and person parameters are generated. This implies theoretically that IRT ability parameters (θ) are "item free" (i.e. would not change if different items were used) and the item difficulty statistics are person free (i.e. would not change if different persons were used) (Adegoke, 2013). IRT parameters have been classified into two broad components which include: the parameter relating to individual examinee and the parameter relating to each of the items of a given test. In the case of IRT, the ability score is denoted by the Greek letter (θ). Therefore, at each ability level, it is certain that an examinee with that ability gives a correct answer to the item. In IRT, P(θ) is denoted for the probability.

The difficulty of an item is known as b parameter which indicates how easy or difficult an item is. The more difficult an item is, the higher the ability level required by the item for examinee to answer it correctly. The discrimination (a) parameter shows how well an item distinguishes among examinees with different ability levels. A test item with positive discrimination index indicates that the high ability students have high probability of answering the item correctly and low ability students have low probability of answering the item correctly. The c parameter indicates that an examinee with low ability can still respond to an item correctly through guessing. For items with four options, an examinee has the probability of getting the item correctly 1 out of 4 that is, the probability of guessing correctly is about 25 %.

By way of design and operations in IRT, the following assumptions have been recognized; unidimensionality, local independence and item characteristic curve (ICC). Items are said to be unidimensional when they measure one trait and differ in their difficulty level. An item is locally independent if the probability of correctly answering that item is entirely determined by a student's ability and not by his or her responses to other items or other sources of unaccounted for. This implies that the probability of getting an item correctly should not be affected by the answer given to the other items in the test (Wallace, Chambers & Prather, 2018). Adegoke (2013) describes parameter 'a' as the discrimination parameter of item commonly known as item slope, while parameter 'b' is the difficulty parameter of item *k*nown as item location parameter, θ (Theta) which is the ability level of a particular examinee. The parameter *c* is the probability of getting the item correct by guessing alone. By

definition, the value of *c* does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. The parameter *c* has a theoretical range of $0 \leq c \leq 1.0$, but in practice, value above 0.35 is not considered acceptable, hence the range $0 \leq c \leq 0.35$ is usually adopted when the 3-parameter model is used. The difficulty parameter also known as location parameter, denoted by *b*, is defined as the point on the ability scale at which the probability of correct response to the item is 0.5. The theoretical range of the values of this parameter is $-\infty \leq b \leq +\infty$. However, typical values have the range $-3 \leq b \leq +3$. The item discrimination parameter also known as slope parameter, denoted by *a*, is the slope of the tangent line of the item characteristic curve at the point of the location parameter. Typical values of *a* range from 0.5 to 2.0. Although, according to Hambleton, Rogers and Swaminathan (1991) values ranging from 0-2 defined to represent items that have appropriate discrimination level.

In the Nigerian Senior Secondary Schools, Physics as a teaching subject is one of the core science subjects to national and technological development of the country. The subject is chosen as a matter of fact that there are few studies on Physics compared to Mathematics, Economics, and so on. According to Erinosho (2013), Physics is essential for understanding the complexities of modern technology as well as fundamental for technological advancement of a nation. To her, Physics has brought significant contributions to many of the inventions that are shaping modern day as well as many of the events being encountered in everyday life. The assessment of Physics for certification purposes at the Senior Secondary Schools' public examinations contains objective, essay and practical type test items. For the sake of this study, the objective format in Physics testing shall be addressed. Multiple choice test items are said to be fundamentally important in the assessment of students.

Han (2012) affirms that for several decades ago, multiple choice test items have remained the most popular choice items from classroom tests to standardized large-scale assessments. Among the reasons for its importance in educational testing by scholars include: easy administration of the test, consistent scoring of the test items, inexpensive to score (whether manually or using automated computer systems), they are adaptable to various levels of learning outcomes, reduce some of the burden of large classes, testing a broader sample of course content in a given amount of testing. This study therefore investigates WAEC (West African Examination Council) and NECO (National Examinations Council) Senior School Certificate Physics Examination items using Item Response Theory. The theory is in the best position in this study to reveal discrimination, difficulty and guessing properties of each item under study. This implies that the quality of Physics multiple choice test items by WAEC and NECO for Senior School Certificate Examination (SSCE) need to be examined.

The WAEC and NECO are two similar examination bodies conducting Senior School Certificate Examination (SSCE) with a common underlying vision of ensuring credibility and fairness in their assessment practices. Over the years, there are cases of poor performances of students in science subjects particularly Physics despite the effort of government at all levels to rekindle the educational system of the country. For many years, the blame of students' poor performances has been shifted to teachers and schools which include incapability to impart the right knowledge to students; much emphasis on theoretical aspect of Physics at the expense of practical; shortage of Physics teachers to schools; and lack or shortage of laboratory equipment in schools. Adeyemo, (2010) explained that the faults of parents have been highlighted which include poor habit to the education of their children; effect of broken home and forcing children to study sciences at all cost even when the children do not have the ability and among others. Furthermore, Achor, Ochonogor and Daikwo, (2011) blame students for poor performance which include students' poor study habit; uncared attitude to their studies and peer group influence. a well-structured assessment framework that incorporates a variety of cognitive levels can provide a more accurate representation of students' true abilities and knowledge retention (Stiggins, 2005). Apart from the factors afore-mentioned, the advent of new test theory known as Item Response Theory (IRT) to educational assessment, show that test items might be faulty thus affecting students' performances. In this study therefore, the researcher is interested in investigating the credibility of psychometric properties of test items constructed by the two examination bodies i.e. WAEC and NECO using IRT in order to ascertain what brings about variation in the performance of student who sit for WAEC and NECO Physics Examinations and to address the shortcomings of CTT as it is no longer giving the services needed in educational assessment..

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

The main objective of this study was to investigate the item parameter qualities (difficulty and discrimination) of both WAEC and NECO SSCE 2016 Physics objectives test items. The specific objectives of this study were to:

i. Determine the dimensionality of WAEC and NECO 2016 Physics test items

ii. determine difficulty level of items in WAEC and NECO 2016 Physics test items;

iii. investigate the discriminating power of WAEC and NECO 2016 Physics test items.

iv. ascertain the difference in guessing parameter of WAEC and NECO Physics test items using Three Parameter Logistic Model of Item Response Theory

Research Questions

With respect to the research objectives, the following research questions were asked:

i. What is the dimensionality of WAEC and NECO 2016 Physics objective test items?

ii. What are the difficulty levels of WAEC and NECO 2016 Physics objective test items?

iii. What are the discriminating indices of WAEC and NECO 2016 physics objective test items?

iv. What is the difference in guessing parameter of WAEC and NECO 2016 physics objective test items?

Research Hypothesis

Ho1: There is no significant difference in the item difficulty level of 2016 WAEC and NECO Physics objective tests.

