International Journal of Advanced Education
www.allpaperjournal.com
Online ISSN: 3107-9601
Received: 09-11-2025, Accepted: 08-12-2025, Published: 23-12-2025
Volume 1, Issue 2, 2025, Page No. 65-70

Development and validation of geography aptitude test for secondary school students using Item Response Theory

Erutujiro Goodluck[1], Onuoha Obinna I[2]
[1] Department of Education Foundations, Nnamdi Azikiwe University, Awka, Nigeria
[2] Department of Education Psychology, Alvan Ikoku Federal University of Education, Owerri, Nigeria

Abstract
Aptitude test assess knowledge, skills and other characteristics that serve to predict learning and success. Unfortunately, there is scarcity of geography aptitude test. The major concern of this study is how to develop and validate geography aptitude test for secondary school students using item response theory. Six research questions guided the study. The design of this study is instrumentation research design. The study was conducted in Delta State, Nigeria. The population of the study is made up of 3,952 SS 1 students offering geography in 2022/2023 academic session in Delta State. The sample for the study consisted of 1400 SS1 students drawn using simple random sampling technique. The initial instrument was developed by the researcher was made up of 60 items drawn from SS1 Geography curriculum. The instrument was administer through the help of research assistants. The assumption of dimensionality was assessed using Dimtest statistic in DIMPACK while local independence of the test items were checked using Yen Q3 statistics. Item parameters were calibrated using Maximum Likelihood estimation in Mplus Version 7. The researcher found out that the instrument is multidimensional and the assumption of local independence of the test items was not violated. Also, the instrument has both convergent and discriminant validity. The entire scale has empirical reliability of 0.81.

Keywords: Validity, reliability, aptitude test, geography, Item response theory

Introduction

Aptitude testing of students is a very important aspect of assessment that should not be overlook by teachers and other stakeholders in education system. Aptitude testing is a great means of refining the thought process and skills of memory. These tests are used to find an individual's potential to get through a certain task, with no previous knowledge or training. Usually, the people who have thirst for research construct tests in areas where they are much more comfortable. Aptitude testing according to Itsuokor (1995) [12], embraces intelligence and achievement, personality, interest and other skills which predispose one to learning. On this basis, Bennett (2015) [4] sees aptitude as an ability to learn, and aptitude test assess knowledge, skills and other characteristics that serve to predict learning and success. According to Bokander and Bylund (2020) [6] aptitude tests cover wide range of human behaviour among which is mechanical aptitude, clerical aptitude, musical aptitude and artistical aptitude. According to Cronbach (1970) [9], aptitude test is a proficiency test that measures ability to perform some task significant in its right. An aptitude test in this regard, is intended to predict success in some occupation or course of training. Cronbach also reported that there are many types of aptitude, musical aptitude, reading aptitude, clerical aptitude, spatial aptitude and numerical aptitude.

Aptitude test commonly used for school educational and vocational counselling is the Differential Aptitude Tests (DAT) which comprises eight distinct sections – verbal reasoning, space – relations including visualization and perception, clerical spread and accuracy, mechanical reasoning, language, spelling and usage (Berk, 1982) [5].Aptitude test is used to investigate not only what a student can do but also what he does. An aptitude test is an instrument used to measure and determine an individual's ability to acquire future training in some specific set of skills. According to Kagan (2019) [13], an aptitude test is used to determine an individual's propensity to succeed in a given activity, and that, aptitude tests assume that individuals have inherent strengths and weaknesses with natural inclination towards success or failure in a specific area based on their innate characteristics while differential aptitude test is a multiple aptitude battery designed to measure ability to learn or succeed in a certain area (Odo & Ugworji, 2016; Wang, 2011) [14].

Despite the aforementioned the benefits of aptitude testing there are scarcity of aptitude tests in various subject in most of our school today. The implication of the above is that many important decision in aptitude testing in most subjects including geography are taken wrongly. The reliability and validity of most assessments tools have not yet lived up to expectation. The main practical and technical problem with assessment is finding suitable criterion measures to provide validity evidence from learning potential measures. It therefore imperative to develop a valid and reliable geography aptitude test for secondary school students.

