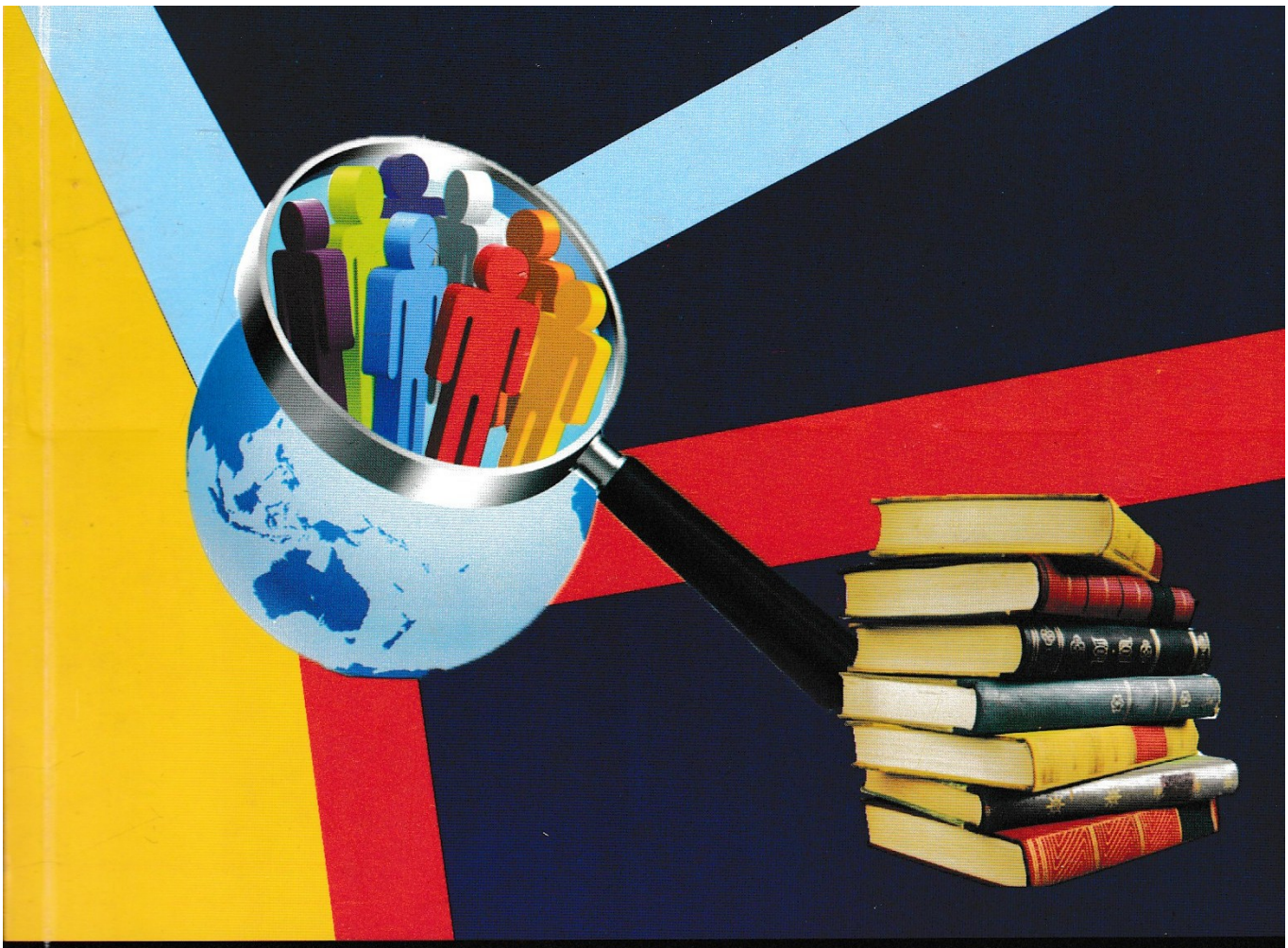

ASSEREN

JOURNAL OF EDUCATIONAL RESEARCH AND DEVELOPMENT (AJERD)

