IOSR Journal Of Humanities And Social Science (IOSR-JHSS)
Volume 26, Issue 11, Series 5 (November. 2021) 01-09
e-ISSN: 2279-0837, p-ISSN: 2279-0845.
www.iosrjournals.org

Development and validation of geography diagnostic test for senior secondary school students using item response theory

Nkechi Patricia-Mary Esomonu[1], Goodluck Erutujiro [2]
[1]*Department of Educational Foundations*
NnamdiAzikiwe University, Awka
[2]*Department of Educational Foundations*
NnamdiAzikiwe University, Awka.

Abstract
The study is on development and validation of geography diagnostic test for senior secondary school students using item response theory. In carrying out the study, ten research questions and two null hypotheses were formulated. The study used instrumentation research design. The population of the study was made of 2,972 senior secodary2 students offering geography in government owned schools in Anambra State, Nigeria. 1400 students were selected for the study through random sampling technique. The initial instrument consisted of 80 multiple-choice items constructed by the researchers. The initial instrument was validated by six experts. After validation, the instrument was administered to the sampled students. The collected data were analyzed using DIMTEST statistics in DIMPACK 1.0 software, maximum likelihood estimation techniques in Mplus 7.0 software and Chi-square of goodness-of-fit in eirt software. Based on the analysis, it was found that: the instrument was unidimensional, the three parameter model of item response theory represents the best fit of the instrument data.. Sixty (60) items in the instrument were found fit and twenty items were rejected. Some of the rejected items discriminated poorly, some had high guessing power, and others had poor item characteristics, poor differential functioning across gender and high standard error of measurement (SEM). Sixty items were good and made the final form of the geography diagnostic test. The empirical reliability of the instrument was 0.98. The final instrument is valid and reliable. Based on the findings of the study, the researchers recommended that teachers should use the final instrument to diagnose persistent learning problems of students offering geography in secondary schools so that remedial help can be provided.

Keywords: Instrumentation, Test, Diagnostic test, Item response theory, Geography

---

Date of Submission: 05-11-2021 Date of Acceptance: 20-11-2021

I. INTRODUCTION

Diagnostic testing is a very important aspect of teaching and learning process that provides quality control measures by assessing learners' strengths and weaknesses for remediation. According to Gani (2012), diagnostic testing improves teaching and learning in education as it identifies the strengths and weaknesses of students and also indicates the effectiveness or ineffectiveness of the education system. Obadere (2017) noted that the outcome of diagnostic testing with a proper remediation will go a long way in reducing failure rate especially in the standardized examinations and improves performance in the area of skills acquisition. In order for diagnostic testing to work effectively, instructors need valid and reliable diagnostic tests.

Diagnostic tests are very important tools in diagnostic testing and assessment of students. They are tools used to assess persistent learning difficulties faced by students in schools. However, when diagnostic tests are poorly constructed they can lead to inaccurate measurement of learning and false information regarding students' performance as well as instructional effectiveness. Results from poorly designed diagnostic tests can also harm students by labeling them in unjustified ways, unfairly denying them of opportunities or simply discouraging them. Hence, the need for measurement and evaluation experts to construct valid and reliable diagnostic tests for assessment of students deficienies. The students deficiency in learning skills in geography were widely reported by the chief examiners in the WAEC report sheets of 2015, 2016, and 2017 (25% of the students weaknesses in geography were in the area of identification skill, 23% in graphic interpretation skill, 18% in spatial relationship skill, 12% in position orientation skill, 10% in measurement skill and 10 % in

calculation skill).The above deficiencies in learning skills in geography account for poor performance of students in the subject particularly in examination (WAEC, 2015, 2016, & 2017).

