Comparative study of classical test theory and item response theory using item analysis results of quantitative chemistry achievement test

Dr Telimoye Leesi Mitee
Port Harcourt Polytechnic, Rumuola, Port Harcourt. Rivers State, Nigeria

ABSTRACT
Classical Test Theory (CTT) and Item Response Theory (IRT) are the two major test theories used to analyse test responses. This study compared the results from the CTT and IRT 3 parameter Logistic model analyses to ascertain which is better fit for the analysis of Quantitative Chemistry Achievement Test (QCAT) items. Fifty objective test items on Senior Secondary 2 quantitative chemistry (calculations in chemistry) were used and Survey research-type of non-experimental design was adopted. Randomly selected sample of 1105 students participated. Item analyses were done based on CTT and IRT using Bilog MG, Dimtest and R software. The QCAT items certified the IRT assumptions on dimensionality (unidimensional), local independence and model-data fit (3 parameter logistic model). Items from CTT analysis with $r_{pbs} \geq 0.20$ *and* $0.30 \leq p \leq 0.80$ *were considered good items and selected. Items from IRT analysis with* $-3 \leq b \leq +3$ *were considered good items and selected. The result revealed that only 11 items survived the CTT analysis while 20 items survived the IRT analysis. The reliability of the items from CTT (0.27) was lower than that of IRT (0.60). IRT is a more effective method of item analysis for objective test item responses and should be used by scale developers.*

Number of Words:198
Key words: Quantitative chemistry, IRT, CTT, Difficulty indices, Discrimination indices.

INTRODUCTION
Multiple-choice objective test is one of the major instruments used to take decisions on students' abilities and placements. It is therefore imperative that multiple-choice objective test on quantitative chemistry be properly developed with appropriate theories to avoid poor performance and wrong assessment of the true abilities of students. One of the major ways the development of a quality multiple-choice objective test can be achieved is through a proper item analysis. Item analysis is a process of examining the responses of examinees to test items to find out if each test item is of good quality. Good test items are retained while poor test items are modified or rejected and deleted from the pool of test items. This helps to ensure that the test is of a good quality.

Classical test theory (CTT) and item response theory (IRT) are the two approaches commonly used to determine the quality of test items. CTT, which is also called the 'true score model', is a test theory which postulates that the observed or obtained score of an examinee on a test is the sum of two unobserved scores (true score which is error free and an error score). This is mathematically represented as X = T + E where X is the observed or obtained score, 'T' is the true score without measurement error and 'E' is the error score. The true score is the expected observed score of an examinee which remains the same (stable) even when the person is given equivalent tests so many times while the error score is the deviation from the true score caused by extraneous influences at the time of measurement such as fatigue, anxiety, stress, etc. Since the true score (to be estimated from the individual's responses on a set of test items) and error score are unknown in the linear equation, it makes X (the observed score) not solvable. Classical test theory is therefore based on the following assumptions:

1. True score (T) and error score (E) from the same test are uncorrelated (zero correlation).
2. The average error score of all the examinees that took the test is zero This means the measurement errors after many repeated measurements cancel out to give an average of zero if the error score is zero.
3. The error scores on parallel tests are uncorrelated.

Item difficulty is the proportion of examinees that correctly answered the item. It is obtained by dividing the number of students who answered the item correctly by the total number of students who took the test. Crocker and Algina (1986) & Ebel (1965) as cited in Ayanwale (2018) suggested that under CTT, items with difficulty indices (p) of less than 0.2 and greater than 0.8 and discrimination indices (r_{pbs}) of less than 0.2 should be deleted. This is because items with difficulty indices of less than 0.2 are too difficult (20% answered the item correctly) and more than 0.8 are too easy (80% answered the item correctly). This implies that difficulty indices close to zero indicate difficult items and close to 1 indicate easy items.

Item discrimination is the extent to which an item differentiates between high and low performing students. For instance, if 10 students took a test and 9 low performing students answered an item correctly and only 1 high performing student answered the item correctly, it means that the item discriminated poorly between high and low performing students. This is probably an indication that the item is not measuring what it was meant to measure. An item that discriminates well should have correlation values close to 1 which implies that students who answered an item correctly should also perform well in the test.

The ability of CTT to detect poor items through simple calculations of item difficulty and discrimination indices makes it very important in test development. It is also very important because it can accommodate small sample sizes. However, CTT has some inadequacies and a major one is that it is group dependent, that is, item difficulty depends on the abilities of the examinees where the sample was drawn (Cappelleri, Lundy & Hays, 2014). This means that test items appear easy if the examinees are high performing students and appear difficult if the examinees are low performing students. The implication is that the obtained item difficulty and discrimination values can only be applicable to the development of tests for populations of examinees that are similar to the examinees that were used to generate the item parameters, making it very limited. Another problem with CTT as observed by Ayanwale (2018) is the fact that the sum of the scores of all the items for an examinee gives the total score for that examinee and this does not depend on the difficulty and discrimination levels of the items. These inadequacies have been taken care of by modern test theories such as Item response theory (IRT).