VOL. 4, JULY 2017 ISSN 2536-6899

ASSEREN
Knowledge Through Research

CONSTRUCTION AND VALIDATION OF A DIAGNOSTIC CHEMISTRY ACHIEVEMENT TEST (DCAT) USING ITEM RESPONSE THEORY

Ijeoma Joy Amajuoyi

&

U. Joseph Eme

Abstract

The development and validation of test instruments have been founded on Classical Test Theory (CTT). Unfortunately, the CTT has been unable to solve a number of testing problems. Consequently, a preferred alternative to CTT is Item Response Theory (IRT). The study therefore, is aimed at constructing and validating a Diagnostic Chemistry Achievement Test (DCAT) using 3-parameter IRT model. The study adopted the instrumentation research design. A sample of 1,150 senior secondary Chemistry students in Abia State, drawn using simple random and stratified random sampling techniques, participated in the study. The instrument for the study consisted of 60 multiple choice items of DCAT with 10 items in each subset developed by the researcher. Three research questions were posed and answered using Factor Analysis, BILOG MG V 3.0 techniques, Chi-square goodness of fit and Kuder-Richardson formula 20 (KR-20). The results showed that the KR-20 reliability coefficient of the entire DCAT was 0.891, DCAT measured one single trait for Chemistry ability and the fit to the model was good. The items parameters obtained were of low to moderate difficulty and discrimination levels. It was therefore, concluded that since the DCAT had satisfactory psychometric properties, the instrument being diagnostic in nature was dependable and valid. The results obtained have implication for Chemistry teachers, Guidance Counsellors and other test users. Based on the findings, it was recommended, among others, that classroom teachers should use the DCAT in identifying the areas of difficulty experienced by students and proffer remediation, while Guidance Counsellors should use the outcome of diagnosis for counselling purposes.

Introduction

The development and validation of classroom achievement tests have long been founded on classical test models. This is primarily due to its simplicity both conceptually and computationally. Classical Test Theory (CTT) introduces three concepts, test score (often called the observed score), true score and error score hence, a simple linear model linking the observable test score (X) to the sum of the two unobservable variables, true score (T) and error score (E) which form a relationship: $X = T \pm E$ (Brennan, 2010). The Classical Test Theory (CTT) assumes that the best set of achievement test items is defined as a set of homogenous items dominated by a single underlying dimension. CTT has limitations that make its use in developing and establishing the psychometric properties of classroom achievement test questionable. The first and foremost limitation is the variant nature of the two item statistics and person statistics. In other words, item difficulty and item discrimination indices that form the cornerstone of many CTT analyses are group dependent, while the person statistic is test specific (Hambleton & Jones in Amajuoyi, 2015). Test difficulty directly affects the resultant test scores. The true-score model upon which much of the CTT is based permits no consideration of examinee responses to any specific item. Consequently, no basis exists to predict how an examinee or a group of examinees, may perform on a particular test items. Conversely, Item Response Theory (IRT) allows the measurement experts greater flexibility of investigating how an examinee responds to an item and hence predicts the examinee's ability.

Item Response Theory (IRT) attempts to model the relationship between an unobservable variable referred to as the examinee's ability and the possibility of the examinee correctly responding to any particular test item (Lord cited in Umobong, 2004). Since the traits are not directly measurable and observable they are called latent traits. The latent trait, generically, is referred to as ability (Θ) which the test item is attempting to measure by estimating the level of difficulty of items a test taker can respond to correctly. Magno (2009) expressed the basic mathematical function of IRT as

$$P_i(\Theta) = b_i + a_i(\Theta);\ i = 1 \ldots n$$

P designates the probability that an examinee with a given latent trait, ability (Θ), will answer item *i* correctly. In the expression n is the number of items, a*i* parameter is the item discriminating index, b*i* is the index of item difficulty. The probability of each response is therefore derived as a function of the latent trait and the item parameters.The basic idea of IRT is that, if the relationship between Θ and P(Θ) is known for each item, the ability of each examinee, and the measurement error associated with each score can be derived mathematically. Put differently, IRT is concerned with how the probability of success on a test item varies as a function of ability (□) and other parameters represented by item difficulty, discrimination power and the degree of guessing (Baker, 2001). IRT models afford the estimates of the achievement test item parameters and person ability that are invariant that is, they neither depend on the subgroup the person belongs to nor on the selection of the specific set of items, provided the data fits the model (Nenty, 2004; Baker, 2001).This is because in IRT, the function of a test is concerned with estimation of an individual location on a dimension represented by a trait rather than being expressed in normative terms. The item parameters are invariant with respect to the ability of subjects in the validating sample (Nenty, 2004). An awareness of the limitations of CTT and the potential benefits offered by IRT has led some measurement practitioners to opt to work with an item response theory framework. The reason for this change of emphasis by the psychometric and measurement community from classical test theory models to item response theory models is a consequence of the benefits obtained through the application of item response models to measurement problems.

