European Journal of Open Education and E-learning Studies
ISSN: 2501-9120
ISSN-L: 2501-9120
Available on-line at: www.oapub.org/edu

DOI: 10.46827/ejoe.v8i3.5054 Volume 8 | Issue 3 | 2023

ASSESSMENT OF THE PSYCHOMETRIC PROPERTIES OF ATTITUDE TOWARDS ASSESSMENT TEST (ATAT)

Megbele, A. M.[1i],
Odili, J. N.[2],
Osadebe, P. U.[2]
[1]Doctoral Student,
Measurement and Evaluation,
Delta State University,
Abraka, Nigeria
[2]Dr., Professor of Measurement and Evaluation,
Delta State University,
Abraka, Nigeria

Abstract:
This study is on the assessment of the psychometric properties of Attitude Towards Assessment Test (ATAT). Four research questions guided the study. An instrumentation research design was adopted. The population consisted of secondary school students in Delta State, Nigeria. The sample size comprised 1,000 students, selected through simple random and non-probability cluster sampling techniques. The test under assessment was developed by Megbele, et al. (2023). The Rasch Rating Scale Model was used to answer research 1, which assessed person and item reliability, item statistics, and ordering of response categories. Two fit indices including the infit and outfit mean square (MNSQ) statistics were used to answer research question 2. The data that were used to answer research questions 1 and 2 were analysed with the aid of the Jmetrik IRT software. The Categorical Confirmatory Factor Analysis (CCFA) was used to answer research question 3 while Chi-Square Goodness of Fit Statistics was used to answer research question 4 on construct validity, evidence of unidimensionality, and local independence respectively. The findings of this study revealed that the three components of the scale had high values of item separation index and reliability as well as an acceptable range of Person separation index and reliability; the difficulty index of items in the test was within an acceptable range; Each of the components of the Attitude Towards Assessment Test (ATAT) (Cognitive, affective and behavioural) had one construct each, which is evidence of unidimensionality for the different components of the Attitude Towards Assessment Test. Based on the findings of the study, it was concluded that all items in the different components that made up the test are reliable, have adequate item difficulty infit, and outfit MNSQ estimates, with evidence of unidimensionality and local independence

[i] Correspondence: email megbele.andrew@delsu.edu.ng

assumptions. The study recommended that the test can be used by examination bodies for the assessment of students in the affective domain.

Keywords: psychometric properties; attitude towards assessment test; categorical confirmatory factor analysis; unidimensionality

1. Introduction

The continuous assessment mode of evaluation currently used in the educational system of Nigeria was introduced to take care of the lapses that characterized the traditional mode of assessment. Some of these lapses are inherent in the domain of assessment. The focus of the traditional mode of assessment was on the cognitive domain of learning, such that at the end of a particular term or school session, students are given a set of questions aimed at assessing the extent to which they have learned a particular school subject. This system of assessment was inherited from the colonial educational system, which based assessment only on the cognitive domain at the expense of the affective and psychomotor domains.

The exclusive focus on the cognitive domain also meant that the instrument used for the assessment was limited to multiple-choice and essay test questions. No room was given to such assessment tools as observation, checklist, and rating scales., which are instruments for measuring the affective and psychomotor domains of learning, covered by the continuous assessment mode of assessment. Assessing the extent to which students have obtained learning in the cognitive, affective, and psychomotor domains makes learning to be comprehensive. This is because, students are not only expected to understand the contents of what is being learned, they are also expected to appreciate learning and practise what was learned. Behaviour change (learning) can only be validated when students are able to practise what they have learned. It is therefore important that the three domains of learning should be systematically and comprehensively assessed during and after learning. The focus of this study is on the affective domain of learning.

The affective domain helps in the process of exploring and adapting human interests, attitudes, values, and appreciation. Affective learning outcomes cannot easily be quantified by traditional testing and rather relies on qualitative self-reflection. The taxonomy of the affective domain contains five levels, from lowest to highest: receiving, responding, valuing, organization, and characterization. This taxonomy was applied to written self-evaluations to assess changes in affective learning. They include receiving, responding, valuing, organisation, and characterization.

In line with the above taxonomies of the affective domain of learning, various attributes can be assessed. These attributes include (but are not limited to) attitude towards education, academic self-efficacy, academic motivation, career aspiration, and test anxiety. No single assessment tool can be used to assess all these behaviours. Hence, the focus of this study is on attitude. As important as students' attitude towards assessment on the outcome of the assessment, a search through the literature revealed

that to the best knowledge of the researcher, no assessment tool exists for the measurement of students' attitude towards assessment, particularly in the context of students in Delta State Nigeria. This is the crux of the study, to assess the psychometric properties of the Attitude Towards Assessment Test (ATAT). The psychometric properties include item person separation indices, item difficulty infit and outfit MNSQ statistics, unidimensionality, and local dependence.

Unlike classical test theory that requires another test for reliability, IRT has a local reliability. That is, an amount of information at each point of underlying continuum (Ceniza and Cereno, 2012). With IRT model, each item of the test contained information. For each parameter logistic model, Bilog MG computed an empirical reliability index. A reliability index within the range of 0.81 to 1.0 indicates high reliability; 0.61 to 0.80 shows a moderate reliability; 0.41 to 0.60 means fair reliability; 0.10 to 0.40 means slight reliability and less than 0.10 means virtually no reliability.