The construction and validation of any test can be approach using either the Class Test Theory or Item Response Theory. The Classical Test Theory In 1904 Charles Spearman showed how to correct correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in making the correction. Spearman's demonstration marked the beginning of classical test theory (Adams, 1998). Subsequently, the framework of classical test theory was elaborated and refined by George Udny Yule, Truman Lee Kelley, and others over the quarter century or so following 1940. Another milestone was laid in 1937 with the publication of the Kuder Richardson formulas. The fundamental principle of classical test theory is that the score that a candidate

obtains on a given examination (observed score) which is symbolized by (*X*), can be classified into the person's true score (*T*) and a random error component (*E*) which is given below

$$X = T + E$$

The candidate's true score, *T*, is defined as the expected value of the observed score over an infinite number of repeat administrations with the same examination. A true score can be thought of as the score that would be obtained if the examination was perfectly measuring the ability of interest (i.e. with no measurement error). A reliability coefficient can provide us with an estimate of the level of concordance between observed and true scores (Olatan, 2005). An observed score is a score got by an examinee during test administration or assessment. It is assumed that each examinee or testee has a true score which would have been gotten in test administration.

The main merits of Classical Test Theory over other test theories is in its simplicity in constructing test for examinees and it can be carried out using smaller samples of examinees. Furthermore, one of benefits of CTT is that candidate's personality and test characteristics cannot be separated, each may simply be explained in the framework of the other. CTT basically deals with the correlation connecting the aforementioned three variables (i.e observed test score, true score and error score) (Sidhu, 2007) [20]. It is by this kind of connection/relationship that test administrators, school teachers or any test experts will use in saying something concerning the characteristics of a the test and test scores. Hence, Classical Test Theory is aimed at giving much preference to test reliability to improve the reliability of psychological tests and assessment.to overcome the above short coming item response theory was development. However, the major item parameters (difficulty, discrimination, distracter indices and reliability) are sample dependent. The sample dependency limit the usefulness of item parameters obtained under CTT. To overcome Item response theory was developed.

Item response theory is psychometric theory that may be regarded as roughly synonymous with latent trait theory. It is also referred to as the strong true score theory or modern mental test theory since IRT is the most recent body of theory with stronger assumption than classical theory. Item response theory originally developed for items that are scored dichotomously -that is, correct or incorrect (Ezechukwu, Oguguo, ENE, & Urgoji, 2020) [11]. The concept and method of IRT extend to a wide variety of polychotomous models for all types of psychological variables that are measured by rating scales of various kinds (Chime, 2012; Vander & Hambleton, 1997) [7, 23].

There are three major assumptions of Item response theory. First assumption of item response theory is dimensionality. The means the probability of responding to an item is dependent on the underlying latent ability. This assumption is the basis of all measurement theory to the extent the sum of item scores is used to assign some overall value of ability to an examinee, as is the case on most tests (Anywale, Isaac-Oloniyo & Abayomi (2020). A second related but separate assumption central to IRT is Local Item Independence (LII), meaning the response to each item is not influenced by the response to any other item. In other words, LII is achieved if examinees" respective ability value (θ) explains fully their performance on all items. Assumption about Item Characteristics Curve (ICC) is the relationship between examinee"s latent ability and probability of the examinee correctly responding to certain item (Okwilagwe & Ogunrinde, 2017) [16]. This relationship is modeled by a mathematical function called item characteristic function and the graph of this function is called Item Characteristic Curve (ICC). In simple terms, it is a linear or nonlinear function for the regression of item score on the trait or ability measured by the test (Okwilagwe & Ogunrinde, 2017) [16]. The assumption can be justified by how well the chosen IRT model accounts for the test data.