The construction of a diagnostic test can be done using either the Classical Test Theory (CTT) or the Item Response Theory (IRT) measurement framework. The classical test theory was developed by Charles Spearman in 1904. It is based on the premise that there is no perfect measurement of ability. In other words, every observed score is made up of two components, the error score and the true score. The error score is normally distributed, uncorrelated with the true score and the expected mean of the error score is equal to zero. The CTT is the most widely used by teachers, researchers and measurement and evaluation experts in construction of tests because it is the easier approach to test construction(Haladyna, 2004).According to Ikona(2016) classical test theory has several advantages. Most researchers are familiar with its basic concepts. That is, researchers who have had any exposure to measurement theory are likely to have encountered CTT. Also, most of the scales that are available and most of the descriptions of those scales are based on principles of CTT. The nearly ubiquitous use of coefficient alpha as an indicator of reliability illustrates this point.

Many researchers like Ugodulunwa and Bulus (2017) developed a quantitative Economics diagnostic test for secondary school students in Jos Plateau State based on classical test theory. Eleje et al (2016) also developed a diagnostic test in Economics for secondary school students based on classical test theory. Chanrasegram, Treagust and Mocerino (2007) designed a diagnostic test in Chemistry for secondary schools in Far Province, Iran based on CTT. Similarly, Tan, Tabe, Goh and Chia (2005) developed a two-tier multiple-choice diagnostic instrument to asses students' level of understanding of ionization energy across the periodic table in Chemistry for secondary school students in Singapore based on CTT. The aforementioned literature indicates no such instrument in geography in Nigeria.

The major limitation of the classical test theory is that the statistics that form its cornerstone, item difficulty and item discrimination indices are both sample dependent. That is the higher item difficulty is obtained from examinees of lower ability, while lower item difficulty is obtained from sample of higher ability. Also, higher scores are associated with test composed of relative easy items and low scores are function of a test composed of items that are more difficult (Innabi &Dodeen, 2018). In terms of discrimination indices, higher values tend to be obtained from heterogeneous examinee sample and lower values are created with homogenous sample. Such sample dependency relationship reduces the overall utility of the item statistics. To overcome the above limitations, another measurement theory called item response theory was developed (Ojerinde, 2013).

Item response theory is a modeling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance. Ayiro (2017) described item response theory as a body of theory describing the application of mathematical models to data from questionnaires and tests as a basis for measuring things such as abilities and attitudes. Item Response Theory (IRT) looks at the examinee's performance by using item distributions based on the examinee's probability of success on a latent variable. Under IRT, parameters of the persons are invariant across items, and parameters of the items are invariant in different populations of persons (Hardiyanti, Mansyur, & Munawwarah, 2018*).* It places item difficulty and student performance on the same scale, tells how much information each item or item score level contributes to the test, and provides standard error of measurement for each item and statistic of fit of each item to the model. This brings greater flexibility and provides more sophisticated information which allows for the improvement of the reliability of an assessment.

Researchers like Esomonu and Eleje (2017) developed a diagnostic test to measure economics quantitative skill of secondary school students based on IRT. Dadughun (2015) also developed a Primary School Mathematics Diagnostic Test based on IRT for use on primary four pupils. Similarly, Young (2014) developed a diagnostic test in English Language for secondary school in Kisumu Municipality, Kenya using item response theory. Furthermore, Chang (2013) also constructed a diagnostic test in Economics for secondary school in Far Province, Iran using item response theory. .Latun (2011) also designed a diagnostic test in Physics for secondary school students in Limpope Province of South Africa using item response theory. The aforementioned literature indicates no such instrument in geography based on item response theory in Nigeria Item Response Theory has three basic assumptions namely; (1) dimensionality, which can be one-dimension (items in a scale measure only one latent construct) or multi-dimension (items in a scale measure more than one latent constructs or trait), (2) local independence of items (that is the relationship among items is only explained by the conditional relationship with the latent construct or trait) and (3) the response of a person to an item can be modeled by a mathematical item response function. Item response function gives the probability that a person with a given ability level will answer correctly a given item. That is, persons with lower ability have less of a chance, while persons with high ability are very likely to answer correctly. The exact value of the probability depends, in addition to ability, on a set of item parameters for item response function. The scale on which these are expressed is called *logit* scale.

The major benefit of item response theory approach in test development is that the parameters of the person do not depend on the parameters of the items, and vice versa. Also in item response theory, the standard

error of measurement gives precision at each level of the ability being measured. This implies that each examinee and item parameter is accompanied by its standard error of measurement thereby making measurement to be more precise, accurate and objective.