Item response theory is a test theory that is interested in the relationship between the ability of an examinee (latent or hidden variable) and the probability that the examinee will answer an item correctly. IRT links this latent trait (ability) in the examinees to some observable characteristics giving each examinee a numerical value or score on an ability scale.

There are four IRT models: one-parameter logistic model (1PLM), two-parameter logistic model (2PLM), three-parameter logistics model (3PLM) and four-parameter logistics model (4PLM). However, only the first three are commonly used. The 1PLM estimates only the difficulty indices of the items. It permits each item to have its own independent difficulty level but assumes equal discrimination indices for all the items. Two parameter logistic model estimates the difficulty and discrimination levels of test items by accepting the fact that each item has its own independent difficulty and discrimination levels but assumes that an examinee cannot correctly answer an item by guessing. Three parameter logistic model estimates item difficulty, discrimination and guessing indices. It assumes that an examinee can answer a test item correctly by guessing. The Four parameter logistic model is also called carelessness.

Item response theory has the following three assumptions which must be applied for a test data to be suitable for IRT model estimation. It is imperative for these assumptions to be satisfied for IRT estimations to be valid:

1. Dimensionality: *Before an* IRT analysis is carried out, it is assumed that the dimensionality test has been carried out to know if the test data is unidimensional or multidimensional. If it is unidimensional, it means that the test is measuring only one dominant latent trait. It assumes that an examinee's performance on a test item depends on only one factor. However, if the test is multidimensional, it is measuring more than one trait. It assumes that an examinee's performance on a test item depends on more than one factor. The determination of the dimensionality of a test is germane because it determines the appropriate software to be used for the IRT analysis.

2. Local Independence: This assumption states that there is no statistical relationship between examinees' responses to two different items in a test (Kyung, 2013 as cited in Eleje, Onah & Abanobi, 2018). This means that the probability of an examinee getting an item correctly is not affected by the answers the examinee has provided for other test items.

3. ***Model-Data fit:*** This analysis is used to check the IRT model (1PLM, 2PLM or 3PLM) that fits or is compatible to the data the most.

Zickar and Broadfoot (2010) observed that IRT is regarded as a superior approach to CTT and this is evident in some studies. Eleje, Onah & Abanobi (2018) found in their study that with respect to very difficult items and poor item discrimination, CTT was not comparable with the appropriate IRT 3PLM used and the reliability value for CTT was found to be lower than that of IRT 3PLM. Another study that had similar result was that of Ayanwale, Adeleke & Mamadelo (2018). They all found in their studies that CTT and IRT item statistics were not comparable; IRT was found to be better than CTT.

However, some other studies did not find disparities between the two approaches. The studies of Nabeel& Chin (2011) and Guler, Uyanik & Teker (2014) did not find much disparity between using IRT 1 or 2 PLM and CTT. Although they found a significant difference between IRT 3PLM & CTT, Meade & Mead (2010) did not also find much empirical evidence to support their hypothesis that IRT would be better than CTT in their study using unrepresentative samples. They rather found little evidence to support that CTT was better than IRT when a sample size is small. Awopeju & Afolabi (2016) compared only 1 and 2 IRT PLM results and found that CTT and IRT were comparable in estimating item characteristics of statistical and psychometric tests. The study of Ojerinde (2013) on IRT and CTT also revealed comparable and almost identical results in the two approaches and recommended that both approaches can be used for item analysis.

Some studies found CTT to be better than IRT, some found IRT to be better than CTT and yet some did not find any disparity between the two. It also appears that no study has compared classical test theory and item response theory using item analysis results of quantitative chemistry achievement test. *This study therefore compared the* item analysis *results from CTT and IRT 3 parameter Logistic model to ascertain which is best fit for the analysis of Quantitative Chemistry Achievement Test.*

Research Questions

1. Which IRT model (1PLM, 2PLM or 3PLM) best fit the QCAT items?

2. Does the QCAT data satisfy the IRT assumptions of dimensionality and local independence?

3. How comparable are the item parameters ***(Difficulty and Discrimination indices)*** of the QCAT under CTT and IRT 3PLM analysis?

METHODS

This study adopted a Survey research type of non-experimental design. Simple random sampling was used to select 20 schools in Obio-Akpor Local Government area of Rivers State. A SS 2 science class was also randomly selected

from each school and all the chemistry students in the classes participated. Fifty-item Quantitative Chemistry Achievement Test (QCAT) was used for data collection. The 50 items were reviewed by experts in chemistry, which gave face validity to the test, and were administered to 1105 senior secondary 2 chemistry students. The students were not timed but were directed to answer all the questions. Calculations in Stoichiometry, *Acid-Base reactions* and Mole Concept were the contents in quantitative chemistry covered in this study.