Basically, there are four assumptions of IRT namely unidimensionality local independence and Item Characteristic Curve (ICC) and monotonicity (Ojerinde, 2013). Unidimensionality which was investigated in this study refers to the existence of one underlying measurement construct (dimension) that accounts for variation in examinee responses (Magno, 2009). Unidimensionality is investigated using factor analysis and eigenvalues equal to or greater than one is considered significant (Lee, Cho, McGugin, Gulick & Gauthier, 2015).

DeVellis (2012) stated that IRT is really a family of models with three-parameter model being the most commonly used. It concentrates on the three aspects of item parameters namely, discrimination parameter (a), item difficulty parameter (b) and the pseudo-chance or guessing parameter (c) (Yu, 2013). The success of a particular IRT model can be assured only when the fit between the model and the test data set is satisfactory (Kenny, 2012).

Although the assumptions and mathematics of IRT are more complex, costly and time consuming, several authors such as Morales (2009); Kaplan and Saccuzzo (2005) have argued that their empirical benefits are sufficient to warrant their usage. Moreover, a closer examination of literature reveals that there is not much empirical study on the development of diagnostic achievement tests based on IRT models.

Diagnosis is central in the process of teaching and learning. As part of the instructional process, Ketterlin-Geller and Yovanoff (2009) defined diagnostic test as one in which

assessment results provide information about students' mastery of relevant prior knowledge and skill within the domain as well as preconceptions or misconceptions about the material . Teachers use this information to adjust instruction by identifying which area students have and have not mastered. This process if adequately followed, results in varied instruction plans that are responsive to students' need. According to Amajuoyi (2015), diagnosis plays a critical role as diagnostic assessment results are utilized for guiding instructional design and delivery decisions for students at-risk of failure. In addition, diagnostic test is considered the best approach for supporting students' achievement through the design of remedial interventions for struggling students. Diagnostic tests have the function of identifying specific difficulties in learning a subject (Aiken & Groth-Marnat 2006).It also plays the role of placement if it is done prior to instruction. It helps in focusing instruction by locating the proper starting point. The placement functions of diagnosis take several forms. It seeks to determine whether or not a student possesses certain entry behaviour skills required for attainment of the objectives of the lesson. It attempts to establish what the student has already mastered thereby allowing him to enroll in a more advance program. Above all it guides instructional decision making.

The psychometric properties of the diagnostic tests are typically confined to a limited area of instruction and hence the low to moderate level of difficulty (Amajuoyi, 2015). In addition, the number of items for measuring a particular sub-skill must be sufficient, at least ten. Buttressing the roles of diagnostic test, Popham (2009) asserted that "legitimate diagnostic tests supply the sort of evidence the teacher needs to make defensible instructional decisions" (p.90). Outcome of such test let teachers know what cognitive skills or bodies of knowledge students are having trouble with. For Popham (2009), a truly diagnostic test is designed;

1) to measure a modest number of significant, high-priority cognitive skills or bodies of knowledge;
2) to include enough items for each assessed attribute to give teachers a reasonable accurate fix on a test taker's mastery of that attribute;
3) to describe with clarity what the test is assessing and
4) not be too complicated or time-consuming.

A growing body of empirical studies advocates that CTT and IRT play a role in testing regardless the acclaimed superiority of IRT in solving measurement problems. Ifamuyiwa (2007) developed and validated a multiple choice diagnostic mathematics test for purposively selected 1200 primary six pupils in Lagos, Ogun and Oyo States of Nigeria. The outcome of the study produced 50-item diagnostic test from an initial 120- item pool following due process of test construction. The internal consistency of the 50-item diagnostic test was 0.78. The test is useful to the practicing primary school teachers in determining and identifying probable problem areas facing primary six pupils in the five content areas of the primary school mathematics curriculum and also serve as a guide for these teachers who might be interested in developing similar tests for pupils.