Ho2: There is no significant difference in the discriminating power of 2016 WAEC and NECO Physics objective test items.

Ho3: There is no significant difference in guessing parameter of 2016 WAEC and NECO Physics objective test items.

METHODOLOGY

Research Design

The study adopted the descriptive survey research design. The population of the study comprised all the public Senior Secondary Schools who enrolled for May/June 2019 Physics SSCE in the three senatorial districts of Osun State. The population of science students was 11,115. The sample size consisted of 1,020 students selected from Senior Secondary Schools in Osun State using multistage sampling procedure. From each of the senatorial district in Osun State, two Local Government Areas (LGAs) were selected by simple random sampling technique, and from each LGA, three schools were selected using simple random sampling technique to make a total of 18 schools. From each school, all the SS 3 Physics students were purposively selected, resulting in 1,020 students.

Instrument

The instruments for the study were WAEC and NECO Physics objective items (Paper II). The instruments were the adopted version of 2016 WAEC and NECO physics objective test items. The WAEC Physics objective test

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

items consisted of 50 multiple choice items with four options A-D. From the four options, one was the key (i.e. the correct answer) and the remaining options were distractors (i.e. incorrect answers). The NECO Physics objective items consisted of 60 multiple choice test items with five options A-E. From the five options, one was the key i.e. correct answer, and the remaining ones were distracters. They are two instruments titled Physics Achievement Test Type I and 2 respectively. The tests were conducted under strict examination condition and administered twice with an interval of two weeks.

Data Analysis

The responses of the examinees to the 2016 WAEC and NECO SSCE Physics tests were respectively subjected to IRT Software which is Normal Ogive Harmonic Analysis Robust Method Package (NOHARM). For research questions one and two, students' responses were subjected to test calibration application of multidimensional item response theory (MIRT), mean and standard deviation. Research hypotheses one and two were tested by Mann-Whitney U test.

RESULTS

Research Question one

To answer research questions two to three, there is need for model-data fit assessment. This is because the value of each of the three parameters of IRT model will be compared to determine the model that produces the best fit to the test data.

Model-data Fit Assessment

Assessment of model-data fit involves two stages: 1. fitting the data test to the three available IRT models and thereafter, the fitness of the three models to the data set are compared. The model that produced the best fit to the data is adjudged the model that fit the data. To achieve this fit, several measures are applied. According to French (2015), prominent among the measures 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. Among the most popular of these indices are the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978), and the sample-size-adjusted BIC (SBIC). These information indices are computed using the -2log likelihood chi-square value and is interpreted such that the model with the lower value exhibits a better fit to the data. In addition, the chi-square and likelihood ratio goodness of fit was used to tests the null hypothesis that three parameter models provide the same fit to a set of data. A statistically significant likelihood indicates a difference when 1PLM and 2 PLM are compared. Also, a significant likelihood indicates a difference when 2PLM and 3PLM of item response theory are compared. Table 1 presents the result of the model-data fit assessment.

Table 1: Model-data fit assessment of 2016 NECO and WAEC Physics tests

NECO							
IRT model	AIC	SABIC	BIC	LogLik	X^2	df	P
Comparing M1PL and M2PL models							

M1PL	65812.62	65921.21	66118.12	-32844.31	2219.655	117	0.000
M2PL	63826.96	64140.47	64708.99	-31734.48			
Comparing M2PL and M3PL models							
M2PL	63826.96	64140.47	64708.99	-31734.48	395.681	60	0.000
M3PL	63551.28	63969.88	64705.97	-31536.64			

Model-data fit assessment of 2016 WAEC Physics tests

WAEC

IRT

Model	AIC	SABIC	BIC	LogLik	X^2	df	P
Comparing M1PL and M2PL models							
M1PL	59825.61	59916.69	60081.85	-29860.81	2524.262	97	0.000
M2PL	57495.35	57756.32	58229.56	-28598.68			
Comparing M2PL and M3PL models							
M2PL	57495.35	57756.32	58229.56	-28598.68	408.73	50	0.000
M3PL	57186.62	57535.16	58167.21	-28394.31			

Table 1 presented the model-data fit assessment, showing the IRT model that is best for the calibration of the NECO and WAEC Physics tests. For the NECO Physics test, the table showed that when the fitness of M1PL and M2PL models to the data were compared, the result showed that the M2PL had AIC = 63826.96, SABIC = 64140.47, BIC = 64708.99 values that were lesser than the AIC = 65812.62, SABIC = 65921.21, BIC = 66118.12 values of the M1PL. In addition, the Likelihood ratio test that M2PL fitted the data better than M1PL was statistically significant χ^2 (117) = 395.681, $p < 0.05$. The results showed that the M2PL model fitted the data better than the M1PL model.

In search for a better model for the test data, the fitness of M2PL model to the Physics data was in turn compared to the fitness of M3PL model to the test data. The result showed that the M3PL model fitted the data better than the M2PL model. M3PL model's AIC = 63551.28, SABIC = 63969.88, BIC = 64705.97 values were respectively lesser than the M2PL model's AIC = 63826.96, SABIC = 64140.47, BIC = 64708.99; the Likelihood ratio test that M3PL model fitted the data better than M2PL model was statistically significant, χ^2 (60) = 395.681, p <

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

0.05. The result revealed that multidimensional three-parameter logistic model fitted the 2016 NECO Physics test. Thus, the test was calibrated using multidimensional three-parameter logistic model.

Furthermore, for the 2016 WAEC Physics test, Table 1 showed that when the fitness of M1PL and M2PL models to the data were compared, the result showed that the M2PL had AIC = 57495.35, SABIC = 57756.32, BIC = 58229.56 values that were lesser than the AIC = 59825.61, SABIC = 59916.69, BIC = 60081.85 values of the M1PL. In addition, the Likelihood ratio test that M2PL fitted the data better than M1PL was statistically significant χ^2 (97) = 2524.262, $p < 0.05$. These results showed that the M2PL model fitted the data better than the M1PL model. In search for a better model for the test data, the fitness of M2PL model to the Physics data was in turn compared to the fitness of M3PL model to the test data. The result showed that the M3PL model fitted the data better than the M2PL model M3PL model's AIC = 57186.62, SABIC = 57535.16, BIC = 58167.21 values were respectively lesser than the M2PL model's AIC = 57495.35, SABIC = 57756.32, BIC = 58229.56; the Likelihood ratio test that M3PL model fitted the data better than M2PL model was statistically significant, χ^2 (50) = 408.73, $p < 0.05$. The result revealed that multidimensional three-parameter logistic model fitted the 2016 WAEC Physics test. Thus, the Physics test was calibrated using multidimensional three-parameter logistic model.

Research Question one: What is the dimensionality of WAEC and NECO 2016 Physics objective test items?

To answer this research question four, students' responses to 2016 WAEC and NECO Physics objective tests were subjected to Stouts Test of Essential Unidimensionality.