According to Maydeu-Olivares, et al. (2011), reliability is the precision of measurement using the ratio of true and observed score variance, equalling the test's average ability. However, there is a fault in this definition due to the fact that reliability is not uniform across the entire range of test scores. For example, students scoring at the high end of the upper level of ability and students scoring at the low end of ability have more variance in their standard error of ability. The scores centred near the mean have less error. Error in this case is not equally distributed among the distribution of scores.

In measuring latent traits, such as ability, item characteristic curves can be modelled for each individual item, showing the item's difficulty and discrimination. While measuring this trait it is necessary to chart a student's ability, this scale can go anywhere from negative infinity to positive infinity with a midpoint of zero and a unit measurement of 1. For practicality in scale construction, the examples in this paper are limited to a range of –3 to +3 (Muis, et al., 2009).

In Item Response Theory, fit (infit and outfit) statistics are used in order to detect the discrepancies between empirical data and the Rasch model prescriptions (Penfield, 2014). By statistically indicating the degree of match between the observed performance and expected performance, fit statistics report how well the empirical data accord with the Rasch model (Linacre, as cited in Sharkness & DeAngelo, 2011). Routinely, fit statistics are reported in both an unstandardized and a standardized form. An unstandardized form refers to mean square (MNSQ) and a standardized form is standardized t (ZSTD).

A fit MNSQ value provides information about "*how confident we can be in the measures (logits) associated with the persons and the items*" (Green, 2013, p. 167). Depending on the MNSQ value, whether items or persons fit the Rasch model or not can be judged. The acceptable MNSQ value of a person and an item ranges from +0.5 to +1.5, which is considered to be productive for measurement (Green, 2013). On the other hand, all data whose MNSQ values are not between +0.5 and +1.5 are classified as misfit data, indicating that the data do not fit the Rasch model.

If MNSQ value is less than +0.5, it means that a person or an item is performing in a too predictable way. For example, if a person with a certain ability responds to all easy questions correctly and responds to all difficult questions incorrectly, the MNSQ value

may be lower than +0.5. "*The MNSQ value of lower than +0.5 is considered to be 'less productive' for measurement*" (Green, 2013, p. 169). On the other hand, if MNSQ value is higher than +1.5, it means that persons or items are performing in an unpredictable way. For example, if an able person responds to an easy item incorrectly, MNSQ value can be higher. Because of the unpredictability, "*the MNSQ value of higher than +1.5 is considered 'unproductive' for measurement*" (Green, 2013, p. 169). In the Rasch model, the 'unproductive' data (MNSQ > +1.5) are usually focused on and investigated rather than 'less productive' data (MNSQ < +0.5).

"*The infit and outfit statistics adopt slightly different techniques for assessing an item's fit in the Rasch model*" (Penfield, 2014). The infit MNSQ assigns relatively more weight to the performances of persons who are closer to the item difficulty value (Ibid.). Thus, if a person incorrectly answers the items particularly close to their ability level, the infit MNSQ value can be affected. On the other hand, outfit MNSQ is more sensitive to the influence of outlying scores (lucky guesses of low performers and careless mistakes of high performers). That is, outfit MNSQ is related to how a person responds to the items that are very easy (item difficulty logit < -2.0) or very hard (item difficulty logit > +2.0). Thus, if a very able person does not respond to a very easy item correctly, the outfit MNSQ value can be affected.

In a general theory of latent traits, it is assumed that a set of k latent traits or abilities underlie the examinee's performance on a set of test items. The k latent traits define a k dimensional latent space, with each examinee's location in the latent space being determined by the examinee's position on each latent trait. The latent space is referred to as complete if all latent traits influencing the test scores of a population of examinees have been specified.

It is commonly assumed that only one ability or trait is necessary to "explain," or "account" for examinee test performance. Item response models that assume a single latent ability are referred to as unidimensional. This assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors that impact test performance, at least to some extent. These factors might include the level of motivation, test anxiety, ability to work quickly, knowledge of the correct use of answer sheets, other cognitive skills in addition to the dominant one measured by the set of test items, etc. What is required for this assumption to be met adequately by a set of test data is a "dominant" component or factor that influences test performance. This dominant component or factor is referred to as the ability measured by the test.

Often researchers are interested in monitoring the performance of individuals or groups on a trait over some time. For example, at the individual (or group) level, interest may be centred on the amount of individual (group) change in reading comprehension over a school year.

Traub (Zheng, 2016) described how the nature of training and education can influence the dimensionality of a set of test items. For example, concerning education (Zheng, 2016) noted:

> *"The curriculum and the method by which it is taught vary from student to student, even within the same class. Out-of-school learning experiences that are relevant to in-school learning vary widely among students. Individual differences in previous learning, quality of sensory organs, and presumably also the quality of neural systems contribute to if they do not define, individual differences in aptitude and intelligence. It seems reasonable then to expect differences of many kinds, some obvious, some subtle, in what it is different students learn, both in school and outside. How these differences are translated into variation in the performance of test items that themselves relate imperfectly to what has been taught and learned, and thus into the dimensionality of inferred latent space, is not well understood."* (p. 17).