Data was analysed using Stout's test of essential unidimensionality implemented in Dimpack, Yen Q3 statistics implemented in R software and Phase 1 and 2 module of BILOG-MG. For CTT, items with difficulty indices of less than 0.30 were considered too difficult, more than 0.80 were considered too cheap and were not selected. Items with less than 0.2 discrimination indices were not also selected because they were considered not to discriminate sufficiently. For IRT, items with discrimination indices equal or more than -3 and equal or less than +3 were considered good items and were selected.

RESULTS

Research Question 1. Does the QCAT data satisfy the IRT assumptions of dimensionality and local independence?

(I) Stout's test of essential unidimensionality, implemented in DIMTEST version 1.0 software, was used to establish the assumption of dimensionality of the test data. It was hypothesised that there will be no significant difference between the Partitioning subtest (PT) and Assessment subtest (AT) of the examinees' responses. Once this is not rejected, assumption of unidimensionality is justified.

Table 1.1: Stout's test of essential unidimensionality of 50 QCAT items

TL	TG bar	T	P-value
5.4844	**3.7439**	**2.7319**	**0.0516**

Table 1.1 presents the result of Stout's test of essential unidimensionality of the 50-QCAT items. The result showed that the null hypothesis was not rejected (T=2.7319, p >0.05). It can therefore be concluded that the test is essentially unidimensional, which implies that only one dimension is accounted for the variation in the responses of the examinees to the QCAT items.

(ii) Yen Q_3 statistics implemented in R software was used to establish local independence of the test data. Yen Q_3 statistics is the correlation of residuals for a pair of items after the person's location estimates are controlled for. After obtaining the residuals, the linear correlation between the residuals from pair of items is then examined to find pairs of items with large residual correlations. Correlation coefficient larger than 0.2 screening criterion suggested by Yen (1993) cited in Ayanwale, Adeleke and Mamadelo (2018) indicates that the paired item violates local independence.