The major item parameters obtained under item response theory are; difficulty, discrimination, guessing and carelessness. Chong (2013) [8] reported that difficulty parameter or the threshold parameter value tells how easy or how difficult an item is. The findings which revealed that the items were selected based on the b-value range of -3 to +3 corresponds with (Baker, 2001) [3] that theoretically, difficulty values can range from - 00 to + 00, in practice, difficulty values usually are in the range of - 3 to + 3. Baker also described the range of values for item discrimination as follows: very low, 01 -. 34, Low, 35 -. 64, moderate, 65 - 1.34 High, 1.35 - 1.69 and Very high, 1.70 and above. Discriminating parameter indicates how well an item discriminate between respondents below and above the item threshold parameter, as in the slope of the item characteristics curves (Reeve & Fayers, 2005). Adedoyin (2010) [1] in his study used chi-square test with probability greater than alpha level of 0.05 significant level to select items that fit model

Researcher like Akaezue (2009) [2] carried out a study on construction and validation physics aptitude test in senior secondary school in Sapele Local Government Area of Delta State. The test has content validity; a reliability of 0.75 was obtained through the use of Kuder – Richardson formula 20. The test has a standard error of measurement of 3.15.The difficulty level of the test ranges from 0.325 to 0.855 while the discriminatory level ranges from 0.15 to 0.48. The researcher found out that the test has high content validity, the items of the test are suitable in terms of its difficulty and discriminatory indices and also the test and its entire subsections have high degree of internal consistency with a low standard error of measurement. Similarly, Osadebe and Nwabeze (2018) [18] construct and validate a Physics Aptitude Test (PAT) for senior secondary schools. The findings revealed that the Physics Aptitude Test (PAT) has adequate face and content validities. It is made up of 50 items. The difficulty and Discrimination indices were appropriate because they are within the standard range of indices for the test. Difficulty indices range from 0.30 to 0.70, Discrimination indices range from. 30 to 0.44. A reliability coefficient of 0.94 was obtained through Kuder Richardson formula 20 as a measure of internal consistency. Since the Physics Aptitude Test (PAT) was highly valid and reliable, it was recommended that it should always be used as an assessment tool for determining the aptitude of senior secondary school students in Physics. Furhermore, Usha, Vijaya and Rajeshkumar (2022) [22] study focused on the construction and validation of mathematical aptitude test. The prime objective of the study was to construct and standardize an aptitude test in Mathematics for Standard IX students. For this purpose, the investigator used the survey method. Fifty items of objective type were prepared. The test was administered to a sample of three hundred students in Thoothukudi. The content validity was found by the

opinion of experts. The reliability coefficient was found to be 0.77.

Literature review showed that there scarcity of geography aptitude test that can used by teachers to obtained students aptitude in the subject for important critical decisions making by teachers, counsellors, and principals. Developing a geography aptitude test will also guide researchers who may be interested in developing similar instrument in other subjects for measuring aptitude. Therefore, the major concern of this study is to develop and validate geography aptitude test for secondary school students using Item Response Theory

Research Questions

The following research questions will guide the study:

1. What is the dimensionality and local independence of the geography aptitude test items?
2. What IRT model best fit the data of geography aptitude multiple choice test items?
3. What are the item difficulty indices of the geography aptitude test?
4. What are the item discrimination indices of the geography aptitude test?
5. What are the item guessing indices of the geography aptitude test?
6. How does the instrument converge and diverge with other constructs?
7. What is reliability of the of the geography aptitude test items?

Method

The design of this study is instrumentation research design. The study will be conducted in Delta State, Nigeria. The population of the study is made up of 3,952 SS 1 students offering geography in 2022/2023 academic session in Delta State. This comprised 4,623 females and 3,329 males (Source: Delta State Post Primary School Service Commission, Asaba, 2023).The sample for the study consisted of 1400 SS1 students drawn using simple random sampling technique.

The instrument was developed by the researcher. The instrument is made up of items drawn from SS1 Geography curriculum. It was generated in such a way that the cognitive aspects of the students' behaviour were included, and various aspects/component of aptitude test were reflected such as numerical aptitude which is in the first section and it has a total of fifteen (15) questions, followed by the verbal section which also has a total of twenty (24) questions. Next to the verbal aptitude section is the quantitative aptitude section which has a total of eleven (11) questions. The last section of this test is the mechanical aptitude test section which also has a total of ten (10) questions. The items in the scale has 1 to 0 scoring format.1 for correct score and 0 for wrong score. The scale was presented to one experts in Measurement and Evaluation in Nwafor Orizu College of Education, Nsugbe as well as two geography teachers at the secondary school level for validation. Thereafter, the instrument was administered to the sampled students with the help of research assistants in the sampled schools as they were guided by the researcher on the administrative procedure. The collected data were using Mplus Version 7.