The assumption of a unidimensional latent space is a common one for test constructors since they usually desire to construct unidimensional tests to enhance the interpretability of a set of test scores. What does it mean to say that a test is unidimensional in a population of examinees? Suppose a test consisting of n items is intended for use in r subpopulations of examinees (such as several ethnic groups). Consider the conditional distributions of test scores at a particular ability level for several subpopulations.

Item local independence is one of the concepts of Item Response Theory (IRT), which has received great attention from researchers and authorities in the area of psychometrics because of its importance in Probability Theory. Ubi, et al. (2011) linked the local independence of items in a test to Probability Theory. According to him, local independence of items conceptualizes that, the probability of an examinee getting examination items correct must not depend on the answers given to other items in the examination. This means that, in a set of mathematics test items, for example, the answer an examinee gives to item number one should not be affected by the answer given to items number two. This is because the ability, that influences responses to any two items in a test, is constant; thus, the relationship between the two items should not differ from zero. If it does, then responses to the item are influenced by factor(s) other than what the test instrument was designed to measure. On such other factors, the examinees who have the same ability level but different response pattern reveal absence of local independence amongst the items constituting the test instrument.

The violation of the LID assumption can have substantial consequences on test parameter estimates and on proficiency estimates. Research studies show that statistical analysis of data with LID is misleading (Chen & Thissen, as cited in Yambi, 2018). Tuerlinckx and De Boeck (as cited in Maydeu-Olivares, et al., 2011) mathematically and empirically demonstrated the impact of LID on difficulty and discrimination item parameters. They showed that if negative LID is not modelled, the discrimination parameters of the interdependent items are underestimated. They also showed that the discrimination parameter (aj) depends on the difficulty of the item it interacts with, but not on the difficulty of the item itself. Due to its effect on the discrimination parameter, the negative LID deflates the item information (as a function of the square of aj), and the standard error of measurement is underestimated. It is therefore essential to ensure the

accuracy of the discrimination parameters, given that they index the item quality and therefore the test quality (Chen & Wang, as cited in Yambi, 2018). LID can also strongly bias the variance estimate of student ability and produce biased proficiency estimates.

Penfield (2014) identified several potential causes of LID. Some of them are independent of the item's content: external assistance (e.g., assistance from a teacher), fatigue (stimuli tend to be more difficult when they appear at the end of a test), practice, item or response format, speediness (if test-takers do not reach item j, they will surely not reach item j+1), and so on. Yambi (2018) calls this last type of local dependency "*surface local dependence*."

Other causes of LID Penfield (2014) relate to the content of items, namely, item chaining (items organized in steps) and explanation arising out of previous answers and stimulus dependence. This stimulus-LID can be produced by an examinee's unusual level of interest in or background knowledge about the common stimuli or by the fact that information used to answer different items is interrelated in the stimulus. Chen and Thissen (as cited in Yambi, 2018) define this category of dependence as "*underlying local dependence*" because it assumes a separate trait common to each set of locally dependent items. These separate traits can therefore be regarded as minor dimensions existing beside the unique essential latent dimension q.

Penfield (2014) identified two forms of independence, namely: statistical independence and stochastic independence. Statistical independence which is important for this present study refers to a situation in which two quantum systems acting randomly are said to be prepared differently. That is, when items are statistically independent, each exhibits its quality and it takes examinees' good display of ability to unfold the characteristic function about them. Based on this assertion, Fishbein and Ajzen (2010) posited that, in order to set mathematic test items that would not violate local independence, the interaction between each test constructs (items) must not be high, but as low as possible with respect to logic operations and manipulations. Items that tend to have the same pattern but with different ability levels should be distributed randomly to span the entire test's length. He concluded by recommending the avoidance of chain items since they can give clues to one another.

2. Research Questions

The study was guided by the following research questions:

1) What are the item and person separation indices of the Attitude Towards Assessment Test as evidence of reliability?
2) What is the item difficulty infit and outfit MNSQ statistics for the Attitude Towards Assessment Test?
3) What is the evidence of unidimensionality of the different components of the Attitude Towards Assessment Test?
4) What is the evidence of local independence of the Attitude Towards Assessment Test?

3. Methods

This study adopted an instrumentation research design. The population of the study consisted of secondary school students in Delta State, Nigeria. The sample size comprised 1,000 students based on the population of the study. A total of 40 students in each local government area of the state were selected to make a total of 1,000 students. This was done through simple random and non-probability cluster sampling techniques. In this case, the schools in each Local Government Area of the state were treated as clusters, such that the researcher randomly selected one school in each Local Government Area to make a total of 25 schools. This was done through a simple random sampling technique of the balloting method. Using this procedure, the researcher wrote the names of all the schools in each local government area on pieces of paper, fold and poured them into a container. He then shuffled them and picked one piece of paper from the container. Schools picked from this process were the selected schools in that Local Government Area. This was done for all Local Government Areas until all the 25 schools (one for each Local Government Area) were selected. This procedure produced 25 clusters, one for each Local Government Area. For each cluster, the researcher randomly selected one classroom out of the various classrooms in the secondary schools. All the students in the selected classroom were involved in the study.