Based on IRT models, Nkpone (2001) reviewed in Amajuoyi (2015) carried out a research which applied latent trait models in the development and standardization of physics achievement test for 2215 Senior Secondary Schools students in Rivers State . The main aims of the study were to develop and establish the psychometric properties of a Physics Achievement Test (PAT) using the two-parameter latent trait models and the classical

test models. The reliability and validity of the instrument was estimated using 1 – and 2 – parameter latent trait models. The result indicated that K – R_{20} was used to estimate the overall reliability co-efficient and it yielded 0.89, the fit to the model was good and the PAT measured a single trait for physics ability. Furthermore, there was significant relationship among the items parameter obtained from 1 – parameter, the 2 –parameter of the latent trait models and the classical test model. Based on the findings, the researcher recommended that latent trait analysis should be followed up by some qualitative analysis.

In another development Stage (2003) reported in DeVellis (2012) compared CTT and IRT methods using data from the Swedish Standard Aptitude Tests. Based on a sample of 2,461 test takers randomly drawn from a pool of 82,506, the study concluded that while three-parameter IRT model fit the data poorly, a model based on CTT performed quite well. Recently, Magno (2009) had a similar study demonstrating the difference between CTT and IRT using actual test data for 219 junior high school students in Philippines. The comparison was across two samples and two test forms on item difficulty, internal consistencies and measurement errors, IRT approach used Rasch model. The result of the study revealed among others that IRT indices and internal consistencies were very stable across samples.

These studies reviewed applied CTT and IRT approaches in item analysis for achievement tests and diagnostic instruments in secondary subjects other than chemistry. Since there is no work in the recent time on diagnostic chemistry achievement test the researchers are aware of, the researchers therefore, wish to contribute to knowledge by applying IRT in constructing and validating a Diagnostic Chemistry Achievement Test (DCAT) with appropriate psychometric characteristics.

Chemistry as a core science subject is required for the study of science and science related courses in the university. It is presently plagued with problems of low enrolment and under achievement. The observed situation has been the concern of government, researchers, policy makers, educators, examination bodies and the society at large. Like other science subjects, chemistry is generally considered to be a difficult subject. It would be observed that efforts like the use of instructional materials, good assessment techniques, among others, employed to improve on performance have not yielded significant results. The need therefore arose for a diagnostic test that would enable the teacher to identify the areas in Chemistry that students are having learning difficulties and consequently recommend remediation interventions. It is expected that a good diagnostic assessment instrument should possess appropriate psychometric properties.

Development and validation of assessment instruments has been based on classical test theory which has weak assumptions. Conversely, IRT presents item statistics that are invariant of the group from which they were estimated, person statistics that are not test dependent, and test models that provide a basis for matching probability of success on test items to ability level. These advantages of IRT, among others, make it justifiable in considering its application in development of a Diagnostic Chemistry Achievement Test (DCAT). IRT is desirable to take into consideration some qualitative information regarding the psychometric characteristics of each item to be included in the final form of the test. Therefore, the problem of the study is that of developing a Diagnostic Chemistry Achievement Test (DCAT) with appropriate psychometric characteristics.

The purpose of the study was to construct and validate a diagnostic Chemistry multiple choice objective achievement test for senior secondary three students using IRT. Specifically, the study was designed to investigate the unidimensionality of DCAT, estimate item parameter using 3 – parameter IRT model and investigate the fit of DCAT items to the IRT models.

Research Questions

The study was guided by the following research questions

1. To what extent are DCAT items unidimensional?
2. What are the estimates of IRT item parameters using 3-parameter model?
3. To what extent do DCAT items fit the IRT models?

Methodology

This study adopted instrumentation research design aimed at developing a diagnostic Chemistry test for assessing students' deficiencies in senior secondary school Chemistry. Instrumentation design was deemed appropriate as it deals with the psychometric principles for test development and validation on the basis of certain test theories (Kpolovie, 2010).

