Investigating Unidimentionality and Item Parameters of 2015/2016 Biology Multiple-Choice Items of Osun State Joint Promotion Examination (OSJPE)

[1]Abdulwasiu Adeyemi RASHEED

[2]Bamidele Abiodun FALEYE
abdulwasiu05@gmail.com

[1, 2]Department of Educational Foundations and Counselling,
Faculty of Education, Obafemi Awolowo University,
Ile-Ife, Nigeria

Abstract

The study investigated the unidimensionality of 2015/2016 Biology Multiple-choice Items of Osun State Joint Promotion Examination (OSJPE). It also estimated the item parameters (difficulty, discrimination, and guessing indices) of the Multiple-choice questionsunder the 3-parameter model of Item Response Theory (IRT). Ex-post-facto research design was adopted. The population used in this study comprised students' responses to the 40 Biology questions in the OSJPE for 2016. The sample size consisted of an intact class of 2000 students' responses in Biology as contained in the Optical Mark Recorder(OMR). Biology questions and students' response sheets constituted the instruments for the study. Data collected were calibrated to establish item difficulty, discrimination, and guessing indices using X-Calibre 4.2.The data was also subjected to factor analysis to establish the unidimensionality of the items. Results showed that the assumption of unidimensionality was satisfied in the 2015/2016 Biology items of OSJPE, since there was one dominating factor (first factor with the eigenvalue of 5.78 that explained 14.45% of the total test variance) among the items set using factor analysis. Twenty-five (25) (62.5%) of the 40 multiple-choice Biology items were considered to have defects of high difficulty level, low discriminating ability, and high guessing tendency under the 3-Parameter Logistic model (3PL). The study concluded that the items were not stable as far as item difficulty, item discrimination, and guessing indices are concerned using IRT Frameworks. It is then recommended that IRT should be maintained in construction of Biology test items that will be gender and location balance, because of its position in the investigation of reliability and in minimizing measurement errors.

Keywords*:* Unidimentionality, item parameters, item response theory, test validation, OSJPE

Introduction

Osun State Joint Promotion Examination is an examination introduced by the Osun State Government to prepare students for better performance in any of the external standardized examinations. It is also for the state government to pay West African Examination Council

(WAEC) fees of those who pass the OSJPE. It is very regrettable to note that despite the huge amount of money expended by the state government to pay students' WAEC fees, the outcome of students' results from WAEC showed that 60% of students failed Biology as it is evident in MAY/JUNE 2017 results. This implied that Biology items of 2016 OSJPE did not function well to predict the students' outcome in 2017 WAEC. This made some people to base their critics of the poor performance in the subject to non-validation of Biology examination items of the OSJPE.

Test validation is a process of identifying the extent to which test items perform its expected function. According to Rezaee and Salehi (2008), test validation is an essential process and it becomes more important when it comes to validating a high-stakes test. Angoff (2007) claims that neither a test nor even the scores produced by the test are validated; rather, the interpretations and inferences that the user draws from the test scores, and the decisions and actions that flow from those inferences are to be validated. Zumbo (2011) also notes that it is not the measure that is being validated; rather the inferences made from a measure must be validated. Brown (2005) in the same line of argument points out that validity is not about the test itself so much as it is about the test when the scores are interpreted for some specific purposes. It is much more accurate to refer to the validity of the scores and interpretations that result from a test than to think of the test itself as being valid.

However, several factors are found to potentially affect the predictive validity of test items. These include factors that are capable of affecting reliability since reliability is an essential (but not sufficient) factor in ensuring validity. The factors are also the nature of the items, their psychometric properties such as discrimination, and distracter abilities of multiple-choice type as well as the unidimensionality of the items (Faleye & Benjamen, 2016). In test validation, areas of item performance such as item parameters (difficulty, discrimination, and guessing indices), which are used for item retention decisions, are examined. Other areas are unidimensionality of items, internal correlation and test bias or assessing the impact of gender, field of study, age, and background knowledge.