The instrument used in the study is an Attitude Towards Assessment Test (ATAT), developed by Megbele, et al. (2023). The test comprises a total of 60 items which are distributed according to the components of attitude, such as Cognitive component (20 items), Affective component (20 items) and Behavioural components (20 items). Some of the items were phrased in negative forms and were represented with letter (N) at the end of the statement. The instrument was structured on a 4-point Likert-type scale of strongly agree, agree, disagree, and strongly disagree, representing 4, 3, 2, and 1 respectively. The expected score for the instrument is between a minimum of 60 and a maximum of 240. The negative items were reverse-coded before the item analysis.

The Attitude Towards Assessment Test (ATAT) was administered to the students directly by the researchers, with the help of 5 research assistants, who were trained on the purpose of the study. The research team visited the schools personally before the testing date to make their intention known to the principal or head of the school and to obtain permission. The research assistants were briefed on the purpose of the study and how to approach testees. The data were collected on the spot from the respondents. The Rasch Rating Scale Model was used to answer research 1, which assessed person and item reliability, item statistics, and ordering of response categories. Two fit indices including the infit and outfit mean square (MNSQ) statistics were used to answer research question 2. The data that were used to answer research questions 1 and 2 were analysed with the aid of the Jmetrik IRT software. The Categorical Confirmatory Factor Analysis (CCFA) was used to answer research question 3 while Chi-Square Goodness of Fit Statistics was used to answer research question 4 on construct validity, evidence of unidimensionality, and local independence respectively.

4. Results

Research Question 1: What are the item and person separation indices of the Attitude Towards Assessment Test as evidence of reliability?

Table 1: Reliability and separation indices in the Attitude Towards Assessment Test

Scale	Person	Item
Cognitive		
Reliability	0.834	0.974
Separation	2.245	6.172
Affective		
Reliability	0.847	0.947
Separation	2.356	4.208
Behavioural		
Reliability	0.859	0.973
Separation	2.463	6.034

Table 1 shows the Rasch-derived item and person separation indices and reliability for each component of the Attitude Towards Assessment Test. The three components of the scale had high values of item separation index and reliability as well as an acceptable range of Person separation index and reliability ($r \geq 0.070$). The cognitive component of the scale had a reliability index of 0.834 and a separation index of 2.245; the affective component had a reliability index of 0.847 and a separation index of 2.356; while the behavioural component had a reliability index of 0.859 and a separation index of 2.463.

Research Question 2: What is the item difficulty infit and outfit MNSQ statistics for the Attitude Towards Assessment Test?

Table 2: Item difficulty, infit, and outfit MNSQ statistics for each item of the Attitude Towards Assessment Test

S/N	Scale	Difficulty	Infit	Outfit
Cognitive Component				
1.	Assessment is used to determine the strength of students.	0.29	0.93	0.82
2.	Assessment is used to determine the weaknesses of students.	0.06	1.06	1.04
3.	Assessment is used by teachers to improve students' learning.	0.25	1.02	0.92
4.	Parents use the outcome of assessment to know when their children are doing well in their studies.	0.16	0.96	0.86
5.	When assessment is carried out on students, they can know when they are doing well.	0.15	1.00	0.94
6.	Assessment is only used to promote students from one class to another.	-0.32	1.16	1.14
7.	Through assessment, teachers' instructional activities can be properly guided.	-0.19	1.04	0.99
8.	Assessment is often carried out at the end of the school term.	-0.03	1.00	0.98
9.	Continuous assessment is a kind of assessment.	-0.16	0.96	0.95
10.	Assignment is a kind of assessment.	-0.07	0.90	0.80