Table 1.2: Summary of Yen Q3 Statistics of correlation of item residual

	qchem1	qchem2	qchem3	qchem4	qchem5	qchem6	qchem7	qchem8	qchem9	qchem10
1	1									
2	0.04									
3	-0.11	-0.02	1							
4	-0.04	-0.02	0.16	1						
5	0	-0.06	0.07	0	1					
6	0.12	-0.01	-0.05	-0.01	-0.01	1				
7	-0.07	-0.04	0.06	0.04	-0.07	-0.02	1			
8	0.07	0.07	-0.05	-0.02	-0.03	0.11	-0.03	1		
9	0.02	0.05	-0.01	-0.09	0	-0.01	-0.09	-0.04	1	
10	0.03	0.07	-0.04	-0.03	0.05	-0.01	-0.04	0.01	0.05	1
11	0.06	-0.04	0.01	0	-0.03	0.05	0.05	-0.08	0.03	-0.03
12	-0.04	-0.09	0.09	0.04	0.03	0	0.08	-0.03	-0.03	-0.05
13	0.01	0.04	0	0.04	0.02	-0.03	-0.04	-0.09	0.05	0.06
14	-0.02	-0.06	0.05	-0.02	0.08	0.01	-0.01	-0.02	0	0.01
15	-0.03	-0.02	-0.01	0.04	-0.03	0.02	-0.06	0.04	-0.06	-0.01
16	-0.04	0.09	-0.1	-0.03	-0.03	-0.03	-0.05	0.06	-0.02	0.03
17	0	0.08	-0.01	-0.07	-0.04	-0.02	-0.07	-0.02	0.05	0.06
18	-0.04	-0.08	-0.08	0.03	-0.07	0.08	-0.03	0.07	-0.03	0
19	-0.03	0.03	-0.02	0	-0.03	-0.03	0.02	-0.02	-0.03	-0.06
20	-0.05	0.05	0.03	-0.05	0	-0.04	0.03	0.04	-0.01	-0.01
21	-0.01	-0.01	0.05	-0.06	0.05	-0.01	0.04	0.04	-0.01	0.01
22	-0.05	0.02	0.01	0.01	0.05	-0.02	-0.06	0.02	0	0.01
23	-0.03	-0.05	0.02	-0.04	0	-0.1	-0.02	-0.06	0.01	-0.02
24	-0.05	-0.05	-0.03	-0.05	-0.04	-0.04	0.01	-0.02	-0.11	-0.05
25	-0.1	0	0.02	-0.08	-0.04	-0.07	-0.03	-0.02	-0.02	-0.03
26	-0.02	0	-0.02	0.01	0.01	0.01	-0.01	0.06	0	-0.06
27	0.02	0	0	-0.02	0.02	-0.03	-0.01	0	0.03	-0.02
28	-0.03	0.01	0.02	0.01	-0.01	0.04	0.02	0	0.01	-0.01
29	0.04	0.03	-0.01	0.01	-0.03	0	-0.03	-0.01	-0.04	-0.04
30	-0.04	0.01	-0.01	0.03	-0.01	-0.01	-0.03	-0.05	-0.04	0
31	-0.02	0	0.04	-0.01	-0.04	-0.02	-0.01	-0.03	-0.04	0.08
32	0.06	-0.02	-0.05	-0.01	0	-0.06	0.01	0.03	0.02	0.06
33	-0.01	0.01	0.01	0.03	-0.03	0.05	0.03	0.04	0.01	-0.01
34	0.01	-0.05	-0.01	-0.04	0	-0.03	0.02	-0.04	0.07	0.03
35	0.02	-0.03	0.01	-0.04	0.01	0.01	0.04	0.05	0.01	-0.03
36	-0.04	-0.01	-0.03	0.02	-0.03	0	0.02	-0.02	-0.09	-0.03
37	-0.04	-0.04	0.02	0	0.05	-0.03	-0.02	-0.03	-0.02	-0.01
38	-0.01	-0.02	0.06	-0.03	0.02	0.01	-0.05	0.03	0.05	0.07
39	-0.03	0	-0.03	-0.01	-0.03	-0.01	-0.02	0.01	-0.04	0.01
40	-0.04	-0.04	-0.01	0.01	-0.01	0.05	-0.09	-0.06	0	0.02
41	0	0.01	0.03	0.06	-0.1	0.04	-0.01	0	-0.02	0.01
42	-0.06	-0.03	0	0.04	0.04	0.05	0.09	0.03	-0.06	-0.04
43	-0.01	-0.03	-0.03	-0.05	0.05	0	0	-0.06	-0.05	-0.06
44	-0.01	0.01	-0.01	-0.06	-0.01	0.01	0.07	-0.01	-0.02	-0.01
45	0.03	-0.05	-0.03	0.03	0.04	0.05	0	0	-0.02	0.07
46	0.04	0	0.03	0.06	0.03	-0.02	0	-0.02	-0.05	-0.01
47	-0.02	0.02	0.02	-0.02	-0.04	-0.03	0.06	0.03	0	0.01
48	-0.01	0	0	-0.02	-0.06	-0.05	0.01	-0.03	-0.06	0.04
49	-0.02	-0.02	0.03	0.01	0.04	-0.02	-0.01	-0.04	0.04	0.03
50	0.02	-0.03	0	-0.01	-0.03	0.03	-0.01	-0.02	-0.05	0