Unidimentionality and parameter invariance are parts of the assumptions of Item Response Theory (IRT). Unidimentionality is an ability of an examination to measure one character of an examinee. According to Dibu, Kunmi, Francis and Patrick (2012), unidimentionality as one of the assumptions of IRT is a phenomenon whereby a single ability is sufficient to explain or account for examinee's performance. It is assumed that a single unidimensional trait underlies the data collected for the estimation of person ability (?). In practice, the unidimentionality assumption is usually approximately met by various ability test such as verbal, numerical, spatial perception and mechanical tests. Items which test bits of knowledge that are learned together, and items that test bits of knowledge which are logically and sequentially related are likely to be unidimensional. Unidimensionality of a set of test items could be examined by factor analytic study (Lumsden, 1961). In Lumsden's method, a factor analysis is performed and items not measuring the dominant factor

obtained in the factor solution are removed.

Parameter invariance indicates a constant characteristic of test items despite changes in the group of examinees answering the items. According to Bruce (2018), parameter invariance refers to population quantities whose values are to be estimated with data collected within a random sampling design. Parameters can in this context refer to the set of item parameters (item difficulty, discrimination, and/or guessing) and the set of examinee parameters (the examinee test scores, or theta [θ] scores, implied by the IRT model) that are tied to a particular measurement. A reasonable assumption about the parameter relating to the individual examinee is that each examinee responding to a test item possesses some amount of underlying ability. Thus, one can consider each examinee to have a numerical value, a score, which places him or her somewhere on the ability scale (Dibu, Kunmi, Francis & Patrick 2012). This ability score is denoted by the Greek letter θ. At each ability level, there is a certain probability that an examinee with that ability, will likely give a correct answer to the item. Under IRT, P(θ) is used to represent this probability. When answering an item, a testee of low ability will have low probability while a testee of high ability will have high ability. The parameters relating to each of the items of the test under IRT of the 3-parameter model include: (i) Discrimination index (**a**-parameter): This measures how accurate a question makes a distinction among testees of different level of ability. An item with a low value discriminates poorly over a wide range of abilities while item with a high value discriminates well, but over a small range of abilities. (ii) *Difficulty index* (b-parameter): It is a measure of the proportion of examinees who answer an item correctly and for this reason; it is frequently called the b-value. Items with high values of b are difficult items, and low-ability examinees have a low probability of correctly responding to the item. Items with low-values of b are easy items with most examinees, even those with low-ability values, having at least a moderate probability of answering the item correctly. (iii) *Guessing index*(c-parameter): It is otherwise known as the pseudo-chance parameter. In multiple-choice items, an examinee who does not know the correct alternative may succeed in responding correctly by random guessing.

The performance of an item in a test is described by the Item Characteristics Curve (ICC) or otherwise known as Item Response Function (IRF). The curve gives the probability that a person with a given ability level will answer correctly. Persons with lower ability, have less of a chance while persons with high ability are likely to answer correctly (Hambleton, Swaminthan, & Rogers, 1991).

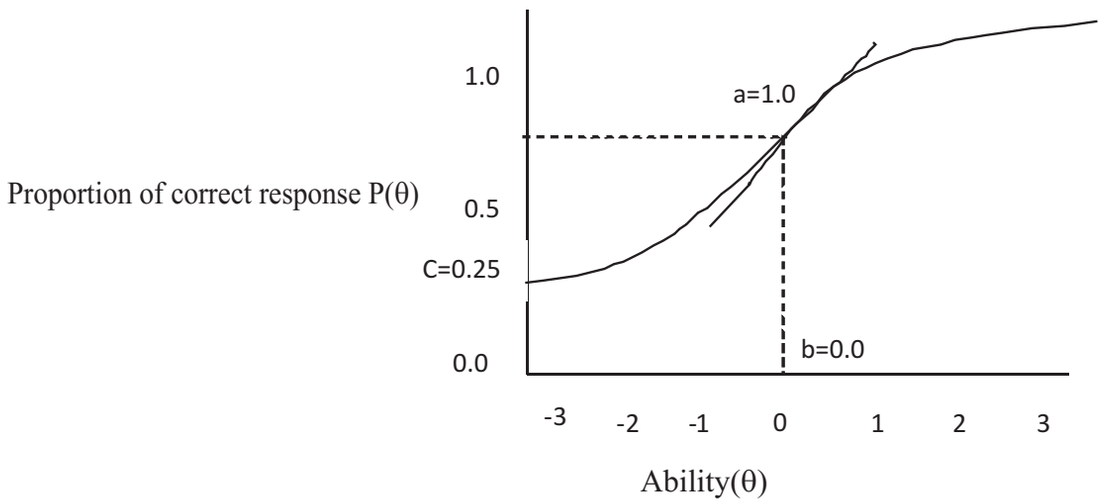


Figure 1: Item Characteristic Curve (ICC)