11.	Classwork is a kind of assessment.	-0.03	0.90	0.78
12.	Assessment is also known as an examination.	0.04	0.90	0.83
13.	Assessment carried out by teachers are known as internal assessment or examination.	-0.01	1.01	1.00
14.	Assessment carried out by WAEC, NECO, or JAMB are known as external assessment or examination.	0.15	0.94	0.85
15.	Assessment is meant for only intelligent students.	-0.06	1.10	1.06
16.	Assessment is a difficult exercise.	-0.37	1.10	1.11
17.	Students with short-term memory should not attempt any assessment.	-0.18	1.08	1.07
18.	Examination malpractice is a threat to assessment outcomes.	0.11	1.02	0.98
19.	Teachers should set only questions they teach in an assessment exercise.	0.02	0.98	0.99
20.	I understand that all learning processes require assessments to determine the outcome of learning.	0.19	0.95	0.92
Affective Component				
21.	I am often tense when I think of assessment.	-0.28	1.01	1.01
22.	Assessment scares me.	-0.10	1.02	1.01
23.	I like assessment.	0.19	1.05	1.04
24.	I would not mind writing an examination every day.	0.07	1.06	1.05
25.	Assessment interests me.	0.11	1.13	1.19
26.	I am always excited when I think of an examination.	-0.01	1.08	1.02
27.	I look forward to the next examination.	0.23	1.00	0.95
28.	Assessment is one of my favourite activities in school.	0.10	0.97	0.96
29.	I feel assessment should be restricted to only intelligent students.	0.08	1.07	1.25
30.	Assessment is always fun for me.	-0.13	1.02	1.02
31.	Assessment makes my heart beat faster.	-0.19	0.98	0.94
32.	I often panic when I have to take a surprise tests.	-0.09	1.00	0.98
33.	I sometimes feel my heart beating very fast during exams.	-0.07	0.92	0.89
34.	If exams could be removed, I think I could learn more.	0.20	1.00	0.96
35.	I dislike people who feel so unhappy with exams.	0.05	0.96	0.99
36.	I do not like teachers who give surprise test.	-0.10	0.93	0.90
37.	It is stressful to write exams.	-0.06	0.87	0.81
38.	It is stressful to prepare for exams.	-0.02	0.96	0.94
39.	The thoughts of writing exams often give me goose bumps.	-0.12	0.99	0.95
40.	Assessment is the only thing I hate about school.	0.13	0.99	0.93
Behavioural Component				
41.	I would not mind taking tests everyday.	-0.12	1.05	1.02
42.	I do my assignments immediately I get home so I will not forget.	-0.02	0.97	0.88
43.	I am one of those who always answers questions in class.	-0.08	0.95	0.96
44.	I read my books every day, so I will not get so tense when an exam is approaching.	-0.17	1.07	0.97
45.	I will love to do classwork immediately after every lesson.	-0.09	0.93	0.81
46.	I would rather do assignments than take an exam.	-0.38	1.26	1.27
47.	I often work very hard to achieve success during assessment.	0.08	0.96	0.93
48.	I don't mind taking a surprise test.	-0.10	1.05	0.98
49.	I always complete my assignments on time.	-0.05	0.88	0.76
50.	I prepare very well for every examination.	0.09	0.83	0.73
51.	I will attend an exam preparatory class if it is available.	0.00	1.02	0.96
52.	I will not mind joining a group discussion class.	0.02	1.01	0.96

53.	I always complete my assignments on time.	-0.07	0.96	0.85
54.	I always participate in group discussion classes in preparation for an important examination.	0.16	0.92	0.92
55.	I will encourage students to prepare very well for their examination.	0.22	0.98	0.87
56.	I always take my school examination seriously.	0.18	1.04	1.00
57.	I only read my books when examination is approaching.	-0.28	1.21	1.21
58.	When taking tests, I take every question seriously.	0.09	0.91	0.78
59.	When taking tests, I often pay attention to the details.	0.06	0.99	0.95
60.	I read and try to understand the instructions before starting to answer the questions.	0.48	0.95	0.80

Table 2 represents item difficulty, infit, and outfit MNSQ statistics for each item of the Attitude Towards Assessment Test. The finding shows that the difficulty index of items in the test was within an acceptable range. The range of difficulty in the cognitive domain was from -0.37 to 0.29. Item difficulty ranged from -0.28 to 0.23 in the affective domain, while item difficulty for behavioural domain ranged from -0.38 to 0.48. In addition, infit and outfit MNSQ for all the items were within the accepted range.

Research Question 3: What is the evidence of unidimensionality of the different components of the Attitude Towards Assessment Test?