The population consisted of 11,666 SS3 students in 216 public senior secondary schools in Abia State. Simple random and stratified random sampling techniques were used to select sample for the study. Two education zones, Umuahia and Aba, were randomly drawn from the three existing education zones of Abia State. These zones had four and nine Local Government Areas (LGAs) respectively; two and four LGAs from Umuahia and Aba education zones respectively were randomly selected for the study. Furthermore, 17 schools were randomly drawn from Umuahia zone while 33 were selected from Aba zone, giving a total of 50 schools. The intact classes of all students offering Chemistry in the 50 schools were used for the study. These made up a sample size of 1150 students.

To trial test the DCAT, 120 out of 528 Chemistry students in eight out of 13 senior secondary schools in Isialangwa North Local Government Area of Abia State were used.

The researchers used the questionnaire titled, "Difficult Topics in Senior Secondary Chemistry Curriculum Questionnaire" (DTSSCQ) to elicit responses from students on the topics they considered difficult in SS1 and SS2 Chemistry curriculum. Thereafter, the mean score of each topic was obtained and topics with mean score of 3.0 and above were considered difficult. DCAT was developed based on the topics identified using DTSSCQ.

DCAT is a six test battery, 4-option multiple choice objective test comprising A, B, C, D, E and F sub-sets. Each test battery consists of items on a topic the students perceived to be difficult. Test battery A was on Chemical Combination; B was on Acids, Bases and Salts; C was on Chemical Reactions; D was on Mass-Volume Relationship; E was on Oxidation and Reduction, and F was on Organic Chemistry. They are made up of 50, 50, 50, 50, 60 and 75 items respectively giving initial pool of 335 items. These test batteries were pre-tested and the results of the item analyses were used to generate two equivalent test forms A and B. Each subsections in both forms A and B DCAT had10 items in each.

In order to establish the face validity of the of the instrument, the DCAT, its marking guide and the Chemistry curriculum were given to three experienced Chemistry teachers who had

taught Chemistry for not less than ten years and three experts in Measurement and Evaluation. These subject specialists were involved in order to confirm the topics drawn from the curriculum and the correctness of the marking guide while the experts in Measurement and Evaluation were a to matcli the items with the topics from which they were developed, establish the adequacy of the items marking guide, appropriateness to the class level, clarity of words, and plausibility of the distracters. The corrections and suggestions of the Measurement and Evaluation experts and subject specialists were taken into consideration and integrated into the drafts for trial testing and field survey. After the final administration and calibration of the test forms, a single version of the test was obtained by assembling items with appropriate psychometric characteristics. The reliability coefficient for internal consistency was computed for the six DCAT subtest using Kuder-Richardson formula-$_{20}$ (K-R_{20}). The returned coefficients were .54, .63, .72, .73, .83, and .70 for subtests A, B, C, D, and F respectively; and the entire DCAT yielded 0.89. These indices were adjudged suitable to regard DCAT as a reliable instrument.

The research instrument, DCAT, was administered to the 1150 students using Chemistry teachers in the sampled schools as research assistance. The date for administering the two form of DCAT was announced seven days to the day of testing and the aim of the test was explained to the students as well as the description of the DCAT subsets. The students were given identification numbers which were also used to number the test booklets. This was necessary for easy matching of the test for each examinee after testing. The teachers at the schools assisted in administering the test at the time that was convenient to the school. The students were given enough time, about an hour thirty minutes to ensure that they attempted the items to the best of their ability since the test was for diagnostic purposes. Test Form B was administered two weeks after under similar test conditions.

The data collected was dichotomously scored. Items correctly responded to were scored 1 while 0 was given to wrong responses. The score per item per respondent was obtained and also the scores per respondent per subsection of the DCAT. The scores in each case were collated and analysed with Maximum Likelihood Estimation Technique of BILOG MG V 3.0 procedures for 3-parameter model, factor analysis for unidimensionality and Chi – square goodness of fit for model fit.

Results

Research Question 1: To what extent is DCAT unidimensional?