Results

Research Question 1: What is the dimensionality and local independence of the geography aptitude test items?

To answer research question one, two basic assumptions of item response theory- dimensionality and local independence were examined. The dimensionality assumption was investigated using DIMTEST statistics.

Table 1: Dimtest Statistics of Geography Aptitude multiple choice test items

TL	TGbar	T	AT	PT	P-value
19.2871	4.9126	15.2807	11	49	0.0000

The result in table 1 indicates that Geography Aptitude multiple choice test items is multidimensional since p-value is <0.05 level of significance. Furthermore, if the difference between the number of items in Partitioning Subtest (PT) and the Assessment Subtest (AT) in a test is significant there is evidence of multidimensionality (Anywale, Isaac-Oloniyo & Abayomi, 2020). For this particular study, the difference between the AT and PT as shown in table 1 is significant ($T=15.2807$, $p<0.05$). This led to the conclusion that the AT items were dimensionally distinct from the remaining items in PT. Therefore, multidimensionality is manifest in Geography Aptitude multiple choice test items.

Local independence assumption was investigated with the Yen's Q3 statistics. Based on this statistics, residuals for any pair of items should be uncorrelated, and generally close to zero. Residual correlations that are high indicate a violation of the local independence assumption, and this suggests that the pair of items have something more in common than the rest of the item set have in common with each other (Marais, 2013). For this study, using Yen's Q3 to screen items for local dependence, 85% item residual correlations were below absolute value of 0.2. This indicates that the local independence assumption of the IRT was not grossly violated.

Research Question 2: What IRT model best fit the data of Geography Aptitude multiple choice test items?

Table 2: IRT Model Best Fit Of Geography Aptitude Multiple Choice Test Items

Model	AIC	-2Loglik
1PL	231533.5	231321.4
2PL	231534.5	231324.4
3PL	230608.8	230018.8
4PL	230630.4	230180.0

The examinee responses were subjected to full information item factor analysis, and compared using Akaike Information Criterion (AIC), and Likelihood Ratio Test (LogLik) in order to establish the best fit model that provided the information for the calibration of item parameters embedded in the test data. The result in table 2 indicates that 3PLM has the smallest information criteria in terms of Akaike information criteria and -2Loglik. Therefore, the 3PLM was used to calibrate the IRT-based parameters estimate of Geography aptitude multiple choice test items.

Question 3: What are the item difficulty indices of the geography aptitude test?

Table 3: Item Difficulty of the Geography Aptitude Test

Item	B	Item	B	Item	b	Item	b
1	0.14	16	-3.90	31	1.19	46	2.22
2	2.61	17	2.17	32	2.11	47	2.81
3	2.71	18	2.81	33	2.41	48	-2.71
4	2.11	19	0.69	34	2.71	49	-2.11
5	-4.21	20	1.17	35	1.09	50	0.11
6	2.10	21	1.81	36	0.14	51	-3.08
7	3.81	22	2.01	37	-1.21	52	4.67
8	-3.61	23	2.71	38	-2.17	53	1.12
9	0.11	24	-1.81	39	0.14	54	-3.95
10	0.23	25	-1.31	40	5.21	55	1.34
11	0.15	26	2.11	41	0.17	56	-1.24
12	1.21	27	2.81	42	1.22	57	-2.14
13	2.14	28	-2.21	43	1.31	58	1.22
14	2.86	29	-2.11	44	-4.07	59	1.71
15	1.31	30	-4.41	45	1.67	60	-2.47

Table 3 shows that fifty (50) items, that is Item 1, 2, 3, 4, 6, 9, 10, 11, 12, 13, 14, 15, 17, 18, 9, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 45, 46, 47, 48, 49, 50, 53, 55, 56, 57, 58, 59 and 60 fall between +3 to -3 of the latent ability logit scale and were accepted. While item 5, 7, 8, 15, 16, 40, 44, 51, 52 and 54 were rejected as their values fall below -3 and above + 3on the logit ability scale.

Research Question 4: What are the item discrimination indices of the geography aptitude test?