The scarcity of diagnostic test in geography as shown in literature could be possibly due to teachers' poor knowledge in development and validation of geography diagnostic test or lack of sufficient time on the part of the geography teachers to develop valid and reliable geography diagnostic instruments. To reduce this gap and problem, this study embarks on construction of a valid and reliable diagnostic test for assessment of students' strengths and weaknesses in geography at the secondary school level using item response theory. The study was guided by the following research questions:

- What is the dimensionality of the Geography Diagnostic Test (GDT)?
- Which of the IRT logistic models represent the best fit for the Geography Diagnostic Test data?
- How many items of the Geography Diagnostic Test fit the parameter logistic model
- What are the difficulty parameters of the items of the Geography Diagnostic Test?
- What are the discrimination parameters of the items of the Geography Diagnostic Test?
- What are the guessing parameters of the test item of the Geography Diagnostic Test?
- What are the standard errors of measurement of items of the Geography Diagnostic Test?

II. RESEARCH METHOD

The design of this study is instrumentation research design. The study was conducted in Anambra State, Nigeria **.**The population of the study was made up of 2,972 SS 2 students offering geography in 2017/2018 academic session in Anambra State. This comprised 1623 females and 1349 males (Source: Anambra State Post Primary School Service Commission, Awka, 2017).The sample for the study consisted of 1400 SS2 students drawn using simple random sampling technique.

The instrument was constructed based on learning skills which geography experts considered necessary for students understanding of geography at the secondary school level. The skills are: identification, calculation, spatial relationship, measurement, position orientation and graphic interpretation skills. The aforementioned skills were obtained from the Senior Secondary School Geography Curriculum and Chief examiners' WAEC report sheets.

Items measuring the skills were generated from various sources, such as locally prepared past examinations and tests, geography selection tests, standard achievement tests, textbooks, and from day to day experiences of geography teachers. The initial instrument consisted of 80 multiple choice test items with 8 items measuring the calculation skill, 20 items for the identification skill, 14 items for spatial relationship skill, 10 items for the position orientation skill, 10 items for the measurement skill and 18 items for the graphic interpretation skills.

Based on the order of priority placed on them by the chief examiners in the WAEC report sheets of 2015, 2016, and 2017 (25% of the students weakness in geography is in the area of identification skill, 23% in graphic interpretation skill, 18% in spatial relationship skill, 12% in position orientation skill, 10% in measurement skill and 10 % in calculation skill). In constructing the items of the instrument, a table of specifications was used. The initial instrument was presented to one expert in Measurement and Evaluation in a University as well as five subject experts at the secondary school level.

The instrument was administered to the sampled students. The difficulty level (b), the discrimination level (a) , the lower asymptote (c) parameters of each item were estimated using DIMTEST statistics in DIMPACK 1.0 software, information criteria statistics in Mplus 7.0 software, Chi-square goodness-of fit in eirt software, and maximum likelihood estimation techniques of Mplus 7.0 software. The following criteria guided the research in the selection of items

The following criteria guided the researchers in selection of final items:

1. p-value of DIMTEST statistics less than 0.05 significant level indicates uinidimensionality (Reckase, 2009).
2. The smallest information criteria statistics in term of Akaike information Criteria (AIC), Bayesian information criteria (BIC) and Sample- Size Adjusted Criteria indicates best IRT model fit (Muthen&Muthen, 2007).
3. Chi-square value of goodness-of- fit of item below 15.507 with probability greater than alpha level of 0.05 indicates fit to the model (Adedoyin, 2010).
4. Item difficulty parameter: Any item between -3 to +3 is good and should be retained (Baker, 2001).
5. Item discrimination parameter: negative values (very poor), low (0.01 to 0.34), moderate (0.35 to 1.34), high (1.35 -1.69) and very high (1.70 and above). Any item above 0.35 is good and should be retained (Baker, 2001).
6. Items with guessing value of 0.26 and above are not good, while items with guessing value of 0.25 and below are desirable (Harris, 2005)**.**

7. Standard error of 0.05 and below indicates high reliability, while error above 0.05 indicates low reliability (Obinne, 2013).