Table 1: ***Eigenvalues associated with the items***

Eigenvalues	% Variance	Cumulative	
DCAT A	4.09	6.82	6.82
	3.40	5.67	12.49
	2.70	4.49	16.98
	2.38	3.97	20.9
	2.25	3.74	24.69
	2.04	3.41	28.10
DCAT B			
	4.12	6.862	6.86
	4.05	6.747	13.61
	3.53	5.889	19.50
	3.07	5.119	24.62
	2.89	4.821	29.44
	2.73	4.550	33.99

From Table 1 the factor analysis procedure applied to DCAT Forms A and B yielded a six-dimensional solution for each test form. Generally, the communalities were moderately high. For DCAT A and B six factors each were extracted accounting for 28.10% and 33.99 respectively of the total variance. The eigenvalue for the first six factors for DCAT A were 4.09, 3.40, 2.70, 2.38, 2.25, and 2.04; for DCAT B the factors were 4.12, 4.05, 3.53, 3.07, 2.89, and 2.73. The values are greater than one providing evidence for one dominant factor. These items loaded positively and significantly on this factor.

Research Question 2: What are the estimates of IRT item parameter using 3-parameter model?

Table 2: ***Item parameter estimates based on 3-parameter IRT model***

Item	a-parameter	b-parameter	c-parameter	Item	a-parameter	b-parameter	c-parameter
1	.650	.049	.000	31	.456	.312	.006
2	.368	.392	.001	32	.352	.244	.001
3	.334	.338	.001	33	.577	.397	.004
4	.375	-.052	.001	34	.456	-.079	.198
5	.575	.223	.001	35	.967	.714	.056
6	.556	-.351	.001	36	.406	.400	.003
7	.478	.402	.001	37	.396	486	.001
8	.544	.514	.001	38	.600	.536	.003
9	.375	.461	.003	39	1.345	.179	.032
10	.430	.106	.001	40	.380	.828	.002
11	.801	1.127	.001	41	.923	.615	.000
12	.563	.823	.001	42	1.335	.358	.006
13	.560	.823	.002	43	..843	-.212	.000
14	.478	.356	.002	44	.833	.519	.000
15	.310	1.152	.001	45	.380	.828	.002
16	.375	.376	.001	46	.730	.548	.001
17	.592	1.008	.001	47	1.625	.622	.001
18	.478	.238	.001	48	.624	-.016	.001
19	.403	.951	.001	49	.624	-.016	.000
20	.746	1.719	.001	50	.634	-.246	.000
21	.746	.091	.001	51	.367	-.148	.001
22	.933	.838	.001	52	.340	.426	.000
23	.545	.167	.002	53	.730	.543	.001
24	.478	.475	.001	54	.511	.335	.001
25	.563	.729	.002	55	.483	.116	.008
26	.370	.140	.004	56	.511	.335	.001
27	1.710	1.174	.004	57	.556	-.140	.052
28	.499	.007	.001	58	.624	1.320	.001
29	.370	.159	.001	59	1.076	-.343	.062
30	.349	.202	.001	60	.487	.029	.004

Table 2 showed that the item parameter values obtained were as follows: for a-parameter, .01 – .34, .35 – .64, 0.65 – 1.34, 1.35 – 1.69, and ≥1.70 were regarded as very low, low, moderate, high, and very high level discrimination respectively. For b-parameter, the values of ≥−2.5,0.0,≤+2.5 were interpreted as very easy, moderately difficult and very difficult respectively. The c-parameter values were interpreted as good proportion of ≤ 2.0, and good proportion of guessing with the value of ≥ 2.0. The item parameter estimates selected from DCAT Forms A and B for inclusion in the final DCAT were presented in Table 3.The result indicated that for item difficulty parameter (b), that 16.7% (10 items) were very easy, 73.3% (4 items) were of moderate difficulty and 10% (6 items) were very difficult. For discrimination parameter (a), 6.7% (4 items) have very low discrimination. 69% (39 items) very low discrimination, 25% (15) moderate discrimination whereas only 3.3% (2) discriminates well but over a small range of ability. It was also observed that at all ability levels, the probability of responding correctly to items, 90% (54) of the items had a probability .20 or below whereas for 10% (6 items) there is no chance of guessing them correctly.

Research Question 3: To what extent do the items fit the IRT model used?