	qchem11	qchem12	qchem13	qchem14	qchem15	qchem16	qchem17	qchem18	qchem19	qchem20
11	1									
12	0.1	1								
13	0.06	-0.02	1							
14	0	0.03	0.02	1						
15	0.01	-0.03	-0.03	-0.06	1					
16	-0.05	-0.06	0.04	-0.11	0	1				
17	-0.05	-0.02	0.07	0.02	-0.02	0.1	1			
18	0	-0.05	-0.04	-0.08	0.1	0.07	-0.07	1		
19	-0.07	0.04	0.05	-0.01	0.03	-0.09	0.03	-0.06	1	
20	-0.02	0	-0.04	0.1	-0.06	-0.04	0	0.04	0.05	1
21	0.03	0.01	-0.02	0.06	-0.05	0.02	-0.07	0.05	-0.08	0.21
22	-0.04	0	0.01	0.06	-0.02	0.05	0.09	-0.04	-0.11	-0.02
23	-0.06	-0.07	0.01	0.05	-0.03	-0.06	-0.04	0	-0.01	0.03
24	-0.09	-0.08	-0.01	-0.04	-0.06	-0.03	-0.09	-0.08	-0.08	-0.01
25	-0.11	-0.08	-0.04	-0.02	-0.1	-0.04	-0.04	-0.06	-0.05	-0.04
26	-0.01	0.01	0.03	0.03	-0.07	0.03	0.02	0	0.06	0.01
27	-0.02	-0.04	-0.02	0.02	0.03	0.03	-0.06	-0.03	-0.02	-0.01
28	0	0.02	-0.01	-0.05	-0.01	0.02	0	-0.02	-0.05	0
29	-0.03	-0.03	-0.04	0	0.03	0	0.04	0.02	0.02	-0.01
30	-0.01	0	0.04	-0.03	0.01	0.06	0.03	0.06	0.01	0.01
31	-0.01	-0.01	-0.03	0.02	-0.02	-0.02	0.01	0.02	-0.05	0.01
32	-0.03	-0.01	0	-0.03	0.01	0.05	-0.01	0.03	0	0.03
33	0.04	-0.01	0	-0.04	-0.04	-0.02	0	0.03	-0.01	0.06
34	0.01	0.03	0.03	0	-0.01	-0.05	-0.01	0	-0.04	0.04
35	0.02	-0.01	-0.03	-0.02	0	0	0.01	0.01	0.03	0.03
36	0.01	-0.03	0	-0.02	0.02	0.03	-0.02	0.02	-0.03	-0.03
37	-0.02	0	-0.01	0.04	-0.03	-0.01	0.02	0.03	0	0
38	0	-0.02	0.04	0.01	0.01	0.01	-0.05	0	-0.01	-0.02
39	-0.01	0.03	0.02	0.05	0.02	-0.06	0.02	0	0.06	0.03
40	0	0.03	0	0.05	0.06	0	0.02	0.03	-0.03	-0.03
41	0.03	0.04	-0.01	0.02	0.01	0	0.01	0.03	-0.01	0.02
42	0.01	0.03	0.03	0.01	0.01	0	-0.01	0	-0.05	0.04
43	0.01	-0.02	0	0.02	0.02	0	0.01	-0.02	0.03	0.01
44	0.03	0.01	-0.03	-0.06	-0.02	-0.02	0.01	0.01	0.02	0.02
45	0.01	-0.04	0.01	0.01	-0.03	0.04	0.02	-0.02	0	0
46	-0.02	-0.02	-0.01	0.03	0.02	0.02	-0.01	-0.05	0.03	-0.01
47	-0.01	-0.02	0	-0.02	-0.01	0.02	-0.03	0.03	0.04	0.03
48	-0.01	0.03	0	0.02	-0.01	-0.04	0	0.01	0	-0.01
49	-0.01	-0.04	0.05	0.09	-0.03	-0.02	-0.02	0	0.01	0
50	0	0.02	-0.01	0.04	-0.04	-0.1	-0.02	-0.02	0.05	-0.03

qchem21	qchem22	qchem23	qchem24	qchem25	qchem26	qchem27	qchem28	qchem29	qchem30
1									
-0.14	1								
0.06	-0.03	1							
0.01	-0.06	-0.03	1						
-0.02	-0.03	0	0.09	1					
-0.02	-0.02	0	-0.03	-0.01	1				
0.03	-0.04	-0.04	-0.02	0.02	-0.12	1			
-0.04	-0.03	0	0.05	0.03	-0.04	-0.08	1		
-0.03	0.01	-0.06	-0.04	-0.01	-0.08	0.05	-0.01	1	
-0.01	0.01	0	-0.02	-0.05	-0.01	0.08	-0.14	0.02	1
-0.01	0.03	-0.03	-0.01	0.01	-0.04	0.07	-0.11	0.04	0.22
0.02	0	0.02	-0.06	-0.01	-0.01	-0.05	0.17	-0.02	-0.09
-0.01	-0.07	-0.03	-0.06	0.02	-0.02	0	0.15	0.04	0.02
0.06	-0.02	0.01	0.01	0.01	-0.05	-0.01	0.02	-0.13	0.04
-0.07	0.02	0.04	-0.04	0.02	0.04	0.05	-0.03	-0.09	-0.04
-0.02	-0.02	-0.05	0.01	0	-0.04	0.03	-0.02	-0.01	0.13
-0.02	0.02	-0.01	-0.05	-0.02	-0.01	-0.02	0.01	0	0.03
0.05	0.02	-0.01	0.01	-0.03	-0.08	-0.03	-0.04	0.05	-0.02
0.03	0.01	0	-0.03	-0.04	0.04	-0.01	-0.03	-0.12	0.07
0.02	-0.01	-0.01	-0.08	-0.03	0.09	-0.06	-0.09	0.01	0
0.05	-0.01	-0.06	0	0.04	0.04	0.01	-0.02	0.09	-0.1
0.01	0.03	-0.04	0.04	0.01	-0.01	-0.06	0.07	-0.06	-0.12
-0.01	-0.02	0.04	0.03	-0.05	0.05	-0.04	-0.08	-0.06	-0.06
0	-0.01	-0.01	0.05	-0.02	-0.04	-0.09	0.03	-0.05	-0.03
-0.01	-0.05	0.03	-0.08	0.01	0.11	-0.05	-0.02	-0.09	0
0	-0.08	0.04	-0.02	0	0	-0.02	0.01	0.01	0.07
-0.03	0.02	-0.05	0.03	-0.04	-0.06	0.05	-0.05	0.06	0.07
-0.03	0.01	-0.04	0.02	0.02	-0.08	-0.01	0.02	0.03	0.01
0.04	0.03	-0.04	0	-0.06	-0.03	-0.03	-0.07	0.01	0.08
0.04	0	0	0	-0.06	0.02	-0.09	-0.02	0	-0.03