To test the assumption of unidimensionality of item respond theory of the test items, the responses of the examinees to the 2016 WAEC and NECO SSCE Physics tests were respectively subjected to Stout's Test of Essential Unidimensionality (STEU) using in DIMTEST 2.0 package, (Stout, 2005). Since unidimensionality of the items is violated, the responses of the examinees to test data were subjected to non-linear factor analysis implemented in Normal Ogive Harmonic Analysis Robust Method Package (NOHARM), (Fraser and MC Donald, 1988) for the extraction of the factors underlie the tests.

Table 2: Unidimensionality of 2016 WAEC and NECO Physics objective tests

	T				**Pvalue**		
NECO	1.7368				0.0412		
WAEC	4.5784				0.0000		

Table 2 presented the result of the assessment of unidimensionality of 2016 NECO and WAEC Physics. For the NECO test, the table showed that items that measured the secondary dimension, Assessment subtest (AT) (these items were empirically extracted with the HCA/CCPROX clustering procedure and DETECT statistics of DIMTEST package using the responses of 30% of the examinees that were randomly selected from the examinees who took the respective tests) were dimensionally distinct from the remaining items of the test ($t = 1.7368$ p-value < 0.05, one-tailed)); therefore, 2016 NECO Physics test was not unidimensional. The result showed that more than one dimension accounted for the variation observed in student's performance on the test. For the WAEC test, items that measured the secondary dimension, Assessment subtest (AT) were dimensionally distinct from the remaining items of the Physics test ($t = 4.5784$ p-value < 0.05, one-tailed)); therefore, 2016 WAEC Physics test was not unidimensional. This result showed that more than one dimension accounted for the variation observed in student's performance on the test. These results showed that the 2016 NECO and WAEC Physics tests were not unidimensional. The implication of the results is that the test items respectively measured more than one ability. Thus, the optimal number of dimensions underlie the test items were assessed.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

Table 3: The number of dimensions of the NECO and WAEC Physics Tests

NECO						
NO OF DIMEN	GFI	RMSR CRITERION	RMSR	DIFF IN RMSR	REDUCTION IN RMSR	PERCENTAGE REDUCTION
1	0.941594	0.031311	0.007075			
2	0.954945	0.031311	0.006214	0.000861	0.121703	12
3	0.958919	0.031311	0.005934	0.00028	0.045122	5
WAEC						
NO OF DIMEN	GFI	RMSR CRITERION	RMSR	DIFF IN RMSR	REDUCTION IN RMSR	PERCENTAGE REDUCTION
1	0.955812	0.031311	0.008698			
2	0.967792	0.031311	0.007426	0.001272	0.146252	15
3	0.971997	0.031311	0.006924	0.000502	0.067561	7

Table 3 shows the number of dimensions underlying the NECO and WAEC Physics tests. Column represents the dimensions hypothesized to underlie the test data. GFI on column 2 is the Tanaka's (1993) goodness-of-fit index (GFI). McDonald (1999) suggests that a GFI of 0.90 indicates an acceptable level of fit and a value of 0.95 indicates "good" fit; GFI = 1 indicates perfect fit. The column labelled RMSR CRITERION is bench mark for judging the overall fitness of a model. The RMSR is the square root of the average squared difference between the observed and predicted covariance. Therefore, small values of RMSR indicate good fit. According to Tate (2003), in situations where several dimension models have acceptable level of fit indices, the most parsimonious model is determined by calculating the percentage reduction in RSMR values of the models being compared. Tate (2003) stated that the highest dimensional model that still produces an approximately greater than or equal 10% decrease in the RMSR over the preceding model is the model that best fit the data set.

NECO

Table 3 shows that 1-dimension, 2-dimension and 3-dimension model fitted the data (RMSR for 1-dimension, 0.008698 was lesser than the criterion 0.031311, this same trend was observed for 2-dimension and 3-dimension respectively and GFI for the three dimensions hypothesized to underlie the test were greater than 0.90). To identify the optimal dimensions underlying, the test data, the fitness of the data to the three hypothesized dimension-model were compared. The table showed that from the first dimension to the second dimension RMSR value decreased by 12%. According to Tate (2003) criteria, this is a significant amount of reduction, showing 2 dimensions significantly fitted the data better than the 1 dimension. Furthermore, the table shows that when that of 3-dimension was hypothesized to underlie the data set, the percentage reduction in RMSR between the 2-dimension and 3-dimension model was approximately 5%. This value was less than the criterion, 10%. Consequently, the highest dimensional model that still produced an approximately 10% or greater percentage reduction in the RMSR over the preceding model is the 2-dimensional model. Thus, 2-dimensional model is the

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

most parsimonious model that fits well. This implies that performance of examinees on the NECO test depended on two dimensions or abilities.

WAEC

Table 3 also shows that 1-dimension, 2-dimension and 3-dimension model fitted the data (RMSR for 1-dimension, 0.007075 was lesser than the criterion 0.031311, this same trend was observed for 2-dimension and 3-dimension respectively and GFI for the three dimensions hypothesized to underlie the test were greater than 0.90). To identify the optimal dimensions that underlie the test data, the fitness of the data to the three hypothesized dimension-model were compared. The table showed that from the first dimension to the second dimension RMSR value decreased by 15%. According to Tate (2003) criteria, this is a significant amount of reduction, showing 2 dimensions significantly fitted the data better than the 1 dimension. Furthermore, the table shows that when that of 3-dimension was hypothesized to underlie the test data, the percentage reduction in RMSR between the 2-dimension and 3-dimension model was approximately 7%. This value was less than the criterion, 10%. Consequently, the highest dimensional model that still produced an approximately 10% or greater percentage reduction in the RMSR over the preceding model is the 2-dimensional model. Thus, 2-dimensional model is the most parsimonious model that fits well. This implies that performance of examinees on the WAEC test depended on two dimensions or abilities.

Research Question Two: What are the difficulty levels of WAEC and NECO 2016 Physics objective test items?

To answer this research question, student's responses to the test items were dichotomously scored. The students' scores were used to estimate the overall difficulty levels of WAEC and NECO Physics objective tests. The result is presented in Table 4:

Table 4: The Difficulty indices of WAEC and NECO 2016 Physics objective test items.