Figure 1: Scree Plot for the Attitude Towards Assessment Test

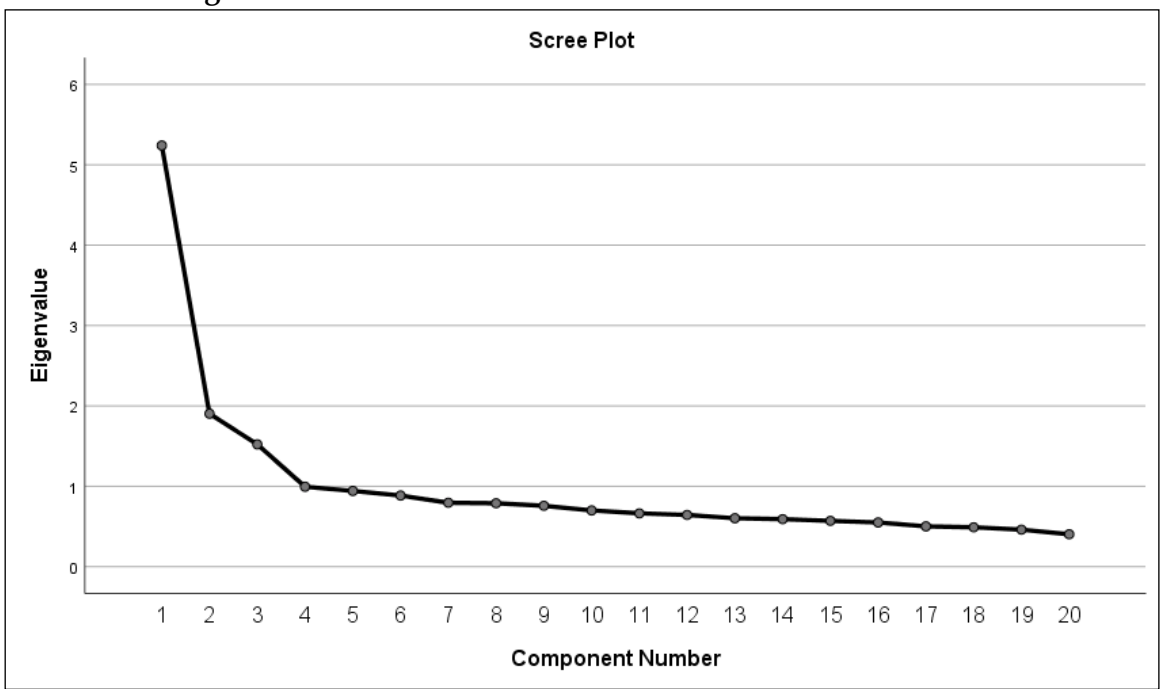

Figure 1 shows the scree plot for the Attitude Towards Assessment Test. From the figure, a careful examination of the scree plot shows that there are three components before the breaking point or elbow joint. This therefore succinctly shows that each of the components of the ATAT (cognitive, affective and behavioral) had one construct each,

which is evidence of unidimensionality for the different components of the Attitude Towards Assessment Test.

Research Question 4: What is the evidence of Local Independence of the Attitude Towards Assessment Test?

Table 3: Chi-square Goodness of fit indices for the Attitude Towards Assessment Test

S/N	Scale	$S\text{-}X^2$	Sig.	Remark
Cognitive Component				
1.	Assessment is used to determine the strength of students.	46.663	.000	√
2.	Assessment is used to determine the weaknesses of students.	70.106	.000	√
3.	Assessment is used by teachers to improve students' learning.	15.226	.085	X
4.	Parents use the outcome of assessment to know when their children are doing well in their studies.	27.677	.001	√
5.	When assessment is carried out on students, they can know when they are doing well.	7.599	.575	X
6.	Assessment is only used to promote students from one class to another.	31.107	.000	√
7.	Through assessment, teachers' instructional activities can be properly guided.	8.948	.442	X
8.	Assessment is often carried out at the end of the school term.	100.919	.000	√
9.	Continuous assessment is a kind of assessment.	13.655	.135	X
10.	Assignment is a kind of assessment.	17.434	.042	√
11.	Classwork is a kind of assessment.	11.764	.227	X
12.	Assessment is also known as an examination	19.649	.020	√
13.	Assessment carried out by teachers are known as internal assessments or examination.	11.750	.228	X
14.	Assessment carried out by WAEC, NECO or JAMB are known as external assessment or examination.	30.326	.000	√
15.	Assessment is meant for only intelligent students.	45.982	.000	√
16.	Assessment is a difficult exercise.	25.469	.002	√
17.	Students with short-term memory should not attempt any assessment.	71.814	.000	√
18.	Examination malpractice is a threat to assessment outcomes.	29.041	.001	√
19.	Teachers should set only questions they teach in an assessment exercise.	31.600	.000	√
20.	I understand that all learning processes require assessments to determine the outcome of learning.	24.803	.003	√
Affective Component				
21.	I am often tense when I think of assessment.	37.264	.000	√
22.	Assessment scares me.	59.835	.000	√
23.	I like assessment.	36.609	.000	√
24.	I would not mind writing an examination every day.	28.821	.001	√
25.	Assessment interests me.	58.556	.000	√
26.	I am always excited when I think of an examination.	74.409	.000	√
27.	I look forward to the next examination.	65.810	.000	√
28.	Assessment is one of my favourite activities in school.	15.364	.081	X
29.	I feel assessment should be restricted to only intelligent students.	72.492	.000	√

30.	Assessment is always fun for me.	23.510	.005	√
31.	Assessment makes my heart beat faster.	38.582	.000	√
32.	I often panic when I have to take a surprise test.	37.025	.000	√
33.	I sometimes feel my heart beating very fast during exams.	34.985	.000	√
34.	If exams could be removed, I think I could learn more.	50.934	.000	√
35.	I dislike people who feel so unhappy with exams.	26.325	.002	√
36.	I do not like teachers who give surprise test.	30.403	.000	√
37.	It is stressful to write exams.	59.731	.000	√
38.	It is stressful to prepare for exams.	38.639	.000	√
39.	The thoughts of writing exams often gives me goose bumps.	36.404	.000	√
40.	Assessment is the only thing I hate about school.	59.429	.000	√
Behavioural Component				
41.	I would not mind taking tests everyday.	29.108	.001	√
42.	I do my assignments immediately I get home so I will not forget.	35.005	.000	√
43.	I am one of those who always answers questions in class	11.906	.219	X
44.	I read my books every day, so I will not get so tense when an exam is approaching.	15.159	.087	X
45.	I will love to do classwork immediately after every lesson.	35.187	.000	√
46.	I would rather do assignments than take an exam.	4.575	.870	X
47.	I often work very hard to achieve success during assessments.	26.840	.001	√
48.	I don't mind taking a surprise test.	32.307	.000	√
49.	I always complete my assignments on time.	49.615	.000	√
50.	I prepare very well for every examination.	68.899	.000	√
51.	I will attend an exam preparatory class if it is available.	45.297	.000	√
52.	I will not mind joining a group discussion class.	24.694	.003	√
53.	I always complete my assignments on time.	44.471	.000	√
54.	I always participate in group discussion classes in preparation for an important examination.	19.053	.025	√
55.	I will encourage students to prepare very well for their examinations.	77.305	.000	√
56.	I always take my school examinations seriously.	32.957	.000	√
57.	I only read my books when an examination is approaching.	36.807	.000	√
58.	When taking tests, I take every question seriously.	34.944	.000	√
59.	When taking tests, I often pay attention to the details.	26.349	.002	√
60.	I read and try to understand the instructions before starting to answer the questions.	55.089	.000	√