The shape of ICC is gotten by item difficulty, discrimination, and guessing indices of an item. Figure 1 above shows an example of the 3PL model of the IRF with a detailed explanation of the parameters. The parameter b represents the item location which, in the case of attainment testing, is referred to as the item difficulty. The ICC has its maximum slope at point θ. The item used is of moderate difficulty, since b=0.0, is located at the center of the distribution. It is emphasized in figure 1 above, the placement of item's difficulty and the person's trait on the same continuum as pointed out by the model. Getting the item correctly depends on the level of person's trait and item's difficulty. The level of a person's trait needs to be equal or greater than item difficulty for the successful performance of the person. The item parameter a, denotes the item discrimination that measures how accurate a question makes a distinction among testees of different level of ability. This parameter is shown in the figure above as the slope of the ICC where the slope is at its maximum. The example item has a=1.0, which discriminates fairly well: persons with low ability do indeed have a much smaller chance of correctly responding than persons of higher ability.

The parameter c, determines the effects of guessing on the probability of a correct response. This explains the probability of how a very low ability person will get this item correct by chance, mathematically represented as a lower asymptote. A multiple-choice item of four options might have an ICC like the example of item used: there is an equal probability of $\frac{1}{4}$ on all options of guessing the correct answer by an extremely low ability person, so the *c* would be approximately 0.25. This makes all the options equally plausible, because if one option made no sense, even the lowest ability person would be able to throw it away, so IRT parameter estimation methods take this into account and estimate a, b, and c, based on the observed data. For example, in the three parameter

logistic (3PL) model, the probability of a correct response to an item 1 is;

$$P_i(\theta)=C_i + \frac{1- C_i}{1+e^{-ai(\theta-bi)}} \quad \text{---------- 1}$$

Where θ is the person (ability) parameter and *ai, bi* and *ci*, are item parameters.

In test situation involving IRT, examinees' performance on examination items can be explained by defining examinees' trait, estimating scores for examinees on these traits and using the scores to explain performance (Opasina, 2009, Adedoyin, 2010). The modern Rasch Measurement Model developed in 1960 by George Rasch is a unidimensional model that belongs to Item Response Theory model. Test experts are expected to generate good items that can be used to examine the ability of students from whether homogeneous or heterogeneous settings, as the value of such a measure would be domiciled in its quality. This would be primarily possible through measuring tools (tests or examination) whose unidimensionality is assured and level of item parameters that are appropriate with the person parameter. This study will examine these qualities to measure the level of item performance in 2016 Biology multiple-choice items of OSJPE for the purpose of future use. Biology as a major subject for the students at the secondary school level, this study seeks to contribute to the existing body of knowledge by providing empirical evidence on psychometric parameters such as difficulty, discrimination, and guessing indices of Biology items used to prepare students for WAEC and NECO. It is therefore important in the study to examine the item statistics of Biology multiple-choice items using X-calibre being modern software for identifying and selecting good items for precise trait estimation and reliability of Biology scores.

The main objective of this study was to investigate unidimensionality and item parameters (difficulty, discrimination, and guessing indices)of Biology items of Osun State Joint Promotion Examination (2015/2016).

Research Questions:

These research questions were raised to achieve the objectives of the study:

(i) What is the dimensionality of the 2016 multiple-choice Biology items of OSJPE?

(ii) What are the item parameters (difficulty, discrimination, and guessing) of Biology items of OSJPE (2015/2016)?

Methodology

The study adopted non-experimental research design of the descriptive research involving ex-post-facto type. This design is considered suitable for the study since the occurrence of an event in this had already taken place. The population for the study comprised students who responded to the 40 Biology multiple-choice questions in the 2016 OSJPE conducted by the Osun State Ministry of Education. The sample size consisted of an intact class of

2,000 students' responses in Biology as contained in the Optical Mark Recorder (OMR). The research instruments used were the 40 multiple-choice Biology questions of 2016 OSJPE and response sheets of all the students that wrote the 40 items multiple-choice Biology questions of OSJPE (during the 2015/2016 session) in the selected schools as contained in the Optical Mark Recorder (OMR) and their keys. All the students' responses to 40 multiple-choice Biology items during the 2015/2016 OSJPE in the selected schools constituted the sample for the study.