qchem31	qchem32	qchem33	qchem34	qchem35	qchem36	qchem37	qchem38	qchem39	qchem40
1									
-0.22	1								
0	0.05	1							
0.08	0.06	-0.03	1						
-0.05	0.09	-0.01	-0.08	1					
0.17	-0.03	-0.03	-0.04	0.02	1				
0.04	-0.01	-0.04	-0.08	0.1	0	1			
0.04	-0.08	0.04	0.04	-0.12	-0.05	-0.03	1		
0.02	0.02	-0.01	0.02	0.01	0.13	0.02	-0.09	1	
0.05	-0.03	-0.05	-0.01	-0.14	0.02	0.02	0	0.09	1
0.03	-0.05	0.06	-0.06	-0.01	-0.01	-0.09	0.12	-0.01	0.01
-0.09	0.15	0.04	0.04	0.03	-0.05	-0.03	-0.01	0.05	-0.04
-0.03	-0.02	-0.03	-0.01	0.05	-0.03	0.03	-0.02	-0.05	-0.06
0.01	0.08	0.05	0.05	-0.03	-0.11	0.09	-0.05	-0.1	-0.07
0.02	-0.11	0.09	-0.04	-0.1	0.01	-0.06	0.05	0.05	0.09
0.02	-0.06	-0.02	0.01	-0.03	0.01	-0.01	-0.01	0.07	0.1
0.02	0	-0.06	0.01	0.03	0.05	0	-0.03	0.08	-0.08
0.09	0.05	0.03	0.03	-0.05	0.06	0.03	-0.14	0	0.05
0.12	0.02	0.02	-0.04	0	0.05	0.06	-0.05	-0.08	-0.06
0	0.07	-0.02	-0.08	-0.04	-0.06	0.06	-0.01	0.04	-0.03

qchem41	qchem42	qchem43	qchem44	qchem45	qchem46	qchem47	qchem48	qchem49	qchem50
1									
-0.06	1								
-0.09	0.06	1							
-0.16	0.08	0.02	1						
-0.03	-0.07	-0.02	-0.11	1					
-0.03	-0.04	-0.06	-0.02	0.32	1				
0.1	0.02	-0.05	-0.06	-0.13	-0.17	1			
-0.01	0.05	-0.02	0.01	-0.05	-0.11	0.02	1		
-0.03	-0.03	-0.04	0.09	-0.03	-0.08	0.02	0.27	1	
-0.05	0.01	0.05	0	0.01	-0.06	-0.06	0.08	0.21	1

Table 1.2 depicts the correlation of item residual output of local independence of 50-QCT achievement test with the aid of Yen Q3 statistics. Comparing responses on pairs of items, the result revealed that of all the fifty items, items 30 and 31, 45 and 46, 48 and 49 and 49 and 50, violated the assumption of local independence because their correlation item residual was substantially greater than 0.2 cut-off point. Consequently, 46 (92%) items obeyed the assumption.

Research Question 2. Which IRT model (1PLM, 2PLM or 3PLM) best fits the QCAT items?

Model-data fit assessment was established (likelihood-based values statistics) with Bilog-MG to know which of the model (that is one parameter logistic model, two parameters logistic model or three parameters logistic model) best fits the data from QCAT items.

Table 2: Likelihood-based values statistics of 50-QCAT

	Model		
	1PL	2PL	3PL
-2 Log Likelihood	31602.87	31313.94	31065.24

Table 2 shows the values obtained for -2LogLikelihood (-2LL) for each model to establish which model best fit the QCAT items. IPLM was compared to 2PLM and 2PLM indicated better fit. To get the best fit, 2PLM was also compared to 3PLM and the result revealed that 3PLM best fitted the data with the least value. It was concluded that 3PLM best fitted the QCAT items and was therefore used to establish the item parameters of the 50-QCAT items.

Research Question 3. How comparable are the item parameters *(Difficulty and Discrimination indices)* of the QCAT *under CTT and IRT analysis*?