Table 3 shows the Chi-square Goodness of fit indices for the Attitude Towards Assessment Test. The result showed evidence of local independence. The result shows that the chi-square score for the cognitive component ranged from 7.599 to 100.920; the affective component ranged from 15.364-74.409; while behavioural component ranged from 4.575-77.305. These values of the chi-square statistics suggest an evidence of local independence assumption.

5. Discussion

The first finding revealed that the three components of the scale had high values of item separation index and reliability as well as an acceptable range of Person separation index and reliability ($r > 0.70$). This finding suggests that the instrument is reliable and can be

used to assess students' attitude towards assessment. The cognitive component of the scale had a reliability index of 0.834 and a separation index of 2.245; the affective component had a reliability index of 0.847 and a separation index of 2.356; while the behavioural component had a reliability index of 0.859 and a separation index of 2.463. Undoubtedly, the above values are high values and indicate the adequacy of the scale items in separate individuals. The distinction between the different levels of ability of individuals on one hand, and the adequacy of the individual's sample in separating between the scale items and the definition of the characteristic continuum which items measure on the other hand. In a more precise sense, individuals are distributed appropriately on their attitudes towards assessment tests.

In any case, the reliability coefficient of the attitudes scale based on the current study according to the IRT was in line with the coefficients of reliability computed for attitudes scales in some previous studies, such as the (Polichnowski, 2008) and (Dimond et al., 2011), (Abdullah and Abu Fakhida, 2009), and (Al-Ghamdi, 2009). The above finding agrees with Demirtaşli, et al. (2016), who validated the Scale of Attitude towards Educational Measurement and Evaluation using polytomous Item Response Theory (IRT) models and identified its psychometric features. The authors found that the validity and reliability features of the scale are fairly good. The finding also agrees with Al-Dlalah, et al. (2021), who developed an attitude scale about e-learning among Isra University students according to the item response theory (IRT) in measurement, and found that the reliability of the scale was 0.94%, while the scale had multiple indications of validation.

The second finding showed that the difficulty index of items in the test was within an acceptable range. The range of difficulty in the cognitive component was from -0.37 to 0.29. Item difficulty ranged from -0.28 to 0.23 in the affective component, while item difficulty for the behavioural component ranged from -0.38 to 0.48. In addition, infit and outfit MNSQ for all the items were within the accepted range. This finding is in line with Al-Dlalah, et al. (2021), who developed an attitude scale about e-learning among Isra University students according to the item response theory (IRT) in measurement, and found that the test has adequate item difficulty infit and outfit MNSQ estimates. The finding also agrees with Cordier, et al. (2018), who used Rasch analysis method of item response theory to evaluate the reliability and validity of the Swallowing Quality of Life questionnaire, and found that the infit values of items in the questionnaire are within acceptable range.

The third finding revealed that each of the components of the ATAT (cognitive, affective and behavioural) had one construct each, which is evidence of unidimensionality for the different components of the Attitude Towards Assessment Test. The assumption of a unidimensional latent space is a common one for test constructors since they usually desire to construct unidimensional tests to enhance the interpretability of a set of test scores. What does it mean to say that a test is unidimensional in a population of examinees? Suppose a test consisting of n items is intended for use in r subpopulations of examinees (e.g., several ethnic groups). A test can be unidimensional within one population of examinees and not unidimensional in another. Consider a test with heavy cultural loading. This test could appear to be

unidimensional for all populations with the same cultural background, but, when administered to populations with varied cultural backgrounds, the test may have more than a single dimension underlying test performance. Examples of this situation are seen when the factor structure of a particular set of test items varies from one cultural group to another.

One of the ways of estimating unidimensionality in tests is factor analysis. According to Field (2013), factors or traits or underlining constructs can be extrapolated or established through the use of eigenvalues and variance, scree plot, and communalities. Several researchers have used factor analysis to determine the unidimensionality of a test and were successful. For instance, Kpolovie and Emekene (2016) validated the advanced progressive matrices for Nigerian sample using Item Response Theory. They used factor analysis to determine the unidimensionality of the scale and found that the unidimensionality of the underlining construct of the APM scale, namely intelligence or fluid ability, and that all 36 items of the scale measure one construct, the fluid ability of the test taker as confirmed by the scree plot. They concluded that all the items APM unquestionably measure just one general intelligence factor in Nigeria just as it does in all other countries where the test is actively in use.