Table 4: Item Discrimination of the Geography Aptitude Test

Item	a	Item	A	Item	a	Item	A
1	2.14	16	2.71	31	4.21	46	2.81
2	-3.14	17	2.83	32	4.01	47	3.91
3	3.14	18	3.01	33	3.27	48	2.91
4	4.21	19	3.42	34	3.14	49	0.00
5	4.82	20	3.07	35	4.22	50	2.75
6	2.95	21	0.00	36	3.61	51	2.81
7	3.11	22	0.14	37	2.91	52	3.01
8	2.86	23	-2.14	38	2.77	53	0.10
9	3.14	24	0.00	39	2.81	54	4.01
10	3.01	25	2.74	40	3.01	55	4.22
11	2.92	26	2.88	41	3.91	56	3.77
12	2.70	27	3.19	42	-4.22	57	3.01
13	3.19	28	3.70	43	3.33	58	2.99
14	3.81	29	2.78	44	4.22	59	2.78
15	2.96	30	2.94	45	4.31	60	3.37

Table 4 shows that item 2, 23 and 42 discriminate in a negative direction and were rejected. Item 21, 24 and 49 do not discriminate and were rejected. Item 22 and 53 have low discrimination capacity and were rejected. Item 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59 and 60 have their discriminating values above 2.0 indicating high discriminating values.

Research Question 5: What are the item guessing indices of the geography aptitude test?

Table 5: Item Guessing Parameter of the Geography Aptitude Test

Item	c	Item	C	Item	c	Item	C
1	0.04	16	0.37	31	0.04	46	0.01
2	0.31	17	0.04	32	0.02	47	0.02
3	0.00	18	0.02	33	0.01	48	0.00
4	0.03	19	0.00	34	0.00	49	0.02
5	0.27	20	0.00	35	0.03	50	0.04
6	0.02	21	0.67	36	0.04	51	0.62
7	0.41	22	0.37	37	0.32	52	0.77
8	0.23	23	0.64	38	0.11	53	0.34
9	0.00	24	0.43	39	0.00	54	0.60
10	0.01	25	0.01	40	0.77	55	0.04
11	0.00	26	0.02	41	0.00	56	0.00
12	0.03	27	0.03	42	0.30	57	0.00
13	0.04	28	0.04	43	0.01	58	0.00
14	0.03	29	0.00	44	0.61	59	0.02
15	0.19	30	0.00	45	0.03	60	0.77

Table 5 shows that the guessing values (c-values) of the items range from 0.00 to 0.77. The table 5 also indicates Item 1, 3, 4, 6, 9, 10, 11, 12, 13, 14, 17, 18, 20, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 39, 41, 43, 45, 46, 47, 48, 49, 50, 55, 56, 57, 58, and 59 guessing values are below. 05 and were accepted. While the remaining items 2, 5, 7, 8, 15, 16, 40, 37, 38, 44, 51, 52, 54, 22, 23, 24, 42, 47, 49, 53 and 60 fall above. 05 which indicates that the items are not very good and the probability of guessing the items correctly is very high and were rejected.

Research question 6: How does the instrument converge and diverge with other constructs?

To answer research question 3, the items (40 items) that survive item analysis based on difficulty, discrimination and guessing were assembled and administers another sample of 500 students. In addition, WAEC Geography, 2022 multiple achievement test and social phobia scale were also administered. A multitrait mono method correlational analysis was carried out to find out the relationship between scores of the geography aptitude test scale, geography achievement test and social phobia scale. Variables that converge highly correlate, while those that diverge will naturally will not correlate among themselves. In this study, the results as presented in Table 6 showed that the geography aptitude test showed a positive correlation with the geography achievement test (r =. 86) demonstrating evidence of convergence validity but showed an insignificant negative relationship with the self-esteem scale (r =-.06) demonstrating evidence of divergence validity. More so, the Social phobia showed an insignificant negative relationship with the geography achievement (r = -.11).

Table 6: Multi-trait mono method correlational analysis of GAT, GATS and SPS

Variables	GAT	GATS	SPS
GAT	1		
GATS	.86	1	
SPS	-.06	-.11	1

Geography Achievement Test (GAT)
Geography Aptitude Test Scale (GATS)
Social Phobia Scale (SPS)

Research Questions 7. What is reliability of the of the geography aptitude test items?