	CTT			IRT			
Item	**Discrimination Index (r_{pbs})**	**Difficulty Index (*p*)**	**Remark**	**Discrimination Index(*a*)**	**Difficulty Index (*b*)**	**Remark**	**Guessing Factor (*c*)**
qchem1	0.2	0.21	Poor	1.63	1.9	Good	0.13
qchem2	0.15	0.38	Poor	0.6	4.88	Poor	0.34
qchem3	0.2	0.24	Poor	0.52	5.8	Poor	0.2
qchem4	0.23	0.29	Poor	0.89	2.05	Good	0.15
qchem5	0.15	0.32	Poor	0.58	4.34	Poor	0.26
qchem6	0.2	0.21	Poor	1.92	2.16	Good	0.16
qchem7	0.2	0.27	Poor	1.71	2.31	Good	0.23
qchem8	0.2	0.26	Poor	2.22	2.1	Good	0.22
qchem9	0.14	0.3	Poor	2.5	1.85	Good	0.25
qchem10	0.26	0.4	Good	2.36	1.88	Good	0.35
qchem11	0.2	0.19	Poor	1.9	1.82	Good	0.11
qchem12	0.23	0.24	Poor	1.76	1.75	Good	0.15
qchem13	0.16	0.3	Poor	0.83	5.57	Poor	0.29
qchem14	0.2	0.32	Good	2.2	2.66	Good	0.35
qchem15	0.21	0.19	Poor	2.04	1.71	Good	0.1
qchem16	0.16	0.36	Poor	2.46	2.23	Good	0.34
qchem17	0.2	0.37	Good	2.25	2.41	Good	0.35
qchem18	0.11	0.22	Poor	2.49	2.46	Good	0.2
qchem19	0.23	0.24	Poor	1.63	1.71	Good	0.14
qchem20	0.27	0.39	Good	1.92	2.34	Good	0.36
qchem21	0.2	0.31	Good	1.02	5.85	Poor	0.31
qchem22	0.03	0.35	Poor	0.82	25.21	Poor	0
qchem23	0.2	0.32	Good	2.3	1.63	Good	0.25
qchem24	0.2	0.2	Poor	2.09	1.26	Good	0.06
qchem25	0.21	0.19	Poor	2.04	1.34	Good	0.05
qchem26	0.13	0.35	Poor	0.81	4.02	Poor	0.31
qchem27	0.02	0.16	Poor	1.31	5.96	Poor	0.16
qchem28	0.09	0.24	Poor	0.82	18.46	Poor	0.24
qchem29	-0.01	0.13	Poor	1.38	5.96	Poor	0.13
qchem30	0.2	0.3	Good	1.04	5.88	Poor	0.29
qchem31	0.21	0.38	Good	0.82	24.18	Poor	0.37
qchem32	0.2	0.26	Poor	1.54	3	Good	0.25
qchem33	0.12	0.21	Poor	1.01	5.81	Poor	0.2
qchem34	0.14	0.29	Poor	0.77	4.51	Poor	0.26
qchem35	0.1	0.3	Poor	0.58	5.81	Poor	0.27
qchem36	0.2	0.4	Good	1	4.52	Poor	0.38
qchem37	0.12	0.35	Poor	1.15	5.91	Poor	0.35
qchem38	0.02	0.15	Poor	0.82	18.45	Poor	0.15
qchem39	0.2	0.34	Good	0.99	5.85	Poor	0.33
qchem40	0.1	0.31	Poor	1.19	5.94	Poor	0.31
qchem41	0.04	0.14	Poor	1.2	5.95	Poor	0.14
qchem42	0.14	0.18	Poor	1.17	5.95	Poor	0.18
qchem43	0.03	0.24	Poor	0.86	4.03	Poor	0.21
qchem44	0.09	0.37	Poor	0.82	21.06	Poor	0.36
qchem45	0.14	0.23	Poor	0.63	5.1	Poor	0.2
qchem46	0.16	0.29	Poor	1	5.94	Poor	0.28
qchem47	0.1	0.26	Poor	1.35	5.96	Poor	0.26
qchem48	0.14	0.2	Poor	0.82	20.81	Poor	0.2
qchem49	0.2	0.35	Good	0.82	18.44	Poor	0.35
qchem50	0.06	0.19	Poor	1.09	5.75	Poor	0.19

Table 3 shows the estimates of the item parameters for all the 50 QCAT items analysed using CTT and IRT frameworks. The result shows that the CTT and IRT frameworks estimated all the item parameters of the QCAT items. Bench mark for good items for CTT was set at $0.30 \leq p \leq 0.80$ and $r_{pbs} \geq 0.20$ and that of IRT was set at $-3 \leq b \leq +3$. Eleven (11) items (10, 14, 17, 20, 21, 23, 30, 31, 36, 39 and 49) that had difficulty and discrimination indices within the bench mark for CTT were considered good items and selected, while 20 items (1,4,6,7,8,9,10,11,12,14,15,16,17,18,19,20,23, 24,25 and 32) that had difficulty indices within the bench mark for IRT were considered good. IRT analysis produced more good items. The result also revealed that the reliability of the test obtained through IRT model (0.60) was better than that of CTT (0.27). The reliability of the test from IRT analysis was moderate while the one from CTT analysis was very low.