For item calibration, the 40 multiple-choice Biology items were scored using the dichotomous approach. In this case, every correct response out of four options length was scored as one (1) and the remaining wrong options attracted zero (0). The possible maximum obtainable score was 40 marks. For item validation, the researcher assumed that being an examination conducted by the state ministry of education, experts would have been consulted for content coverage based on the school syllabus. However, the results emanate from this study focused on the reliability aspect of the items. Students' responses to the 40 multiple-choice Biology items were treated as a data file and the control file consisted of students' ID, keys, numbers of alternatives, domain and inclusion status. To answer question 1, the coded data were subjected to factor analysis to establish the unidimensionality of the items. To answer question2, Biology items were calibrated to ascertain item quality with reference to item difficulty, discrimination, and guessing indices (using Item Parameter Output) under the 3-paramter model using X-calibre4:2 header. X-calibre is software that is capable to generate outputs that is reliable and dependable.

Results

Research Question 1: What is the dimensionality of 2016multiple-choice Biology items of OSJPE?

To answer this question, the responses of the students on 40 multiple-choice items of OSJPE Biology examination were subjected to factor analysis. This analysis was done to determine whether all these measures shared some common variance and, thus, could be said to tap the same underlying construct using the principal components analysis (PCA) to extract the initial factors. The extraction of factors was based on the suggestion by Zwick and Velicer (1986) that the eigenvalue-greater-than-one should be selected as the extraction rule. This rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity be excluded from the analysis. The results of the factor analysis are represented in Table 1 and Figure 2.

Table1:

Total variance explained by result of factor analysis

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	5.778	14.446	14.446	5.778	14.446	14.446
2	2.188	5.470	19.916	2.188	5.470	19.916
3	1.588	3.971	23.886	1.588	3.971	23.886
4	1.362	3.405	27.291	1.362	3.405	27.291
5	1.241	3.103	30.394	1.241	3.103	30.394
6	1.177	2.942	33.336	1.177	2.942	33.336
7	1.143	2.857	36.194	1.143	2.857	36.194
8	1.110	2.774	38.968	1.110	2.774	38.968
9	1.074	2.684	41.652	1.074	2.684	41.652
10	1.027	2.567	44.219	1.027	2.567	44.219
11	.998	2.495	46.714			
12	.986	2.465	49.179			
13	.963	2.408	51.587			
14	.937	2.343	53.930			
15	.917	2.292	56.222			
16	.877	2.192	58.414			
17	.870	2.176	60.590			
18	.851	2.127	62.717			
19	.835	2.088	64.805			
20	.814	2.034	66.839			
21	.798	1.996	68.835			
22	.796	1.991	70.826			
23	.767	1.916	72.742			
24	.751	1.877	74.619			

25	.742	1.855	76.474
26	.724	1.810	78.284
27	.701	1.753	80.037
28	.690	1.725	81.762
29	.684	1.711	83.473
30	.678	1.694	85.167
31	.673	1.683	86.849
32	.649	1.624	88.473
33	.640	1.601	90.074
34	.624	1.559	91.633
35	.603	1.508	93.141
36	.577	1.442	94.583
37	.560	1.400	95.983
38	.549	1.372	97.354
39	.545	1.363	98.717
40	.513	1.283	100.000

From Table 1, the result of principal component analysis produces ten (10) items with eigenvalues greater than one. These ten factors explain 44.22% of the variance. The first eigenvalue was 5.78 higher than the next eigenvalue (i.e. 2.19, 1.59, 1.36 etc.). The first factor explained 14.45% of the variance; the second factor explained 5.47% of the remaining variance. The remaining variances were explained by the other 30 factors. Hence, there is one dominating factor in the factor structure of item set. Since there is a dominating factor that explained 14.45% of the variance the assumption of unidimensionality is established. The result of the eigenvalue test produced the scree plot to determine whether the dimensionality could be inferred.