The reliability of the geography aptitude test scale and its sub-scale dimensions were presented in table 7 for discussions.

Table 7: Reliability of the Final Instrument and Its Sub-Scale Based On Empirical Reliability

Scale Dimension	No of items	Items	Empirical Reliability
Numerical Sub-scale	10	1, 3, 4, 6, 9, 10, 11, 12, 13, 14	.89
Verbal Sub-scale	16	17, 18, 19, 20, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35 36, 39	.78
Quantitative Sub-scale	8	41, 43, 45, 46, 47,48, 49, 50	.86
Mechanical Sub-scale	6	54, 55, 56, 57, 58, 59,	78
Entire Scale	40		.81

Table 7 show the reliability of the overall scale as well as sub-scale dimensions of the construct. Empirical reliability coefficients for geography aptitude test scale was calculated as 0.81 for the scale total score with 40 items, 0.78 for the numerical sub-dimension 1 with 10 items, and 0.86 for the verbal sub-dimension 2 with 16 items. 0.86 for the quantitative sub-dimension 3 with 8 items. While 0.86 for the quantitative sub-dimension 3 with 6 items.

Discussion of Findings

The researcher found out the instrument is multidimensional. This implies that the various dimensions of geography aptitude test: numerical, verbal, quantitative and mechanical sub-dimensions were adequately represented in the instrument. The above findings in line with The finding agreed that with the study of Anywale, Isaac-Oloniyo and Abayomi (2020) study on assessment of dimensionality of Osun State unified Mathematics Achievement Test items is multidimensional in nature. Findings from the work of Oguoma, Metibemu and Okoye (2016) [15] on dimensionality assumption test on 2014 Mathematics achievement items of West African Senior Secondary Certificate Examination (WASSCE) concluded that the test items of WASSCE mathematics were inherently multidimensional in nature. Furthermore, Okwilagwe and Ogunrinde (2017) [16] also found that fifty (50) items of 2013 WASSCE and sixty (60) items of National Examinations Council (NECO) Geography respectively violated assumption of unidimensionality and that there were more than one dimension that accounted for the variation observed in examinees to the geography test items.

In term of local independence, the researcher also found out that the assumption was not violated using Yen Q3 statistic. The finding also agreed that of Ubi (2006) [21] study that assessed 800 scripts of candidates who wrote Mathematics in the University Matriculation Examination in Cross River State, Nigeria and discovered that a good number of items in the examination were locally independent.

Furthermore, the study shows that fifty (50) items fall between +3 to -3 of the latent ability logit scale and were accepted. The researcher also found out 52 items have their discriminating values above 2.0 indicating high discriminating values. Forty (40) items have guessing values are below. 05 and were accepted. This finding agrees with the finding of Chime (2012) [7], who developed and validated an Economics Achievement Test and found that the test had high psychometric properties in terms of facility and discrimination index. According to Sidhu (2007) [20], the discrimination index of an item indicates whether or not the item is measuring the same ability as the test measures and shows how well an item discriminates between the able and weak students. The above findings are in line with earlier studies with that of Odo and Ugwoji (2016) [14], who developed and validated Biology Achievement Test (BAT) for Assessment of Students in Enugu State. They used Kuder Richardson formula 20 for the reliability and obtained a coefficient of 0.88. This value is similar to value of reliability coefficients obtained by Adams (1998) which was 0.96. Also, Olaitan (2005) [17] stated that tests with reliability coefficient of 0.70 and above are considered sufficiently reliable to be of practical use.

Conclusion

The geography aptitude test is valid and reliable with forty (40) items. Furthermore, the various dimensions of geography aptitude test-numerical, verbal, quantitative and mechanical sub-dimensions were adequately represented in the instrument. The aforementioned sub-scales are valid and reliable.