Item	MDIFF	b	Item	MDIFF	b	Item	MDIFF	b	Item	MDIFF	b
1	-2.06	0.71	28	-4.08	1.59	1	-0.52	0.46	28	-1.90	3.07
2	0.13	-0.12	29	-2.02	6.68	2	-2.18	0.84	29	-6.73	1.97
3	-0.81	0.44	30	-3.47	1.19	3	-1.39	0.82	30	-0.90	0.55
4	-0.12	0.08	31	-2.83	1.81	4	-2.48	1.61	31	-1.51	0.80
5	-3.29	1.57	32	-0.66	0.36	5	-0.91	0.26	32	0.34	-0.15
6	-0.43	0.20	33	-1.74	7.04	6	-7.67	1.89	33	-0.34	0.79
7	-4.31	1.23	34	-4.98	2.73	7	-0.68	0.28	34	-1.01	0.83
8	-2.07	1.00	35	-0.99	4.08	8	-9.86	1.75	35	-1.09	1.39
9	-23.06	1.91	36	-48.96	2.28	9	-1.50	2.44	36	-0.52	0.23
10	-24.19	1.81	37	-2.30	9.40	10	-3.73	1.39	37	-1.76	0.72

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

11	-5.73	1.87	38	-1.55	1.55	11	-1.69	3.76	38	-0.94	0.78
12	-3.01	1.31	39	-2.68	2.74	12	0.55	-0.54	39	0.17	-0.25
13	-2.17	2.44	40	-2.23	2.08	13	-0.10	0.04	40	-1.20	1.21
14	-1.79	1.06	41	-1.80	4.08	14	-5.55	1.25	41	-6.70	1.66
15	-2.90	1.12	42	-4.07	1.56	15	-1.94	0.90	42	-1.72	3.73
16	-4.29	1.94	43	-3.19	1.53	16	-3.71	1.51	43	-13.37	1.92
17	-1.55	4.21	44	-1.50	3.30	17	-0.93	0.48	44	-10.86	1.77
18	-9.13	1.59	45	-1.77	2.26	18	-4.83	1.31	45	-4.46	1.67
19	-0.66	2.37	46	-2.19	1.51	19	-4.25	1.92	46	0.18	-0.07
20	-2.13	4.58	47	-2.19	5.75	20	-3.21	1.05	47	-1.25	2.11
21	-1.20	1.04	48	-2.40	2.55	21	-0.95	0.53	48	-0.91	3.94
22	-2.51	1.33	49	-35.64	2.29	22	-3.06	1.21	49	-5.70	1.95
23	-3.69	0.93	50	-1.50	4.16	23	-0.66	0.57	50	-3.40	2.95
24	-8.31	1.72	51	-0.63	0.59	24	1.18	-0.59	Mean		1.25
25	-2.05	0.84	52	-56.66	2.19	25	-6.24	1.95	SD		1.066
26	0.12	-0.16	53	-1.44	0.94	26	0.21	-0.12			
27	-15.74	2.09	54	-4.97	1.86	27	-17.20	2.02			
IT55	-48.99	2.17									
IT56	-3.30	2.35									
IT57	-0.34	0.21									
IT58	-1.46	1.69									
IT59	-2.39	1.65									
IT60	-0.70	1.32									

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

Mean		2.11									
SD		1.765									

Table 4 shows the difficulty parameters of the 2016 NECO and WAEC Physics tests. From the table only four items of WAEC are outside the range of -3 to +3 defined according to Hambleton, Rogers and Swaminathan (1991) to represent items that have appropriate difficulty level. On the other hand, 11(18%) of NECO Physics items are outside the specified range of -3 to +3 indicating inappropriate difficulty. The columns' labeled b is the intercept used in estimating the multidimensional difficulty of the tests items. The overall difficulty called multidimensional item difficulty parameter is represented by: MDiff $= \frac{-d}{MDiff}$ the table showed that NECO items were more difficult ($\bar{X}$ = 2.11, SD = 1.765) than the WAEC items ($\bar{X}$ = 1.25, SD = 1.066). The results showed that the NECO test was more difficult than the WAEC test. The implication is that the examinees performed better on WAEC test than did on NECO test.

Research Question Three: What are the discrimination powers of 2016 WAEC and NECO Physics objective tests?

To answer this research question, statistical models of mean and standard deviation were used to determine the overall discrimination levels of WAEC and NECO 2016 Physics objective tests. The result is presented in Table 5

To determine the discrimination powers of 2016 WAEC and NECO Physics objective tests, statistical models of mean and standard deviation were used to determine the overall discrimination levels of WAEC and NECO 2016 Physics objective tests. The result is presented in Table 5. Table shows that 21 (35%) NECO physics items were outside the range of 0-2 defined according to Hambleton, Rogers and Swaminathan (1991) to represent items that have inappropriate discrimination level. Twenty-seven (27%) WAEC items were outside the specified range

Table 5: Discrimination parameters of 2016 NECO and WAEC Physics tests items

Item	**X**	**Y**	**Z**	**Item**	**X**	**Y**	**Z**	**Item**	**X**	**Y**	**Z**
1	-2.90	-0.41	2.92	28	-2.53	0.41	2.56	55	2.16	22.52	22.62
2	-1.07	0.24	1.10	29	0.14	0.27	0.30	56	-0.61	-1.26	1.40
3	-1.81	0.09	1.82	30	-2.93	-0.08	2.93	57	-1.46	0.70	1.62
4	-1.60	0.14	1.61	31	-1.57	-0.03	1.57	58	0.85	0.16	0.86
5	-2.06	-0.37	2.09	32	-1.69	0.69	1.82	59	-1.43	-0.19	1.45
6	-2.06	0.45	2.11	33	0.16	0.19	0.25	60	0.43	0.31	0.53
7	-3.48	-0.37	3.50	34	0.55	-1.74	1.82	Mean			3.43
8	-0.18	2.05	2.06	35	0.24	0.06	0.24	SD			5.488

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

9	4.94	10.99	12.05	36	-0.05	21.52	21.52				
10	-8.12	-10.61	13.36	37	0.23	-0.07	0.24				
11	-3.03	-0.44	3.06	38	-0.83	0.57	1.00				
12	-2.30	0.01	2.30	39	-0.31	-0.93	0.98				
13	0.65	0.61	0.89	40	-0.94	-0.51	1.07				
14	-0.75	1.51	1.68	41	0.38	-0.23	0.44				
15	-2.59	-0.09	2.59	42	-2.54	0.57	2.60				
16	-2.20	-0.20	2.21	43	-1.58	1.35	2.08				
17	0.08	0.36	0.37	44	-0.45	0.06	0.45				
18	-3.99	4.13	5.74	45	0.74	-0.24	0.78				
19	0.01	0.28	0.28	46	-1.43	-0.25	1.45				
20	0.29	-0.37	0.47	47	0.34	0.17	0.38				
21	-1.11	-0.31	1.15	48	0.84	-0.43	0.94				
22	-1.89	-0.11	1.89	49	0.74	-15.54	15.56				
23	-3.96	0.01	3.96	50	-0.36	0.04	0.36				
24	-4.80	0.56	4.83	51	-1.00	0.36	1.06				
25	-2.39	0.44	2.43	52	-22.53	-12.71	25.86				
26	0.00	0.70	0.70	53	-0.53	1.44	1.53				
27	-7.49	-0.94	7.55	54	-2.40	-1.18	2.67				