DISCUSSION

The findings revealed that the QCAT items satisfied the IRT assumption on dimensionality and it was found to be unidimensional. This is probably so because the QCAT items must have been constructed in such a way that to enable them measure only one dominant latent trait. The result is consistent with the studies of Ayanwale, Adeleke & Mamadelo (2018) and Eleje, Onah & Abanobi (2018) who also found that their test items were unidimensional.

The findings also revealed that the QCAT items satisfied the IRT assumption of local independence. This was perhaps possible because the QCAT items must have been developed in a way to ensure that the correct answer the students gave to an item is independent of the answer the student gave to other items in the test.

Another aspect of the findings revealed that the IRT 3PLM best fitted the QCAT items. The use of the IRT model that best fitted the QCAT items is crucial because it ensures the validity and reliability of the test. This agrees with (Kyung, 2013 as cited in Eleje, Onah & Abanobi, 2018) who observed that the items in a test will be valid if the most compatible model to the data is used for the analysis.

The results on IRT assumptions on local independence and model data fit are consistent with the studies of Nabeel and Chin (2011); Ayanwale, Adeleke & Mamadelo (2018) and Eleje, Onah & Abanobi (2018) whose results revealed that the IRT 3PLM best fitted their test data and also satisfied the IRT assumption on local independence.

Results also showed that more items had difficulty indices within the bench mark for IRT than CTT analysis, which implies that more items survived under IRT analysis and more items were rejected under CTT analysis. This is likely because IRT is a more effective approach to item analysis than CTT. The result corroborates the studies of Zickar & Broadfoot (2010), Eleje, Onah & Abanobi (2018), Ayanwale, Adeleke & Mamadelo (2018) which also revealed that more items were retained from IRT analysis than from CTT analysis. However, the result is contrary to the studies of Meade and Mead (2010), Nabeel& Chin (2011), Ojerinde (2013), Guler, Uyanik & Teker (2014), Awopeju and Afolabi (2016)which did not find disparities between number of survived items under CTT and IRT.

Another aspect of the result revealed that the reliability of the test analysed using IRT was better than that of CTT. The result is consistent with that of Eleje, Onah & Abanobi (2018) who found in their study that the reliability of the test obtained through IRT model was better than that of CTT model. However, the result is contrary to that of Ojerinde (2013), Guler, Uyanik & Teker (2014).

CONCLUSION

This study compared the results from the CTT and IRT 3 parameter Logistic model analyses to ascertain which is more appropriate for the analysis of Quantitative Chemistry Achievement Test items. More test items were rejected from the CTT analysis as a result of poorer difficulty and discrimination indices.

The IRT 3PLM which was found to fit the QCAT items the most produced test items with a higher reliability than that of CTT. IRT can therefore be

considered a better approach of item analysis than CTT and should be preferred while developing achievement test.

RECOMMENDATIONS

Achievement test developers should adopt the IRT approach to test development. However, they should ensure that the IRT model that best fits the test items should be used. They should also ensure that the items for analysis satisfy the IRT assumptions on dimensionality and local independence before using IRT approach.

REFERENCES

Adedoyin, O. O., &Adedoyin, J. A. (2013). Assessing the comparability between CTT and IRT models in estimating test item parameters. *Herald Journal of Education and General Studies*, 2(3), 107-114.

Awopeju. O. A., &Afolabi, E. R. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal*, 12(28), 263-270.

Ayanwale, M. A. (2018). Efficacy of Item Response Theory in the validation and Score Ranking of Dichotomous and Polytomous Response Mathematics Achievement Tests in Osun State, Nigeria. Unpublished PhD Thesis, University of Ibadan.

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An assessment of item statistics estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory approach. *International Journal of Educational Research Review*, 3(4), 55-67. `

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). *Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures*. Retrieved fromhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4096146/

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and ıtem response theory using diagnostic quantitative economics skill test ıtem analysis results. European Journal of Educational and Social Sciences, 3(1), 71-89.

Guler, N., Uynik, G.K., &Teker, G.T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, 2(1), 1-6. http://iassr2.org/rs/020101.pdf

Mead, A.D., & Meade, A.W. (2010). Item selection using CTT and IRT with unrepresentative samples. Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA. Retrieved from http://mypages.iit.edu/~mead/Mead_and_Meade-v10.pdf

Nabeel, A., & Chin, H. L. (2011). The Relationship Between CTT and IRT Approaches in Analyzing Item Characteristics. The Malaysian Online Journal of Educational Science,1(1),52-58. Retrieved from https://www.researchgate.net/publication/316172903_The_Relationship_between_CTT_and_IRT_Approaches_in_Analyzing_Item_Characteristics

Ojerinde, D. (2013). Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results. Retrieved from https://ui.edu.ng/sites/default/files/PROF%20OJERINDE'S%20LECTURE%20(Autosaved).pdf

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance and R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in organizational and social sciences*, 37-60. New York, NY: Routledge.