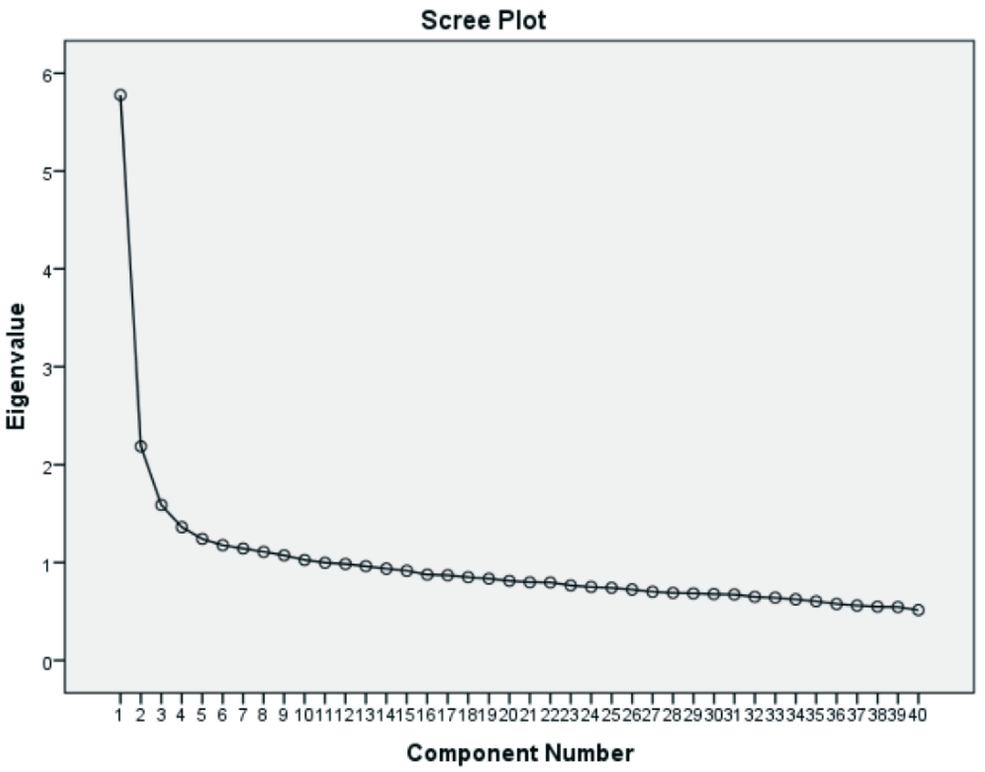


Figure 2: Scree Plot of factor analysis result using principal component analysis

Looking at Figure 2 above, the eigenvalue of the first factor was larger compared to the second factor, and the eigenvalue of the remaining factors are all about the same. Hence, the assumption of unidimensionality was satisfied in the 2015/2016 multiple-choice Biology items of OSJPE, since there was one dominating factor (first factor with the eigenvalue of 5.78 that explained 14.45% of the total test variance) among the items set using factor analysis.

Research Question Two: What are the item parameters (difficulty, discrimination, and guessing) of Biology items of 2016 OSJPE?

To answer this research question, the OSJPE 40 Multiple-Choice Biology items were dichotomously scored and calibrated with X-Calibre 4.2-IRT statistical software using three-parameter logistic model (3PL). The data and control file were created under notepad; the data file contained responses of 2000 students to 40 multiple-choice Biology items while the control file contained students' item option length, domain inclusion (V) and multiple (M) formats. The calibration was performed using maximum E-M loops of 65 at 0.001 convergence criterion, the item parameters of 2016multiple-choice Biology items

of OSJPE generated using IRT framework [i.e. item difficulty (threshold or b), item discrimination (slope or a), and guessing index (c)] are presented in Table 2

Table 2:

Item parameters for dichotomously scored 2016 Biology items of OSJPE

Seq.	**3-Parameter Model**		
	A	B	C
1	3.904	0.892	0.557
2	5.276	0.878	0.537
3	4.204	1.692	0.530
4	3.238	0.855	0.329
5	2.475	1.566	0.308
6	1.788	0.490	0.333
7	3.075	0.837	0.298
8	1.849	0.620	0.383
9	0.945	1.137	0.305
10	2.095	0.817	0.338
11	1.268	0.531	0.288
12	3.617	1.791	0.298
13	3.498	1.786	0.308
14	0.347	3.088	0.252
15	1.467	3.507	0.279
16	2.002	0.630	0.283
17	2.088	0.560	0.314
18	1.502	0.472	0.245
19	1.830	0.517	0.283
20	1.447	0.676	0.314
21	1.501	0.289	0.202

22	1.958	0.544	0.245
23	1.349	0.358	0.210
24	3.183	1.805	0.283
25	0.977	0.364	0.210
26	0.793	0.643	0.209
27	1.692	0.519	0.265
28	1.436	0.393	0.215
29	0.924	0.793	0.227
30	1.602	0.303	0.177
31	2.589	2.363	0.200
32	0.784	1.459	0.216
33	1.528	0.487	0.268
34	3.341	1.798	0.333
35	1.883	1.490	0.230
36	3.391	1.760	0.225
37	2.554	2.400	0.234
38	2.529	2.429	0.279
39	1.107	0.406	0.326
40	0.339	-0.402	0.252

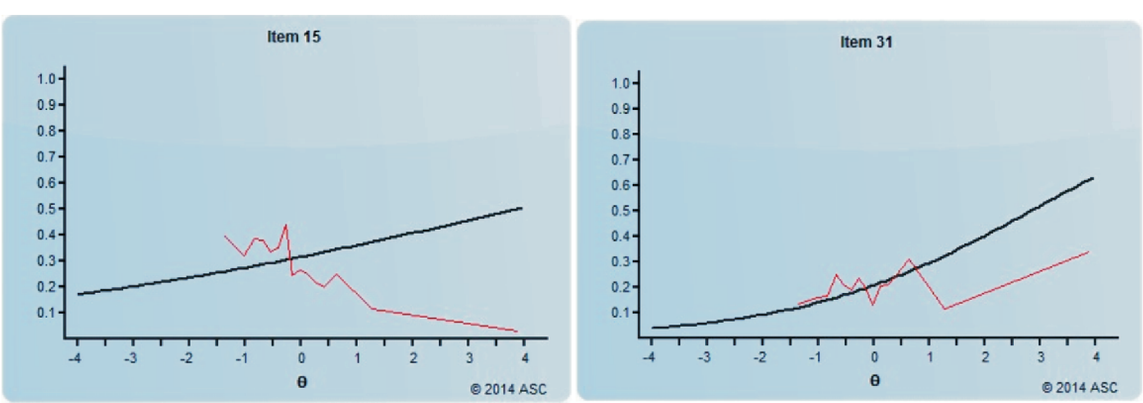
The results are interpreted based on the set standard for interpreting difficulty and discriminating indices as shown in the table 3 below. The bold values of item difficulty and discrimination indicate the items that are the set standard of difficulty level, low discriminating ability, and high guessing tendencies because the values are not within the set standard of moderate difficulty level ($-1 \leq 1.00$), moderate discrimination ability ($0.65 \leq a \leq 1.7$ and above) and low guessing tendency (< 0.3).

Table 3:

Distribution of Items based on Difficulty values and Discrimination indices

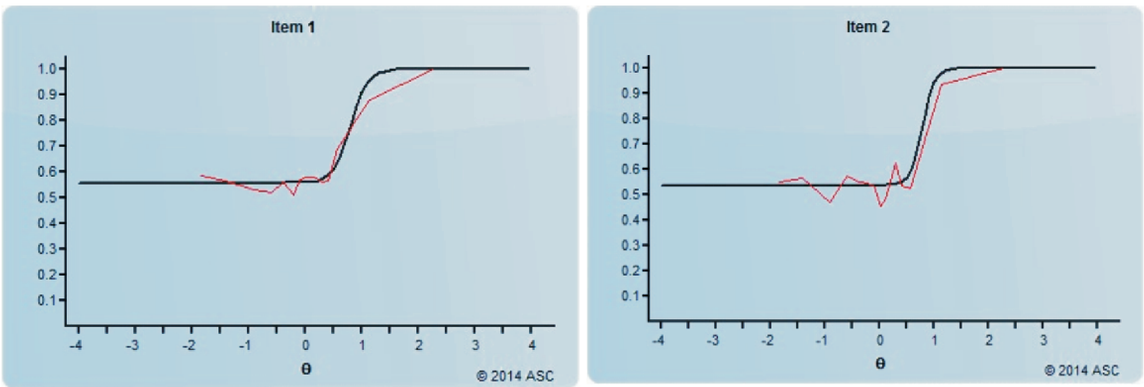
Item Difficulty (b-value)	**Total 3PL**	**Discrimination (a-value) 3PL**	**Total**
Easy (-3.00 ≤-1.00)	0(0%)	Excellent (a ≥1.70)	22(55%)
Moderate (-1.00 ≤ 1.00)	25(62.5%)	Good (1.35 ≤ a ≤ 1.69)	9(22.5%)
		Moderate (0.65 ≤ a ≤ 1.34)	7(17.5%)
Difficult (1.00 ≥ 2.00)			
	15(37.5%)	Marginal (0.35 ≤ a ≤ 0.64)	1(2.5%)
		Poor (a ≤ 0.34)	1(2.5%)

Based on the difficulty index under 3PL model of IRT, the result indicates that none of the items were found to be easy, 25(62.5%) were moderate and the remaining 15(37.5%) were difficult. The findings on the basis of item discrimination indices, the results indicates that only 1(2.5%) of the items was poor, 2(5%) of the items were of marginal discriminating ability, 6(15%) were found to be moderate, 8(20%) were found to be good and the remaining 23(57.5%) of the items were found to be excellent. Based on guessing indices (c), theoretically, c ranges from 0.0 to 1.0, but is typically <0.3. Fourteen (35%) of the items were found to have high guessing tendency while the remaining items were of low guessing tendency.



a= 0.194, b= 4.000 a= 0.456, b=2.975

Figure 3: Item Characteristics curves that show items that have high difficulty level and low discriminating power.



c= 0.557 c= 0.537

Figure 4: Item Characteristic Curves which show the items that have high guessing tendency.

Discussion

The focus of this study is to investigate unidimentionality and item parameters of Biology items of the 2015/2016 OSJPE. The first question relates to the verification of IRT model assumptions; the unidimensionality assumption of the model was established. Since there was one dominating factor in the factor structure of the items set, the assumption of unidimensionality was satisfied in the 40 multiple-choice Biology items of 2016 OSJPE which relaxes the assumption of item response theory (Embretson & Reise, 2000).

The second question dealing with the determination of the OSJPE item parameters was addressed as the parameters were determined using the X-calibre 4.2 and presented in Table 2 with items of special interest highlighted with bold figures. The findings further revealed that, based on the established standards, 62.5% of the items were problematic under 3PL IRT model. The Item Characteristic Curves of the test items show different behaviours (some were easier while others were difficult and of high guessing tendency). Under 3PL, none of the item was found easy, 25 (62.5%) of the items were of acceptable difficulty level, and 15(38%) items were difficult. Based on guessing under the 3PL, 26 (65%) of the items were of low guessing tendency, and 14(35%) items were of high guessing tendency. For example, item 1 in Figure 4 presents the ICC, it shows that the value obtained on the ability scale (difficulty parameter estimate), that is 0.5 probability of examinees getting the item right is low (-4). This means the item is very easy for the students.

However, items 15 and 31 in Figure 3 showed that the value obtained on the ability scale (difficulty parameter estimate) 0.5 probability of examinees getting the item right is high (4 and 2.975 respectively). This means the item is difficult for the students. These difficult items should be rejected from the entire test. Going by this result, 26 (65%) of the items were of acceptable difficulty, the remaining 14 (35%) of the items having turned problematic based on difficulty parameter should be modified or eliminated from the test completely. This finding is consistent with the findings of (Pande et al., 2013) and (Surachi & Rana, 2015) whose findings revealed that majority (75%) and (78%) of the items respectively, were acceptable as far as difficulty was concerned. The reason for this consistency may be due to the facts that the test construction may have been made to undergo certain psychometric considerations by the developers to allow for items that are relatively not too easy and difficult respectively in the test.