III. RESULTS AND DISCUSSION

Research Question 1

What is the dimensionality of the Geography Diagnostic Test (GDT)?

Table 1: Dimensionality Test Statistic of the Geography Diagnostic Instrument

TL	TGbar	T	p-value
14.6138	4.8041	9.7610	0.0000

Table 1 shows the p-value of the DIMTEST statistics is less than 0.05 significant level. This indicates that the underlying latent ability of examinees' responses to the instrument is unidimensional.

Research Question 2

Which of the IRT logistic models represent the best fit for the Geography Diagnostic Test data?

Table 2: Model Fit Information Criteria for 1PLM, 2PLM, 3PLM and 4PLM

Information Criteria	1PLM	2PLM	3PLM	4PLM
Akaike(AIC)	106394.437	101153.687	100956.711	101579.869
Bayesian(BIC)	106819.219	101992.763	101215.325	103258.022
sample-sizeAdjusted	106561.913	101484.583	101452.935	102241.501

Table 2 shows that 3PLM has the smallest information criteria in terms of Akaike information Criteria (AIC), Bayesian information criteria (BIC) and Sample-Size Adjusted. Therefore, the 3PLM represents the best fit for the data than the 1PLM, 2PLM and 4PLM. Thus the 3PLM was used in this study to estimate the item parameters.

Research Question 3

How many items of the Geography Diagnostic Test fit the three parameter logistic model?

Table 3. Summary of Item Fit to Three Parameter Logistic Model

Items that:	Frequency	percentage
FIT THE MODEL	60	75%
DO NOT FIT THE MODEL	20	25%

Table 3 shows that twenty items that is, Items 3, 6, 10, 12, 16, 20, 22, 31, 33, 34, 40, 43, 44, 48, 63, 66, 70, 72, 76 and 80 did not fit the 3PLM because their p-values are below 0.05 level of significance and the item Chi-square values are above 15.507 while Sixty items that is, Items 1, 2, 4, 5, 7, 8, 11, 13, 14, 15, 17 , 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 32, 35, 36, 37, 38, 3, 41, 42, 45, 46, 47, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 67, 68, 69, 71, 73, 74, 75, 77, 78 and 79 fit the 3PLM mode and therefore accepted because their p-values are greater than 0.05 level of significance and the items Chi-square values are below 15.507.

Research Question 4

What are the difficulty parameters of the items of the Geography Diagnostic Test?

Table 4. Item Threshold Values (b-values) of the GDT Based on 3PLM

Item	b	Item	B	Item	B	Item	b
1	-1.850	21	0.618	41	0.721	61	-1.850
2	-1.521	22	-2.180	42	-0.671	62	1.521
3	0.870	23	-1.479	43	0.061	63	-0.870
4	0.842	24	0.539	44	0.486	64	-0.840
5	-1.008	25	0.839	45	-0.957	65	1.008
6	0.239	26	0.153	46	-1.132	66	0.239
7	-0.858	27	0.396	47	-1.130	67	-0.858
8	-1.308	28	0.686	48	-1.421	68	-1.308
9	2.504	29	0.852	49	0.502	69	-2.504
10	-1.495	30	0.022	50	0.483	70	-1.495

11	-0.720	31	-2.424	51	-1.293	71	-0.720
12	0.340	32	-0.952	52	0.095	72	0.340
13	-0.625	33	0.635	53	-1.070	73	-0.625
14	-1.054	34	-3.606	54	1.531	74	-1.054
15	-1.286	35	0.579	55	1.266	75	-1.286
16	0.082	36	-0.919	56	-0.770	76	0.082
17	-1.543	37	-0.961	57	-0.918	77	1.543
18	-1.263	38	0.029	58	0.409	78	-1.263
19	-0.703	39	-1.037	59	-0.545	79	0.703
20	0.878	40	-2.946	60	-0.773	80	0.878

Table 4 shows that forty two items (42) items fall between -3 to 0 while thirty eight (38) items fall between 0 to +3. Based on this information, none of items was rejected in term of difficulty level.