Second Table

Item	**X**	**Y**	**Z**	**Item**	**X**	**Y**	**Z**
1	-1.04	-0.40	1.12	26	-1.71	0.40	1.76

2	-2.57	-0.43	2.60	27	-8.06	-2.76	8.52
3	-1.68	-0.31	1.71	28	-0.61	-0.12	0.62
4	-1.51	-0.30	1.54	29	-0.51	-3.38	3.42
5	-3.43	0.75	3.51	30	-1.26	1.08	1.66
6	-3.26	-2.42	4.06	31	-1.81	-0.50	1.88
7	-2.35	0.46	2.39	32	-2.27	0.27	2.28
8	-0.57	-5.61	5.64	33	-0.01	0.43	0.43
9	0.53	0.31	0.61	34	0.44	-1.13	1.21
10	-2.12	-1.64	2.68	35	-0.75	-0.22	0.78
11	0.30	0.33	0.45	36	-2.23	0.23	2.24
12	-0.99	-0.23	1.02	37	-2.46	-0.05	2.46
13	-2.88	0.49	2.92	38	-1.18	-0.21	1.20
14	1.18	-4.27	4.43	39	-0.63	0.22	0.67
15	-0.63	-2.07	2.16	40	-0.86	-0.49	0.99
16	-1.41	-2.00	2.45	41	-3.34	-2.28	4.04
17	-1.92	-0.29	1.94	42	0.46	0.01	0.46
18	-3.58	0.87	3.68	43	-4.66	-5.17	6.96
19	-2.16	-0.48	2.21	44	-5.56	-2.61	6.14
20	-2.99	-0.57	3.05	45	-2.30	-1.36	2.67
21	-1.75	-0.26	1.77	46	-2.35	0.46	2.39
22	-2.41	-0.76	2.52	47	0.43	-0.40	0.59
23	-1.02	0.53	1.15	48	0.05	0.22	0.23
24	-2.01	0.13	2.01	49	-0.90	-2.78	2.92

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

25	-0.52	-3.16	3.20	50	1.14	0.13	1.15
Mean			2.37				
SD			1.716				

Table 5 shows the discrimination parameters of the 2016 NECO and WAEC Physics test items. The columns labelled a1 and a2 represent the discrimination parameter of the items at dimension 1 and 2 respectively. The column labelled MDISC is the overall discrimination of the items which is calculated by:

$MDISC_i = (\sum_{k=1}^{m} a_{ik}^2)^{1/2}$).

From Table 5, the result showed that NECO items discriminates more ($\bar{X}$ = 3.43, SD = 5.488) than the WAEC items ($\bar{X}$ = 2.37, SD = 1.716). The result showed that the NECO test items were better than the WAEC items. The implication is that the NECO items were able to distinguish more between high and low examinees than WAEC items.

Research Question four: What is the difference in guessing parameter of WAEC and NECO Physics objective test items?

To answer this research question, students' responses were subjected to test calibration application of multidimensional item response theory. To determine the difference in guessing parameter of WAEC and NECO Physics objective tests, students' responses to the two tests were compared. The result is presented in Table 6

Table 6: The Guessing Parameters of WAEC and NECO Physics objective test items

Item	NECO	Item	NECO	Item	NECO	Item	WAEC	Item	WAEC
1	0.21	28	0.13	55	0.23	1	0.09	28	0.06
2	0.51	29	0	56	0.16	2	0.31	29	0.2
3	0.43	30	0.13	57	0.13	3	0.13	30	0.12
4	0.02	31	0.16	58	0	4	0.22	31	0.1
5	0.1	32	0.17	59	0.09	5	0.29	32	0.18
6	0.32	33	0	60	0	6	0.16	33	0
7	0.17	34	0.1	Mean	0.13	7	0.23	34	0.2
8	0.35	35	0	SD	0.11	8	0.25	35	0
9	0.11	36	0.26			9	0	36	0.15

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

10	0.21	37	0			10	0.16	37	0.24
11	0.2	38	0.1			11	0	38	0.13
12	0.18	39	0.07			12	0.07	39	0
13	0.11	40	0			13	0.26	40	0.1
14	0.13	41	0			14	0.06	41	0.13
15	0.31	42	0.17			15	0.36	42	0
16	0.08	43	0.27			16	0.29	43	0.35
17	0	44	0			17	0.33	44	0.22
18	0.19	45	0			18	0.06	45	0.27
19	0	46	0.14			19	0.06	46	0.19
20	0	47	0			20	0.16	47	0
21	0.04	48	0.08			21	0.23	48	0
22	0.18	49	0.13			22	0.23	49	0.19
23	0.12	50	0			23	0	50	0.12
24	0.11	51	0.08			24	0.22	Mean	0.16
25	0.23	52	0.17			25	0.24	SD	0.11
26	0	53	0.15			26	0.22		
27	0.19	54	0.14			27	0.26		

Table 6 shows that 2016 WAEC Physics test items are more vulnerable to guessing ($\bar{X}$ = 0.16, SD = 0.11) than 2016 NECO Physics test items ($\bar{X}$ = 0.13, SD = 0.11). This result showed that the vulnerability to guessing of 2016 NECO Physics test items were better than the vulnerability to guessing of WAEC Physics test items. The implication of the result is that examinees are more likely to guess correctly the answer to the WAEC test than the NECO test items.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

Research Hypotheses

Research Hypothesis One: There is no significant different in the difficulty level of 2016 WAEC and NECO Physics objective test items

To test whether the difference observed in the difficulty parameters (MDIFF) of the NECO and WAEC test items presented in Table 6 was significant, Mann-Whitney U test was conducted. The result is presented in Figure 1 and Table 7.