Under 3PL model, only 1 (2.5%) item was of marginal and poor discriminating ability, 7 (17.5%) of the items were of moderate discriminating ability and 31 (77.5%) of the items differentiate between students of different abilities. Similarly, 13 items presented a marginal and poor discriminating ability or cannot differentiate substantially between low and higher achieving students, therefore, the itemsneed to be reviewed or should be rejected. The findings of this study seem to agree with that of (Pande et.al., 2013) whose study revealed 75% of the multiple-choice questions (MCQs) items in formative examination in Physiology were within acceptability of discrimination. The reason for the consistency in the results may be that the test items have the ability to take care of the test takers abilities as it is important that test items must discriminate well to allow for floor and ceiling effect to occur.

Similarly, on the overall, based on the IRT parameters generated, 25 (62.5%) of the items were considered to have defects and are, therefore, needed to be reviewed especially if the items are to be used in the subsequent examinations. The poor performance of these 25 (62.5%) items and the students could have been due to poor understanding of difficult topics, ambiguity in wordings of the questions or even inappropriate key, it may also be due to personal variations in students' intelligence level (Bichi, 2015).

Conclusion

Findings of this study indicates that the items of the Osun State Joint Promotion Examination are not stable as far as item difficulty, item discrimination, and guessing indices and differential item functioning are concerned using the IRT Frameworks. As a result of these item defects, the 40 multiple-choice Biology items of 2016 OSJPE were considered to have low psychometric quality and this could not provide an objective result without bias and standardized view of a range of candidates' competencies such as knowledge and ability. Item analysis results generated may be influenced by many other factors which include examinees having a poor understanding of difficult to topics, ambiguity in wordings of the questions or even inappropriate key, instructional procedure

applied, it may also be due to personal variations in students' intelligence level. Similarly, the results of this study shows the need to improve the OSJPE biology items especially in developing valid and reliable items, through the mandatory involvement of experts in tests, measurement and evaluation in the process of constructing /developing the multiple-choice biology achievement test. Based on the findings, it is recommended that IRT should be maintained in construction of Biology test items that will be gender and location balance, because of its position in the investigation of reliability and in minimizing measurement errors. Also, in compliance with the standards and current practice in development and validation of test items, the 'problematic' items identified in this study, having failed to satisfy the set quality criteria, should be modified, dropped, or completely eliminated from the test.

References

Adedoyin, O. (2010). Investigating the invariance of person parameter estimates base on classical test and item response theories. *International Journal of Education, 2*(2), 107-113.

Angoff, W. H. (2007). Validity: An evolving concept. In H. Wainer & Braun, H. (Eds.), *Test validity*(pp. 19-32). Hillsdale, NJ: Erlbaum.

Bichi, A. A. (2015). Item analysis using a derived science achievement test data. *International Journal Science and Research, 4*(5), 1655-1662.

Brown, J. D. (2005). *Testing in language programs, a comprehensive guide to English Language assessment*. New York: Mcgraw-hill College.

Bruce, B. F. (2018). Parameter invariance. *The SAGE Encyclopaedia of Educational Research, Measurement and Evaluation*. Retrieve from http://www.methods.sage.pub.com.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Dibu, O., Kunmi, P., Francis, O. & Patrick (2012). *Introduction tottem response theory: parameter models, estimation and application*. Abuja: Marvelouse Mike Press LTD.

Faleye, B. A. & Benjamin T. A. (2016). Continuous assessment practices of secondary school teachers in Osun State, Nigeria. *Journal of Psychology and Behavioral Science, 4*(1), 44-55.

Hambleton, R. K., Swaminathan, H. & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Lumsden, J. (1961). Construction of unidimensional test. *Psychological Bulletin*, *58*, 122-131.

Opasina, O. C. (2009). *Development and validation of alternative to practical Physics test using Item Response Theory Mode*(Unpublished PhD Thesis).University of Ibadan, Ibadan, Nigeria.

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Angrekar, S. (2013). Correlation between difficulty and discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal Medical Education, 7*(1), 45-50.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: META Press.

Rezaee, A. & Salehi, M. (2008). The construct validity of a language proficiency test: a multitrait multimethod approach. *TELL, 2* (8), 93-110.

Surucchi, & S. R. (2015). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in Biology. *Idian: J. Res*., 3(6), 56-58.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.

Zumbo, B. D. (2011). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.