Table 3: ***Analysis of Likelihood ratio Chi-square misfitting items for the two test forms***

DCAT A (N = 60)			DCAT B (N = 60)		
Items	Chi-square	df.	Items	Chi-square	df.
5	16.0	9.0	29	16.9	8.0
7	12.0	8.0	43	18.0	8.0
22	13.3	9.0	51	17.0	9.0

Alpha = .01; Chi-square critical = 20.1 and 21.7 for df. 8 and 9 respectively

Table 3 showed that five items of DCAT A and 3 items of DCAT B were identified as misfitting the 3-parameter IRT model. From the result of the analysis, the answer to research question 3 is that the data fit well since only 6 items out of the 120 items for DCAT A and DCAT B put together were identified as misfitting.

Discussion

The result in Table 1 revealed that that the first six eigenvalues of DCAT A and DCAT B respectively were extracted based on the criterion of those eigenvalues greater than unity. This implies that a dominant factor exists among all the items of the test, that is, the items are measuring the same latent trait. This finding is consistent with those of Lee et al (2015) who extracted 12 out of 48 items on Venderbilt Expertise Test for cars (VETcar) whose eigenvalues were each equal to or greater than1 and loaded positively and significantly on the factor. They concluded that the test items had one dominant dimension which sufficiently explained the item variance of the test. The finding implied that the items on the two test forms measure the same underlying dimension, that is, chemistry ability. Therefore, the researcher concluded that the unidimensionality assumption for 3-parameters IRT model held for the data used for the study the total variance of the two forms of DCAT are nearly the same indicating that the items in the two test forms measure common underlying trait, Chemistry ability.

The result in Table 2 answered the question on the item parameter estimate based on 3-parameter IRT model. The average of the item parameter was computed and their suitability determined based on the guidelines for interpreting item parameters by Baker (2001). According to the authors, b-parameter of ≥ -2.5 is very easy, 0.0 is moderately difficult and ≤ +2.5 is very difficult. For discrimination index, item with 0.1 - 3.4 values had very low discrimination, 3.5 – 0.64 is low discrimination, 0.65 – 1.34 is moderate discrimination and 1.34 – 1.69 have high level of discrimination. The c-parameters is adjudged suitable when it is less than or equal to 0.20. The psychometric characteristics of DCAT in Table 2 revealed that 73.3% represented by 44 of the item are of moderate difficulty, low to moderate discrimination and low proportion of guessing. These indices are satisfactory for DCAT being a diagnostic test that is meant to identify areas of difficulties in chemistry. This finding collaborates with the assertion of Baker (2001) that the psychometric properties of diagnostic tests are typically confined to a limited content area of instruction and hence the low level of difficulty. Baker added that the psychometric properties of a test

are determined by the purpose for which the test was designed to serve.

Table 3 revealed the extent to which the IRT model assumptions are valid for the data obtained in respect to DCAT and how well testing data fit the 3- parameter model used in this study. Details of the analysis with Chi square goodness of fit yielded three items for DCAT A and three for DCAT B which were not significant at 1% alpha level. In other words, these items are misfitting to three-parameter IRT model. Misfit provides invaluable diagnostic tool for test developers. This finding is coherent with Fan (1998) cited in DeVellis (2012) who found one or two items as misfitting with a sample of 6000 examined on 60 mathematics test items and 48 reading test items. He therefore, concluded that the result of the data fits the two and three parameter IRT models exceptionally well. The result of the present study is coherent with Fan's study, in which the ability estimate of the three-parameter model used to reproduce the data set hinged on the robustness of the data. Similar views are held by authors like Nenty (2004), and Kenny (2012). However, the finding disagrees with the findings of Stage (2003) reported in DeVellis (2012) as he compared CTT and IRT methods using data from the Swedish Standard Aptitude Tests that while three-parameter IRT model fit the data poorly, a model based on CTT performed quite well.

Conclusion

In view of the research questions answered, there is considerable evidence that DCAT is of high face validity and the reliability of the subtests and the entire test are substantial. Therefore, study concluded that DCAT measures a single underlying construct,which chemistry ability, only sixitems put together,misfitted the three-parameter model used. Providing evidence that the data fit the IRT model and most of the items of DCAT are satisfactory in terms of item difficulty parameter (b), item discrimination parameter and guessing parameter. The average b-estimate is 2.32, average a-estimate 3.67 and average c-estimate is .049. These indices indicate that the instrument, being diagnostic in nature, is dependable and valid and can be used for diagnosis in chemistry.