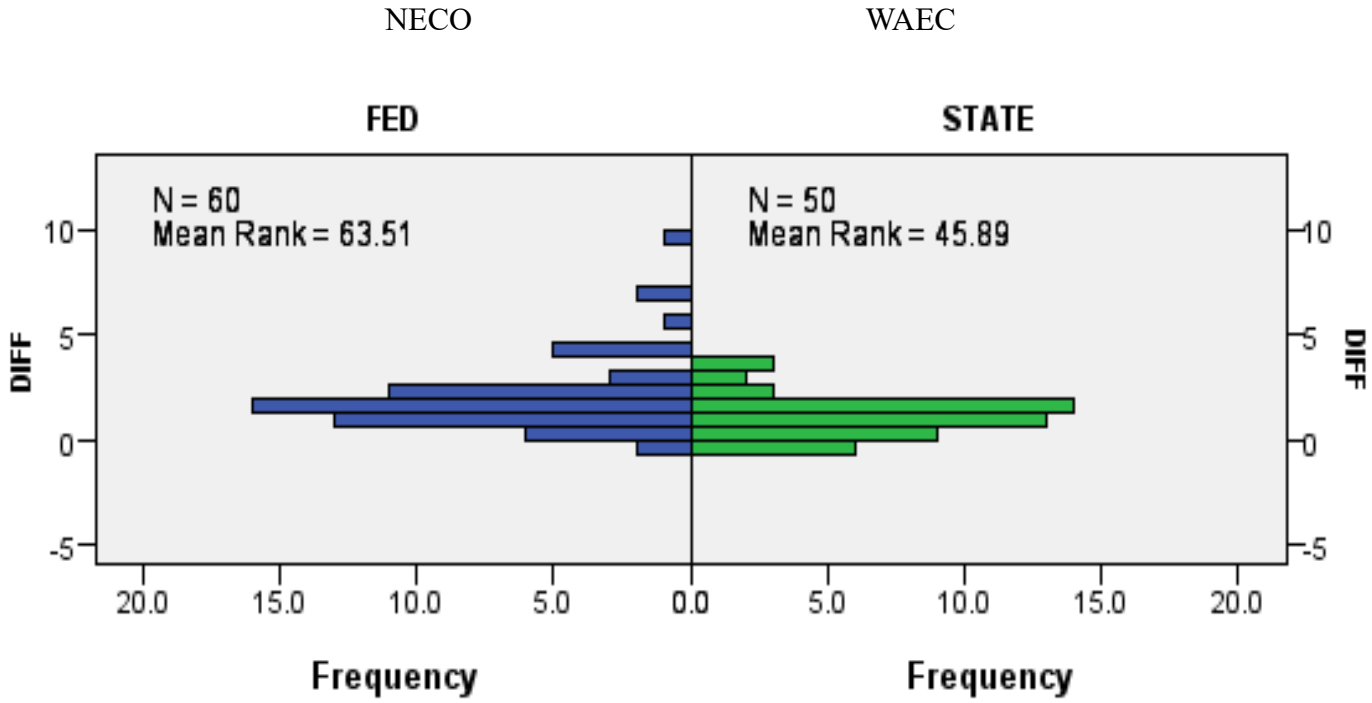

Figure 1: Distribution of Difficulty Indices of 2016 WAEC and NECO test items

Table 7: Mann-Whitney U test of the WAEC and NECO items difficulty parameters.

Null Hypothesis	U test	Sig	Decision
The distribution of the difficulty of NECO and WAEC test items are the same	2.885	0.04	reject the null hypothesis

significant at 0.05

Table 7 shows that the distribution of the difficulty parameters of NECO and WAEC tests presented in Figure 3 were significantly different from one another ($U = 2.885$, $p < 0.05$). Hence the Hypothesis "There is no significant difference in the item difficulty level of 2016 WAEC and NECO Physics objective test items" under which the test was conducted was rejected. This result showed that the difficulty level of NECO items were significantly different from the difficulty level of WAEC Physics test items. The implication of the result is that the 2016 WAEC and NECO Physics test items showed different level of difficulty among the examinees.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

Research Hypothesis Two

There is no significant difference in the item discrimination level of 2016 WAEC and NECO Physics objective test items in Osun State.

To test whether the difference observed in the discrimination parameters (MDISC) of the NECO and WAEC test items presented in Table 8 was significant, Mann-Whitney U test was conducted. The result is presented in Figure 2 and Table 7.

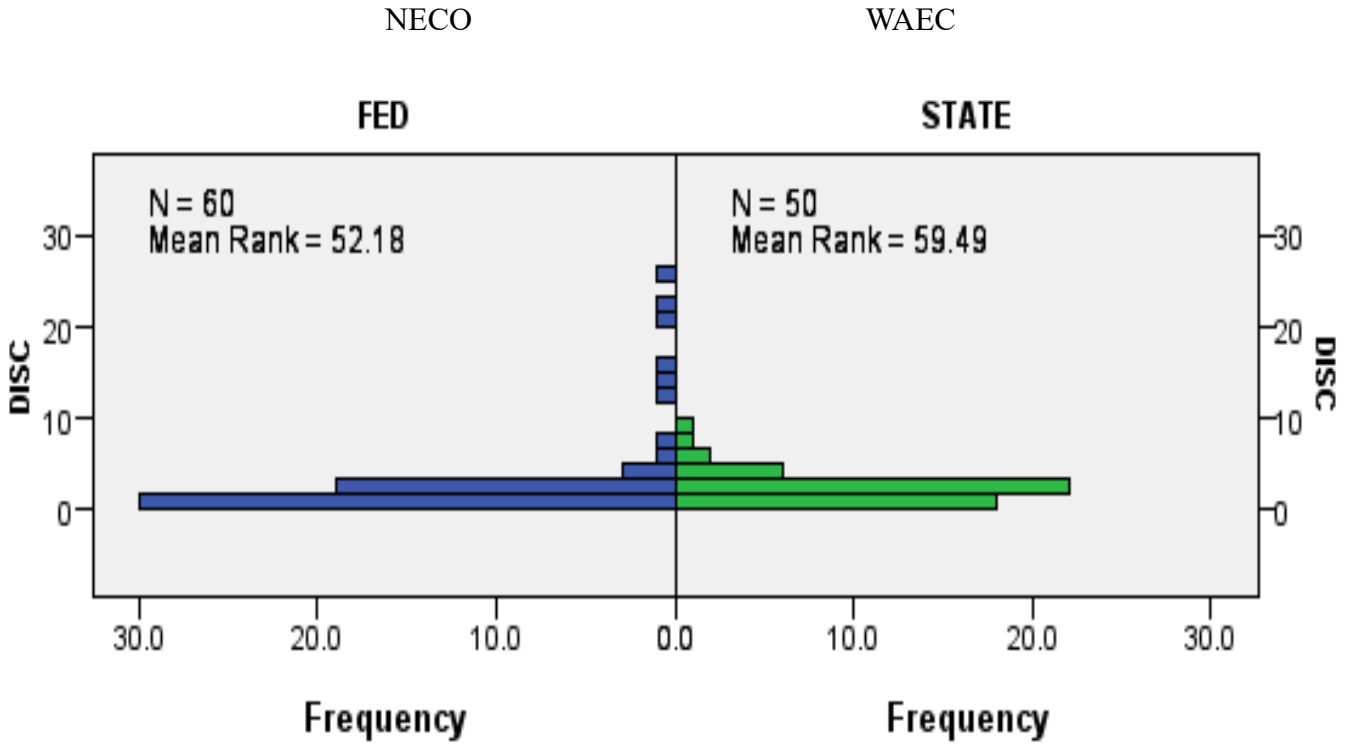

Figure 2: Distribution of Discrimination indices of 2016 WAEC and NECO test items

Table 8: Mann-Whitney U test of 2016 WAEC and NECO Physics items discrimination parameters.