The fourth finding showed that there is evidence of local independence. The result shows that the chi-square score for the cognitive component ranged from 7.599 to 100.920; the affective component ranged from 15.364-74.409; while behavioural component ranged from 4.575-77.305. These values of the chi-square statistics suggest evidence of local independence assumption. This finding agrees with Ubi, et al. (2011), who assessed the item local independence in University Matriculation examination in Nigeria for the years 2000 to 2003. Data analysis using Tertrachoric Correlation analysis revealed that JAMB-UME mathematics test items were significantly locally independent. It was thus concluded that JAMB-UME mathematics 2000 to 2003 met the assumptions of Item Response Theory (IRT) on local independence and thus is recommended, among others, that test practitioners and examination bodies in Africa should endeavour to adhere to the requirements of item local independence, while preparing test items using the proper procedures for item construction while preparing their questions.

6. Conclusion and Recommendations

Based on the findings of the study, it was concluded that all the items in the different components that made up the test are reliable, have adequate item difficulty infit, and outfit MNSQ estimates, with evidence of unidimensionality and local independence assumptions. The researchers recommended the following based on the findings of the study:

1) The Attitude Towards Assessment Test should be used to assess secondary school students' attitude towards assessment
2) The test can be used by teachers prior to any examination
3) The test can be used by examination bodies for the assessment of students in the affective domain

4) Other researchers who wish to carry out a study on students' attitudes towards examination can use the test
5) Some of the items that are weak should either be modified or discarded before the use of the test.

Conflict of Interest Statement
The authors declare no conflicts of interest.

About the Author(s)
Megbele, A. M. is a doctoral student of Measurement and Evaluation in the Department of Guidance and Counselling, Delta State University, Abraka, Nigeria. He earned his Master's degree in Measurement and Evaluation in 2021 and his Bachelor's degree in Biology Education from Delta State University, Abraka, Nigeria in 1996. The author is interested in item response theory, test development and psychometrics.
Odili J. N. is a Professor of Measurement and Evaluation in the department of Guidance and Counselling, Delta State University, Abraka, Nigeria. He obtained his B.Sc. (Ed) Biology degree from the University of Ilorin in 1986, M.Ed. and PhD in Measurement and Evaluation from the University of Nigeria Nsukka in 1992 and 2004 respectively. He established the effect of language manipulation on differential item functioning of Biology Multiple Choice Test, using Item Response Theory of Measurement. He has provided information that guides item writers to bridge the gender gap in performance in Biology. He has supervised several PhD and M.Ed. students in the field of Measurement and Evaluation.
Osadebe P. U. is a Professor of Measurement and Evaluation, in the department of Guidance and Counselling, Delta State University, Abraka, Nigeria. He specializes in test construction and standardization. He has supervised several PhD and M.Ed. students in the field of Measurement and Evaluation.

References

Ceniza, J. C., & Cereno, D. C. (2012). Development of mathematic diagnostic test for DORSHS. Available from http://www.doscst.edu.ph/index.ph[/academics/graduateschool/publication/category/5-volum-1-issue-1-2012?

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying Item Response Theory (IRT) Modeling to an Observational Measure of Childhood Pragmatics: The Pragmatics Observational Measure-2. *Front. Psychol.* 10, 408. doi:10.3389/fpsyg.2019.00408

Field, A. P. (2005). Discovering Statistics Using SPSS, Sage Publications Inc.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior.* New York: Taylor & Francis.

Green, R. (2013). Statistical analysis for language testers. New York: Palgrave Macmillan.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. Struct. Equation Model. *A Multidisciplinary J.* 18 (3), 333–356. doi:10.1080/10705511.2011.581993

Megbele, A. M., Odili, J. N., & Osadebe, P. U. (2023). Development of Attitude Towards Assessment Test for Secondary School Students in Delta State. *Canadian Journal of Educational and Social Studies,* 3(4), 120-132.

Muis, K. R., Winne, P. H., & Edwards, O. V. (2009). Modern Psychometrics for Assessing Achievement Goal Orientation: A Rasch Analysis. *Br. J. Educ. Psychol.* 79 (3), 547–576. doi:10.1348/000709908X383472

Penfield, R. D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. Educ. Meas. Issues Pract. 33 (1), 36–48. doi:10.1111/emip.12023

Sharkness, J., & DeAngelo, L. (2011). Measuring Students' Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys. *Res. High Educ.* 52 (5), 480–507. doi:10.1007/s11162-010-9202-3

Ubi, I. O., Umoinyang, I. E., & Joshua, M. T. (2011). Item local independence in selection examination in Nigeria. *Education for Today: Journal of Faculty of Education,* 7(1),175-188.

Yambi, T. A. C. (2018). Assessment and Evaluation in Education. Available from: https://www.researchgate.net/publication/342918149_ASSESSMENT_AND_EVALUATION_IN_EDUCATION

Zheng, Y. (2016). Online Calibration of Polytomous Items under the Generalized Partial Credit Model. *Appl. Psychol. Meas.,* 40 (6), 434–450. doi:10.1177/0146621616650406