Archives of Educational Studies
Vol. 3, No. 2, Dec., 2023, 242-257
https://ares.pk/

Construction, Analysis and Calibration of Multiple-Choice Questions: IRT versus CTT

Iqra Batool[1], Ashfaque Ahmad SHAH, PhD[2] & Sehrish Naseer, PhD[3]

[1]Department of Education, University of Sargodha, Punjab,Pakistan.
E-mail :iqraaslam634@gmail.com
[2]Department of Educational Development, University of Baltistan, Gilgit Baltistan, Skardu, Pakistan. Email:multanxa@gmail.com
[3]Department of Education, University of Sargodha, Punjab, Pakistan.
E-mail :sehrishnwarraich@gmail.com

Abstract

The current study examined the construction, analysis and calibration of multiple-choice questions. This quantitative study employed developmental and descriptive methods of research. A convenience sampling technique was used to select a sample of 200 students from the University of Sargodha. The researchers developed a test of multiple-choice items at a master's level from the "Methods of Teaching" course. This test was used as an instrument to collect data from the respondents. Iteman and X-Calibre were considered suitable tools for item analyses for assessment management applications used to analyze the data. Results showed that the test was identified as fairly difficult, having a modest level of item discrimination index. Student raw scores ranged from 7 to 49 marks. CTT proposed to reject seven items whereas IRT removed six based on the item difficulty index. CTT proposed to reject 18 items due to low ability to differentiate between high and low achievers. Six items were flagged with K. Under the S-pbis in CTT, 18 items were rejected and according to IRT's parameter 'b', there were 6 items that were rejected. Results of the current study established that using IRT for item analysis may be useful in determining the grades of the course and the number of students passing the cut-score. It was recommended that before applying IRT, verify if the test items are locally independent one-dimensional and the ICCs fit the model.

Keywords. Construction, Calibration, Classical Test Theory (CTT), Item Response Theory (IRT), Item difficulty, Item discrimination

Introduction

In development of new instrument, latent trait is considered as an important measurement. For this purpose, IRT models were used to determine whether these items or a subset can be combined to form a good measurement tool. With IRT, it has been evaluated that how much amount of information each item provides; If some items do not provide much information, these may be eliminated. IRT models estimate the difficulty of each item; it informs about the level of the trait that is assessed by an item. Those items that provide information across the full continuum of the latent trait scale are incorporated; these items provides how much information an instrument provides as a whole for each level of the latent trait. Researchers intended to study the application of IRT in comparison with CTT on a rigorously self- developed test of multiple-choice items on a master's level course of "Methods of Teaching".

Different statistical models are used to characterize items as well as test taker individually. Item response theory (IRT) is used for the designing, analysing and scoring of tests, questionnaires and similar instrument which measures the abilities, attitudes and other variables. It is also known as latent trait theory, strong true score theory and modern mental test theory (Molenaar, 2002). It is a theory of testing based on the relationship between performance of an individual on test items and the levels of performance of test takers on an overall measure of the ability.

IRT is recognized as modern theory which is a requirement of recent era. This is the era of technological world where fast and validated results are required. CTT (Classical test theory) as compared to IRT is considered to be forcing linear model on something that is known to be nonlinear. In short, for valid, reliable, accurate and precise results and computer adaptive testing, IRT is found important and necessary application in recent century. Classical test theory (CTT) and item response theory (IRT) are two theoretical frameworks that are frequently used to evaluate the psychometric properties of tests or the degree to which the statistical properties of a test support its intended interpretation and use (Algina & Penfield, 2009). Traditionally, researchers have relied on CTT in developing and scoring personality assessment instruments, whereas

In educational testing, IRT has been dominated at a large scale (Embretson & Reise, 2000). However, the use of IRT modelling is more common in personality related studies (Waller & Reise, 2010). Although CTT and IRT subsume different measurement models and statistical methods, statistics generated from each framework tend to be complementary in practical testing applications (Thissen & Orlando, 2001). In fact, some researchers have argued that CTT analysis should usually precede IRT analyses so that items of poor psychometric quality can be screened out prior to IRT analyses (Morizot et al., 2007). Although CTT and IRT are frequently used to develop personality measurement instruments, both can also be productively used to refine the existent inventories. In the present study, we primarily used IRT measurement models to evaluate and shorten a measure of locus of control.

IRT focuses on item scores rather than on test scores. The estimated and true development score scales for each test were compared to examine the variability and the separation of grade distribution for examinees at different grade level (Topczewski, 2013). Item Response Theory (IRT) is better than CTT as it:

- Provided good results in psychometric test as compared to CTT
- Determined the level of trait of the test
- Provided a greater level of information per item (Virginia, 2001)

Uni-dimensionality and Multidimensionality depends upon the dispersion of data i.e., latent trait (θ) showed the narrowness of the data. The narrower data was unidimensional and data having more dispersion was multidimensional (Williamson, 2010).