Research Question 5

What are the discrimination parameters of the items of the Geography Diagnostic Test?

Table 5. Discrimination Parameters of the Items of the GDT based on 3PLM

Item	a	Item	A	Item	A	Item	A
1	1.153	21	1.460	41	1.648	61	1.153
2	1.554	22	0.106	42	1.684	62	1.554
3	-0.837	23	1.392	43	2.567	63	-0.837
4	1.784	24	1.615	44	2.178	64	1.784
5	2.055	25	1.958	45	1.326	65	2.055
6	1.845	26	3.046	46	1.479	66	1.845
7	1.993	27	1.593	47	1.173	67	1.993
8	1.574	28	1.896	48	1.414	68	1.574
9	0.871	29	2.040	49	1.889	69	0.871
10	1.635	30	3.681	50	1.547	70	1.635
11	1.834	31	-0.507	51	1.428	71	1.834
12	2.073	32	1.918	52	3.166	72	2.073
13	1.723	33	1.967	53	1.962	73	1.723
14	1.882	34	-0.518	54	1.577	74	1.882
15	1.679	35	1.735	55	1.442	75	1.679
16	20.466	36	1.795	56	1.702	76	10.666
17	1.827	37	1.591	57	1.513	77	1.829
18	1.690	38	2.961	58	1.802	78	1.690
19	1.927	39	1.418	59	1.434	79	1.927
20	0.193	40	-0.648	60	1.515	80	0.193

Table 5 shows that three (3) items are within the range of 0.01 to 0.34 indicated very low discriminating power, while six items within the range of 0.35 to 1.34 indicated moderate discriminating value. Also, fifty four (54) items within the range of 1.35 to 2.00 indicated high discriminating value. Thirteen Items are within the range of 2.01 and above indicated very high discriminating value. Five (5) Items had negative discriminating values. Therefore Items 3, 20, 22, 31, 34, 40, 63 and 80 were rejected.

Research Question 6
What are the guessing parameters of the test item of the Geography Diagnostic Test?

Table 6. Guessing Parameter of the Test Items of the GDT based on 3PLM

Item	c	Item	C	Item	C	Item	c
1	0.000	21	0.003	41	0.002	61	0.000
2	0.000	22	0.000	42	0.001	62	0.000
3	0.000	23	0.000	43	0.139	63	0.000
4	0.001	24	0.001	44	0.004	64	0.001
5	0.000	25	0.001	45	0.000	65	0.000
6	0.005	26	0.172	46	0.000	66	0.005
7	0.001	27	0.001	47	0.000	67	0.001
8	0.001	28	0.000	48	0.000	68	0.001
9	0.000	29	0.062	49	0.003	69	0.000
10	0.000	30	0.292	50	0.000	70	0.000
11	0.002	31	0.000	51	0.000	71	0.002
12	0.005	32	0.000	52	0.188	72	0.005
13	0.003	33	0.002	53	0.001	73	0.003
14	0.000	34	0.000	54	0.001	74	0.000
15	0.000	35	0.003	55	0.000	75	0.000
16	0.264	36	0.000	56	0.000	76	0.264
17	0.000	37	0.001	57	0.000	77	0.000
18	0.000	38	0.102	58	0.002	78	0.000
19	0.001	39	0.000	59	0.001	79	0.001
20	0.000	40	0.000	60	0.000	80	0.000

Table 6 shows that guessing values (c-values) of the items range from 0.000 to 0.292. Table 6 also indicates seventy seven(77) itemslied between 0.000 to 0.250 while the remaining three (3) Items fall between 0.251 and above which indicates that the items are not very good and the probability of guessing the items correctly is very high. Therefore, Items 16, 30, and 76 were rejected.

Research Question 7
What are the standard errors of measurement of items of the Geography Diagnostic Test?