Null Hypothesis	U test	Sig	Decision
The distribution of the discrimination of NECO and WAEC test items are the same	1.198	0.231	Do not reject the null hypothesis

not significant at 0.05

Table 8 shows that the distribution of the discrimination indices of NECO and WAEC tests presented in Figure 2 were not different from one another significantly (U = 1.198, p > 0.05). Hence the Hypothesis "There is no significant difference in the item discrimination parameters of 2016 WAEC and NECO Physics objective test items" under which the test was conducted was not rejected. This result showed that the discrimination indices of WAEC and NECO Physics test items were not significantly different from one another. The implication of the

result is that the WAEC and NECO Physics test items showed almost equal level of discrimination power among the examinees.

Research Hypothesis Three: There is no significant difference in guessing parameter of 2016 WAEC and NECO Physics objective test items.

To test whether the difference observed in the guessing parameters of the NECO and WAEC test items presented in Table 8 was significant, Mann-Whitney U test was conducted. The result is presented in Figure 4 and Table 9

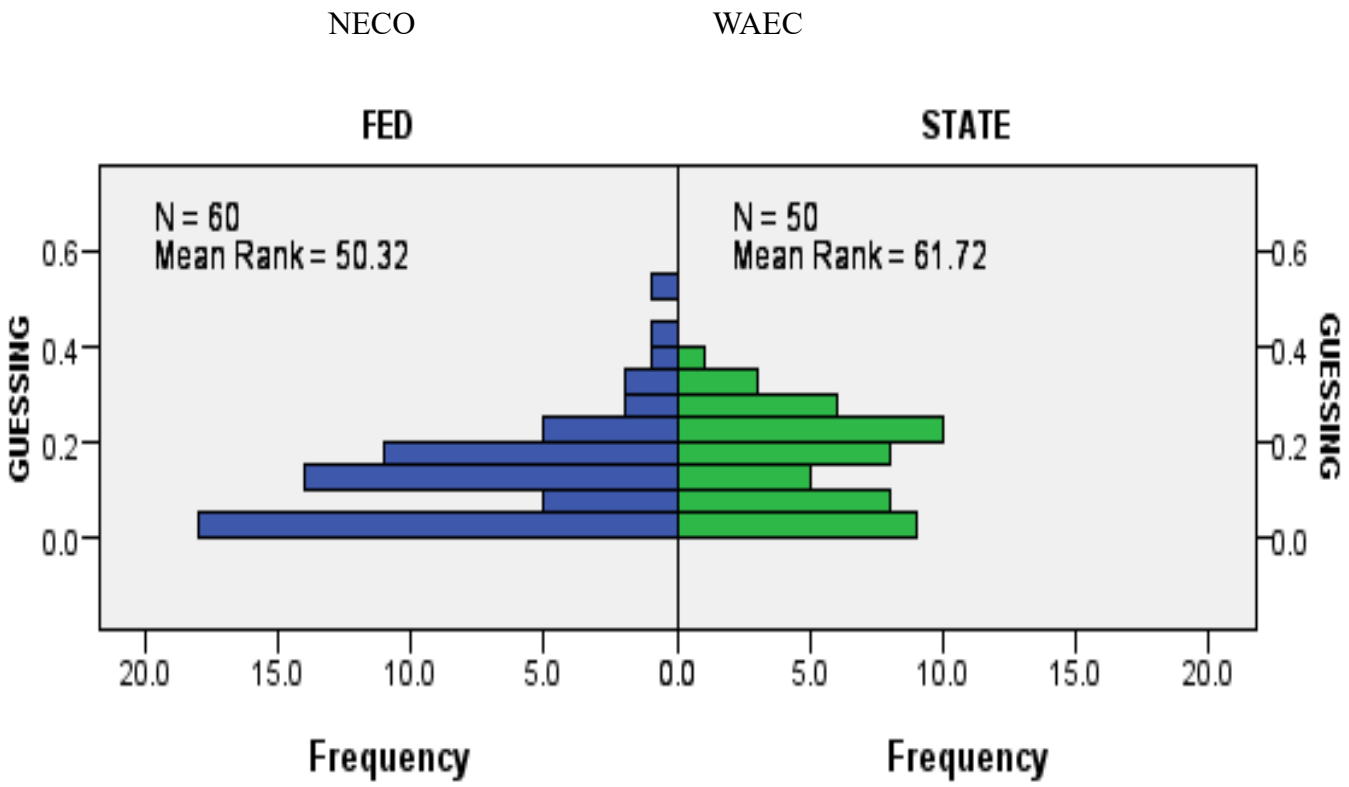

Figure 4: Distribution of guessing parameters of 2016 WAEC and NECO Physics test items

Table 9: Mann-Whitney U test of 2016 WAEC and NECO Physics items guessing parameters

Null Hypothesis	U test	Sig	Decision
The distribution of the guessing of NECO and WAEC test items are the same	1.876	0.061	Do not reject the null hypothesis
not significant at 0.05			

Table 9 shows that the distribution of the guessing parameters of NECO and WAEC test items presented in Figure 4 were not different from one another significantly ($U = 1.876$, $p > 0.05$). Hence the Hypothesis "There is no significant difference in the guessing parameters of 2016 WAEC and NECO Physics objective tests under which the test was conducted was not rejected. This result showed that the guessing indices of WAEC and NECO Physics test items were not significantly different from one another. The implication of the result is that the WAEC and NECO Physics test items showed almost equal vulnerability to guessing among the examinees. The finding of the research question four established the difference in guessing parameters of 2016 WAEC and NECO Physics test items. The average of WAEC and NECO Physics objective items were 0.16 and 0.13 respectively. In WAEC, only item 15(0.36) exceeded the acceptable level of guessing under item response theory. While on

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

NECO, items 2 and 3 (0.33) exceeded the acceptable level of guessing. In practice, value above 0.35 is not considered acceptable in IRT approach.