Recommendations

Based on the findings and implications of the study, the researcher recommends as follows:-

i. The DCAT should be used to identify the areas of strengths and weaknesses of the chemistry students.

ii. Data obtained from this survey will provide clues to the teacher which might suggest some adjustments in instructional techniques with succeeding group of students and remediation programme for present ones.

iii. The DCAT should be used either in part or whole in diagnosing students' difficulty in chemistry. This is very important as the information obtained would serve for guidance purposes as well as decision making regarding the students' education.

iv. Chemistry teachers should lay greater emphasis on the areas where the performance of the students was lowest.

v. Test developers should investigate the dimensionality of measuring instrument to ensure that the items are measuring one underlying trait in order to avoid bias in testing.

vi. Test developers should also assess model fit and employ IRT model to ensure objectivity in measurement.

References

Aiken, L. R. & Groth – Marnat G. (2006). *Psychological testing and assessment.* (12th Ed.). Needham Heights: A Pearson Education Company.

Amajuoyi, I. J. (2015). Construction and validation of a diagnostic Chemistry achievement test for senior secondary three

students using classical test and item response theories. An Unpublished Doctoral Thesis. University of Uyo, Akwa Ibom State.

Baker, F. B. (2001). *The basics of item response theory.* (2nd Ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1 – 21. Retrieved December, 2012 from http://dx.doi.org/10.1080/08957347.2011.532417.20.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd Ed.). California: SAGE Publications, Inc.

Eluwa, O. I, Akubuike, N. E. & Bekom, K. A. (2011). Evaluation of Mathematics achievement tests: A comparison between classical test theory and item response theory. *Journal of Educational and Social Research, 1*(4), 99 – 106.

Ifamuyiwa, A. S. (2007). Development and validation of a multiple choice diagnostic test for primary six pupils. *Journal of Science Teachers' Association of Nigeria, 42*(1&2), 33 – 41.

Kaplan, R. M. & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications and issues.* (6th Ed.). Belmont: Thomas Wadsworth.

Kenny, D. A. (2012). Measuring model fit. Retrieved 20 May, 2013 from http://www.davidakenny.net/cm/fit.html.

Ketterlin – Geller, L. R. & Yovanoff (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research and Education, 14*(16), 1–11. Retrieved on 20th December, 2012 from http://pareonline.net/getun.asp?v=14&n=16.

Lee, W. Y., Cho, S. J., McGugin, R. W., Gulick, A. B. V., & Gauthier, I. (2015). Differential item functioning analysis of the venderbilt expertise test for cars. *Journal of Vision, 15*(13), 13–23. doi: 10.1167/15.13.23.

Kpolovie, P. J. (2010). *Advanced research methods.* Owerri: Springfield Publishers Limited.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Association, 1*(1), 1–11.

Morales, M. A. (2009). Evaluation of Mathematics achievement test: A comparison between CTT and IRT. *The International Journal of Education and Psychological Assessment, 1*(1), 19 – 26.

Nenty, H. J. (2004). From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In: Afemikhe, Q. A. and Adewale, J. C. (Eds.). *Issues in educational measurement and evaluation in Nigeria (In Honor of Wole Falayejo)* pp. 37–384, Ibadan: Institute of Education, University of Ibadan, Nigeria.

Ojerinde, D. (2013). Classical test theory vs item response theory: An evaluation of the comparability of item analysis results. Lecture Presented at the Institute of Education, University of Ibadan, Nigeria.

Popham, W. J. (2009). Diagnosing the diagnostic test, *Educational Leadership, 66*(6),90 -91.

Umobong, M. E. (2004). Item response theory: introducing objectivity into educational measurements. In O. A. Afemikhe & J. G. Adewale (Eds.). *Issues in educational measurement and evaluation in Nigeria (In Honor of Wole Falayejo),* pp. 384 – 398. Ibadan: Institute of Education, University of Ibadan, Nigeria.

Yu, C. H. (2013). A simple guide to the item response theory and Rasch modeling. Retrieved 30th November, 2015 from http://www.creative_wisdom.com.