Table 8: Standard Error of Measurement of Items of the Instrument

Item	SE	Item	SE	Item	SE	Item	SE
1	0.01	21	0.01	41	0.03	61	0.03
2	0.03	22	1.00	42	0.01	62	0.02
3	0.99	23	0.02	43	0.04	63	0.98
4	0.02	24	0.02	44	0.01	64	0.03
5	0.01	25	0.03	45	0.03	65	0.02
6	0.02	26	0.02	46	0.02	66	0.01
7	0.05	27	0.04	47	0.04	67	0.04
8	0.07	28	0.03	48	0.03	68	0.01
9	0.01	29	0.02	49	0.02	69	0.02
10	0.02	30	0.01	50	0.01	70	0.04
11	0.03	31	1.00	51	0.04	71	0.02
12	0.02	32	0.01	52	0.01	72	0.04

13	0.04	33	0.02	53	0.02	73	0.05
14	0.02	34	1.00	54	0.04	74	0.03
15	0.03	35	0.02	55	0.02	75	0.03
16	0.01	36	0.03	56	0.02	76	0.04
17	0.02	37	0.02	57	0.03	77	0.01
18	0.01	38	0.04	58	0.03	78	0.01
19	0.02	39	0.05	59	0.05	79	0.02
20	0.02	40	1.00	60	0.04	80	0.89

Average SEM =0.109

Empirical reliability =1-$(0.109)^2$, Empirical reliability =0.98

The table 7 indicates that 73 items had standard error below 0.05, while 7 items had above 0.05. The empirical reliability of the instrument is 0.98. This showed that the instrument is reliable.

The study indicated that the underlying latent ability of examinees responses to the instrument is unidimensional. A test is unidimensional when the performance of each examinee is assumed to be governed by a single factor referred to as ability. The assumption of unidimensionality means that all the geography diagnostic test items measures or are governed by only one underlying latent ability. Unidimensionality of the instrument also implies that only one latent variable explains all the correlations measured between the items. Furthermore, unidimensionality of the construct leads to stable measurements which is very important in diagnostic testingimportant in diagnostic testing . The above findings is in line with Reckase (2009) that stated unidimensionality of test items exists when a variable (often called a latent variable, as this variable may not be observed) which explains' all the correlations observed between the items in a measuring instrument. The above finding is in line with Esomonu and Eleje (2017) study that economics quantitative diagnostic test for secondary school students based on IRT is unidimensional. Similarly, .Latun (2011) also found physics diagnostic test for secondary school students based on IRT to beunidimensional.

The study revealed that the three parameter model represents the best fit for the instrument data than the one parameter logistic model, two parameter logistic model and four parameter logistic model. Thus, the three parameter logistic model was used in the study to estimate the item parameters and to generate the item characteristic curves of the test items. See Appendix 1. The assessment of model fit is very crucial in development of any instrument using item response theory. When a particular model fits the test data of interest, several desirable features are obtained. Ability estimates are obtained from different set of items will be the same. The property of invariance is only present when a model fit the data. Model fit also reduce the risk of drawing incorrect conclusion regarding the items parameters. This is in line with Stone and Zhang (2003) who reported that assessing item response theory model fit to item response data is one of the most crucial steps before an IRT model can be applied with confidence to estimate item statistics or ability level of examinees. The property of invariance is only present when model parameters are estimated properly. The property of invariance or item free measurement and sample free measurement allows for generalization beyond the specific test (Kose, 2014). Similarly, Sinharay (2005) also noted that substantial lack of model fit could result in overestimation of items and ability estimates.

The study revealed that 60 items fitted the 3PLM model. Within the latent trait test model, the internal validity of the test is assessed in terms of statistical fit of each item to the model. Fit of item to the model also implies that item discrimination is uniform and substantial, that is there are no errors in item scoring. The analysis of fit is a check on internal validity. If the fit statistics of an item is acceptable, then the item is valid (Esomonu&Eleje, 2017).According to Dadughan (2015), fit of item to the model also implies that guessing has had a negligible effect on the test taker.

In term of difficulty, all the test items were appropriate for measuring examinees of different abilities. A good test item should neither be too difficult nor easy for the examinee. This in line with study by Dadughan (2015) that suggested a good test item should not be too difficult for examinee, at the same time it should not be too easy for them. Similarly, 75 items have good discriminating values while 5 items discriminate poorly. The implication of the above is that most of the items in the instrument can discriminate various skill deficiencies of the students offering geography in secondary schools. Furthermore, 78 items had low guessing values, while 2 items had high guessing values. A good item should not have high guessing parameteras this can make examinee of low ability to score very high. This is in agreement with Young (2014) that recommended test items of low guessing value in measuring diagnostic assessment should not be accepted.