DISCUSSION

Assessment of IRT model fit to item response data is an important step to be taken before an IRT model can be applied with confidence to estimate ability level of examinees (Stone & Zhang, 2003). The results of this study have given insight into the comparison of WAEC and NECO Physics 2016 items using IRT. As earlier mentioned, the fitness of the IRT model to the test data was first determined. The best fit was produced from the Three Parameter Logistic Model (3PLM) of (IRT) produced the best fit compared to 1 and 2 Parameter Logistic Models. The analysis also showed that WAEC and NECO Physics 2016 items violated the assumption of unidimensionality meaning that the tests items measured more than one ability. This result is in line with Li, Jiao and Lissitz (2012) report that unidimensionality may be violated when multiple content areas exist in a single test. This finding is also in line with the findings of Lebeau and Liu (2017) report that when there is evidence that the unidimensionality assumption of IRT is not met, there is evidence that more than one ability distribution is present for each individual representing more than one construct. The findings of the difficulty levels of 2016 WAEC and NECO Physics objective tests items having the average of 1.25 and 2.11 respectively. This is contrary to the findings of Babatimehin (2021) study on the determination of the equivalence of both examinations in Chemistry where WAEC and NECO items had averages of 1.08 and 0.70 respectively. Six items of WAEC, (12%) were easy, 40 (80%) items were moderate, and 4 items (8 %) were difficult. While, 2(3.3%) of NECO easy items were found, 48 (80%) items were moderate and 10 (16.7%) items were difficult. The results of the analysis of hypothesis one of this study showed that the difficulty level of 2016 WAEC Physics test items were significantly different from the difficulty level of NECO Physics objective items in Osun State Senior Secondary Schools. This finding is also not in tandem with the finding of Babatimehin (2021), where both difficulty and discrimination levels of WAEC and NECO items are not significantly different. This finding is however similar to Olutola (2015) finding that the mean difficulty index of WAEC items is slightly higher than NECO. The reasons for the results on the discrimination levels may be adduced to the fact that the gap between the time of writing the two examinations is close. While the difference in the difficulty levels of the two examination items may be due to the areas covered (content coverage) in by the items.

The finding of the research question two confirmed the discrimination parameters of 2016 WAEC and NECO Physics objective tests. The average of discriminating power of WAEC and NECO were 2.37 and 3.43 respectively. Table 3 showed that 21 (42%) of the WAEC Physics items discriminated appropriately and 29 (58%) of the items discriminated inappropriate. While on NECO, 26 items (43.3%) discriminated appropriately and 34 items (56.7%) discriminated inappropriate. The results showed that the NECO items were able to discriminate more than the WAEC Physics test. The results of the analysis of the hypothesis two showed that the discrimination indices of WAEC and NECO Physics test items were not significantly different from one another. The two tests showed almost equal level of discrimination power among the students. This implies that the set of students who failed NECO are supposed to fail WAEC and vice-versa. Thus, confirming that the two tests are equivalent. The results of the analysis of hypothesis three showed that there was no significant difference in the guessing levels of the 2016 WAEC and NECO Physics test. The implication of the result is that WAEC and NECO Physics test items showed almost equal vulnerability to guessing among the examinees. Akindele (2003) also noted that items do not have perfect c-values because examinees do not guess randomly when they do not know the answer. Obinne (2008) established that WAEC items were more susceptible to guessing than NECO items.

CONCLUSION

Considering the findings of this study on the discrimination and difficulty levels of 2016 WAEC and NECO Physics items, it is therefore concluded that the psychometric properties of the two examinations in Physics objective tests are statistically comparable.

Limitation of the Study

Although, the WAEC and NECO 2016 Physics items were administered on 2019 final year students who were qualified and registered to write May/June and June/July WAEC and NECO examinations respectively having completed the syllabi for the two examinations. This is in no way a threat to the validity of the study since the emphasis is on the syllabi and the items were administered to the candidates who are qualified to write the examinations.

REFERENCES

1. Achor, E. E., Ochonogor, C. E., & Daikwo, S. A. (2011). Relative abstract nature of the three core science subjects at the senior secondary level in Nigeria as exemplified by classroom interaction patterns. Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 5(1), 152–162.
2. Adegoke, B. A. (2013). Comparison of Items Statistics of Physics Achievement Test Using Classical Test and Item Response Theory Frameworks. Journal of Education and Practice, 4(22), 8331-8667.
3. Adeyemo, S. (2010). Teaching and learning of Physics in Nigerian secondary schools:The curriculum transformation,issues,problems and prospects. International journal of Educational Research and Technology, 1(1), 99-111.
4. Akindele, O. A. (2003). Fundamentals of test and measurement. Lagos: Crown Publishers.
5. Ani, E. (2014). Application of Item Response Theory in the Development and Validation of Multiple Choice Test in Economics. Msc Thesis,University of Nigeria Nsukka.
6. Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. European Scientific Journal, 12(28), 263–284. https://doi.org/10.19044/esj.2016.v12n28p263
7. Babatimehin, T. (2021). Determination of the Equivalence of WAEC and NECO SSCE Chemistry Items Using Linear Test Equating Approaches of Classical Test Theory and Item Response Theory. Bulgarian Journal of Science and Education Policy (BJSEP), 15, 187-207.
8. Baker, B. (2001). The basis of Item Response Theory in United States of America (2nd ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.
9. Dibu, O. O. (2013). Classical Test Theory (CTT) VS Item Response Theory (IRT) and Evaluation of the Comparability of Item Analysis Results. Ibadan: Institute of Education University of Ibadan.
10. Erinosho, S. (2013). How do students perceive the difficulty of physics in Secondary School? An exploratory study in Nigeria. International Journal for Cross-Disciplinary Subjects in Education (IJCDSE), 3(3).
11. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.
12. Han, K. (2012). Fixing the c Parameter in the Three- Parameter Logistic Model. Practical Assessment, Research, and Evaluation, 17.
13. Li, H., Jiao, H., & Lissitz, R. W. (2012). Exploring the impact of item response model and ability estimation methods on test score equating. Educational and Psychological Measurement, 72(4), 590–607.
14. Lebeau, B., & Liu, O. L. (2017). Effects of test mode on student performance and test-taking behavior: Paper-based vs. computer-based testing. Educational Measurement: Issues and Practice, 36(3), 38–46.
15. Obinne, A. E. (2008). Comparison of Psychometric Properties of WAEC and NECO Test Item under Item Response Theory. Unpublished PhD Thesis University of Nigeria Nssuka.
16. Olutola, A. T. (2015). Empirical Analysis of Item Difficulty and Discrimination indices of Senior School Certificate Multiple Choice Biology Tests in Nigeria. A paper presented at the 41st annual

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

conference of International Association of Educational Assessment (IAEA) held on 11th-15th October, 2015 at University of Kansa, Lawrence, Kansas, USA.

17. Reeve, R., & Fayers. (2005). Applying Item Response Theory Modelling for Evaluating Questionnaire Items and Scale Properties (2nd ed.) Oxford University Press, Oxford.
18. Steyer, R. (2014). Classical (Psychometric) Test Theory. USA: international Encyclopedia of the Social and behavioral Sciences.
19. Ugwoke, E. O. (2015). Effectiveness Utilization of ICT for Repositioning Business Education Programme in Tertiary Institutions in Nigeria for National Development. International Journal of Education Research, 11(1), 201-214.
20. Wallace, C., Prather, E., & Duncan, D. (2012). A Study on General Education Astronomy Students Understanding of Cosmology: Evaluating Conceptual Cosmology Survey. An Item Response Approach. The American Astronomical society.
21. Steyer, R. (2014). Test theory: A unified treatment. Psychology Press.
22. Stout, W. (2005). DIMTEST 2.0. London: Oxford University Press.