Recommendations

The researcher wish to make the following recommendations:

1. Geography teachers should use the instrument to assess students' readiness to learn the subject in secondary school.
2. Secondary school counsellors should use the instrument developed to assess readiness to learnt geography and counsel students' choice of geography as a school subject.
3. Researchers should be trained on how to apply item response theory correctly in development and standardization of instrument
4. Government should provide funds to assist researchers who wants to develop and validate instruments for students' assessment in schools.

References

1. Adedoyin OO. Investigating the invariance of person parameter estimates based on classical test and item response theories. International Journal of Educational Science,2010:2(4):70–78. Retrieved November 30, 2012 from http://www.uniBotswana./journal/education/ science.
2. Akaezue NN. Construction and validation of physics aptitude test for senior secondary school student in delta state. Unpublished M.Ed. project. Delta State University, Abraka, Delta State, 2009.
3. Baker FB. The basics of item response theory. New York, NY: ERIC Clearinghouse, 2001.
4. Bennett RE. The changing nature of educational assessment. Review of Research in Education,2015:39(1):370–407. http://doi.org/10.3102/0091732X14554179.
5. Berk RA. Handbook for detecting test bias. Baltimore: John Hopkins University Press, 1982.
6. Bokander L, Bylund E. Probing the internal validity of the LLAMA language aptitude tests. Language Learning,2020:70(1):11–47. https://doi.org/10.1111/lang.12368.
7. Chime UM. Development and validation of economics achievement test for senior secondary school students. Unpublished Master's Thesis. Department of Science Education, University of Nigeria, Nsukka, 2012.
8. Chong HY. A simple guide to the Item Response Theory (IRT) and Rasch modeling. Retrieved March, 2013 from http://www.creativewisdom.com, 2013.
9. Cronbach LJ. Essential of psychological testing. New York: Steal Press, 1970.
10. Egbule JF. Principles of psychological testing. Owerri: Whyte and Whyte Publishers, 2002.
11. Ezechukwu RF, Oguguo BC, ENE CU, Urgoji CO. Psychometric analysis of economics achievement test using item response theory. World Journal of Education,2020:10(2):59–68.
12. Itsuokor DE. Essentials of test and measurement. Illorin: Woye and Sons (Nig) Ltd, 1995.
13. Kagan J. Aptitude test. Retrieved on the 17 June, 2020 from Investopedia.com, 2019.
14. Odo CI, Ugwoji JN. Development and validation of biology achievement test for assessment of students in Enugu State. International Journal of Innovative and Applied Research,2016:4(5):23–28.
15. Oguoma RO, Metibemu CC, Okoye MA. An assessment of the dimensionality of 2014 West African secondary school examination mathematics objective test scores in Imo State, Nigeria. African Journal of Theory and Practice of Educational Assessment,2016:4:18–33.
16. Okwilagwe MA, Ogunrinde EA. Assessment of unidimensionality and local independence of WAEC and NECO 2013 geography achievement tests. African Journal of Theory and Practice of Educational Assessment,2017:5:31–44.
17. Olaitan SO. Understanding curriculum. Nsukka: Ndudim Press, 2005.
18. Osadebe PU, Nwabeze PC. Construction and validation of physics aptitude test as an assessment tool for senior secondary school students. International Journal of Assessment Tools in Education,2018:5(3):461–473.
19. Shogbesan YO, Ogunsakin IB. Item response theory (IRT): A modern statistical theory for solving measurement problem in 21st century. International Journal of Scientific Research in Education,2018:11(3):627–635.
20. Sidhu KS. New approaches to measurement and evaluation. New Delhi, India: Sterling Publication Pvt. Ltd, 2007.
21. Ubi IO. Item local independence, dimensionality and trend candidates' mathematics performance on University Matriculation Examination in Nigeria. Unpublished Ph.D Dissertation. University of Calabar, Nigeria, 2006.
22. Usha G, Vijaya R, Rajeshkumar M. Construction and validation of mathematical aptitude test for secondary level learners. Journal of Positive School Psychology,2022:6(6):2883–2887.
23. Vander LWJ, Hambleton RK. Handbook of modern item response theory. New York: Springer-Verlag, 1997. https://doi.org/10.1007/978-1-4757-2691-6.
24. Wang L. The differential aptitude test: A review and critique. Annual meeting of Southwest Educational Research Association, Austin, Texas: January 29, 1993.