. The study revealed that 73 items had standard error below 0.05 which indicates high reliability. The standard error of measurement allows researchers to determine the probable range within which the individual's

true score fall. The result is in agreement with Obinne (2013) that standard error of 0.05 and below is described as high reliability, while error 0.05 is described as low reliability. The result is also in agreement with Meredith et al (2007) that if reliability increases, the standard error of measurement becomes smaller. According to Chatterji (2003), standard error of measurement is a statistical estimate of the amount of random error in the assessment of results or scores.

IV. CONCLUSION

In conclusion, sixty items scaled through all the checks in the analysis and were fund to be good. Five items were rejected due to poor discrimination. Two Items were rejected due to high guessing power. Thirteen Items were rejected due to poor item characteristics curve of item response theory. Two items were rejected because they function significantly different across gender, while seven Items were rejected due to high standard error of measurement (SEM).Some of these errors ran concurrent among some items hence in summary only twenty items were rejected. The sixty items that were good and made the final form of the geography diagnostic test is valid and reliable. The study has provided a valid diagnostic geography test for classroom interaction. Based on the findings of the study, the researchers recommended that teachers should use the final instrument to diagnose persistent learning problems of students offering geography in secondary schools so that remedial help can be provided. The test developed is with the authors.

ACKNOWLEDGEMENTS

We thank the principals of secondary schools that were used in the study as well as students that responded to the geography diagnostic test. We also acknowledged the research assistants that helped in administration of the instrument to the students in the area of study.

REFERENCES

[1] P. S. Gani, "*Diagnostic assessment of senior secondary two students' achievement in quantitative aspect of Economics in Akwanga Educational Zone, Nasarawa State".* Unpublished M.Ed. Thesis. Department of Educational Foundations, University of Jos, 2012.

[2] A. T. Obadare, "*Diagnostic assessment: A tool for quality control in education".*Retrieved from https://aai.ku.Edu/sites/aai.Ku.du/files/docs/ conference/NCME17, 2017.

[3] A. T. Obadare, "*Diagnostic assessment: A tool for quality control in education".*Retrieved from https://aai.ku.Edu/sites/aai.Ku.du/files/docs/ conference/NCME17, 2017

[4] West African Examination Council. "*Chief examiners' report".* Retrieved from waeconline.org.ng/e-learning/geography/24.html, 2015.

[5] West African Examination Council. "*Chief examiners' report".* Retrieved from waeconline.org.ng/e-learning/geography/16.html, 2016.

[6] West African Examination Council. "*Chief examiners' report".* Retrieved from waeconline.org.ng/e-learning/geography/2.html, 2017

[7] T. M. Haladyna, *Developing and validating multiple-choice test items*, Mahwah, NJ: Lawrence Erlbaum, 2004.

[8] K. Ikona, "*Unidimensionality of mock mathematics for Cross River State secondary schools".* Seminar paper present to the Department Of Educational Foundations University of Calabar, Calabar – Nigeria, 2016.

[9] L. I Eleje,., N. P. M Esomonu,., N. N Agu,., R. O Okoye,., E.Obasi and E. F Onah,., "Development and validation of diagnostic economics test for secondary schools*". World Journal of Education,vol. 6*, no.3, pp90-112,2016.

[10] A. L Chanrasegaran,., D. F.Treagust and M.Mocerino,. "The development of a two tier multiple choice diagnostic instrument for evaluating secondary school students ability to describe and explain chemical reaction using multiple levels of representation". *Chemistry Education Research and Practice, vol.8*, no.3, pp 293-307, 2007.

[11] D. K. Tan, K. Tabe, , N.Goh and B. Chia, "Development and validation of ionization energy diagnostic instrument". *Chemistry Educational Research Practice, vol.6, no.*4, pp180-197, 2005.

[12] H Innabi and H. Dodeen, "Gender differences in mathematics and science achievement in Jordan: A differential item functioning analysis". *Journal of School Science and Mathematics*, vol. 3, no.4, pp127-137, 2018.

[13] D. Ojerinde, "*Classical test theory vs item response theory: An evaluation of the comparability of item analysis results".* A paper presented at the Institute of Education University of Ibadan, 2013.

[14] A. O. Ayiro, "*Introduction to test theory and item analysis".* A paper presented at ASSEREN conference at the University of Jos.2017.

[15] H. Hardiyanti, B. Mansyur and O.Munawwarah, " Differential item functioning of National Standard School Examination for chemistry subject through Lord's Chi -Square method based on gender of students in Indonesia*". Proceedings of ISER International Conference*, Tokyo, Japan, 2018.

[16] N. P. M Esomonu and L. I. Eleje, " Diagnostic quantitative economics skill test for secondary schools: Development and validation using item response theory". *Journal of Education and Practice, vol.8,*no.22, pp110-125, 2017.

[17] S. I Dadughan, "*Development and calibration of primary school mathematics diagnostic test based on item response theory".* A Ph. D Thesis submitted to University of Nigeria, Nsukka, 2015.

[18] B. A Young, "*Development and validation test in English Language for secondary school in Kisumu Municipality using item response theory*".Unpublished M.Ed. Thesis University of Kassel, Wizenhausen, Germany, 2014.

[19] W. P, Chang, " *Development and validation of diagnostic test in economics for secondary school in Far Province, Iran using item response theory"*. Unpublished M.Ed. Thesis.Arsanya University, Iran. 2013.
[20] O. S.Latun, " *Development and validation of diagnostic test in physics for secondary school students in Limpopo Province of South Africa using item response theory".* Unpublished M.Ed. Thesis. University of Pretoria South Africa, 2011.
[21] M. D.Reckase, "*Multidimensional item response theory"*. London: Springer, 2009.
[22] L. K Muthen and B. O.Muthen, "*.Mplus user's guide".* Los Angeles, CA: Muthen and Muthen Co, 2007.
[23] O. O Adedoyin, , "*Investigating the invariance of person parameter estimates based on classical test and item response theory"*. Retrieved from http://www.uniBotswana./journal/ education/science. 2010.
[24] F. B. Baker, "*The basics of item response theory"* .NewYork: ERIC. 2001.
[25] D. Harris, "*Educational measurement issues and practice: Comparison of 1-,2-, and 3-parameter IRT models"*. Retrieved from 10.1111/j.1745-3992.1989.tb00313.x, 2005.
[26] A. D. E,Obinne, "A psychometric analysis of two major examinations in Nigeria: Standard error of measurement". *International Journal of Education Science, vol.3*,no.2, 137-144, 2011.
[27] C. A. Stone and B .Zhang, "Assessing goodness of fit of item response theory models. A comparison of traditional and alternative procedures". *Journal of Educational Measurement*, vol.*40, no.4*, pp331-532, 2003.
[28] I. A Kose, , "Assessing model data fit of unidimensional item response theory models in stimulated data". *Journal of Education and Review*,vol. *9,* no.17, pp642-649, 2014.
[29] S, Sinharay, ".Assessing fit of unidimensional item response theory models using bayesian approach". *Journal of Educational Measurement*, vol. 42, no.4, pp375-394, 2005.
[30] A. D. E Obinne, "*Test item validity: Item response theory perspective for Nigeria".* Retrieved from www.emergingresource.org, 2013.
[31] D. G.,Meredith, P. G .Joyce and R B. Walter, "Educational *research: An introduction (8th ed.*)". United State of America: Pearson Press, 2007.
[32] M., Chatterji, "*.Designing and using tools for educational assessment*". Retrieved fromhttp://www.columbia.edu/~mb1434/EdAssess.htm. 2003.

PROF.NKECHIPATRICIA-MARY ESOMONU, et. al. "Development and Validation of Geography Diagnostic Test for Senior Secondary School Students Using Item Response Theory." *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, 26(11), 2021, pp. 01-09.