The low scores achieved by an item does not verify that the item was not good and high scores achieved by an item were meant that item was very good. The scores are just the representation of the respondents' opinions on a certain topic (Trang, 2013). IRT requires larger data for better fit and interpretation. It is clear that the choice between CTT and IRT is considered to be a better choice for the data when it is being analysed at item level. As it improves the reliability of the scores, it never means that IRT had superiority on CTT; both CTT and IRT often occupy similar results. When scores developed by IRT can be correlated with the scores developed by CTT, its indicated that the two sets of scores correlated at a higher level; there would be hardly a minor difference of the both data sets (Nasir, 2014). The advantages of IRT includes improvement in content

coverage, reduce the construction of irrelevant variance, improved efficiency, fewer test items and shorter test time (Tayn, 2010).

Objective type tests are used to measure the higher order skills because it was observed that the item which falls in the domain of higher cognitive skill are difficult and thus those students who failed to give the correct responses achieved a penalty in the terms of losing the scores. (Zaman et al., 2008). This section gives information about IRT and CTT. It raises questions like why IRT is applied, how it is applied and when it is needed to apply; why IRT is preferred over CTT. It also contains the brief descriptions about different model of IRT. It includes the advantages of using IRT over CTT. In the last section, it includes the information about Iteman and xcalibre which is used for the application of IRT on the given test items.

Objectives of the Study

The major objectives of the study were to:

- develop quality test items from the subject "*Methods of Teaching"* at masters level using table of specification
- do item analyses by using IRT and CTT
- compare IRT and CTT on the basis of the analysis

Research Questions

- What was the level of item difficulty and item discrimination on multiple-choice tests (MCQs) using a comparison of both types of analysis (CTT and IRT)?
- What are the main differences between IRT and CTT on the basis of the analysis?

Methodology

This section contains the information about the design and procedure that was adopted for the current study. Main purpose of the study was to determine the ‘Construction Analysis and Calibration of Multiple-choice Questions from "Methods of Teaching". This study carried out by using quantitative research method; study was descriptive in nature. The population includes all the male and female students studying "Methods of Teaching" at master’s level. The sample of the study consisted 200 male and female students studying at the Department of Education, University of Sargodha. Convenience sampling technique was used to draw sample from the population. A test of multiple-choice questions in course of

"Methods of Teaching" was developed by the researcher and subsequently used as a research instrument. The researcher prepared a comprehensive course outline from various courses related to "*Methods of Teaching"* which comprised 12 chapters. Table of specification was made according to the outline; objectives were chosen on the basis of cognitive domain of the Bloom's taxonomy. Sixty questions were made; Answer key was prepared to assess the test.

For the sake of validation of research instrument, three experts reviewed it and suggested improvements which were incorporated. The improved version of the test was presented again to the same experts to verify if the modifications were made correctly. Forty respondents were conveniently selected by the researcher from the whole population of the study for the pilot testing. The respondents were given the test individually at a time. The respondents were seated in the traditional way of testing. The time given to the respondents was one hour i.e., one minute per item. After initial item analysis, the test was further modified and thus final version of the research instrument was prepared which comprised 49 items. The next phase was to collect data from the sample. Sample of 200 students were taken from BEd, MEd, BS (Education) and MA Education programmes. Test was conducted according by following traditional way and the testing protocol. Tests of the students were marked according to the revised answer key. After cleaning, data was analysed by using MS Excel, Iteman and XCalibre software and the results were reported, analysed and discussed to draw the conclusion.

Iteman and X-Calibre are software programmes designed to provide comprehensive item and test analysis reports. The purpose of these reports is to provide assistance in testing programs and to evaluate the quality of test items and tests as a whole, by examining their psychometric characteristics. These reports are generated in Word/RTF design, allowing researchers to type/paste in item texts and comments and provide a complete report to stakeholders or content experts. Reliability of the test was calculated in traditional way by using formula:

$$reliability = \frac{k}{k-1}(1 - \frac{\sum PQ}{\sigma^2})$$

Reliability of the test for pilot testing was 0.58 and overall reliability of the final test was 0.56. Every item was analysed separately. Item was analysed by using formula:

$$pvalue = \frac{R}{T}$$

Where P = difficulty level

R = number of students who got item right

T = total number of students

The discrimination index of each test item is calculated by the help of the following formula:

$$D = (RU - RL)/(0.5T)$$

D= Discrimination index

RU= students in the upper group who get the item right

RL= students in the lower group who get the item right

T= Total number of the students (Linn and Gronlund, 2000)

According to formula, 27% students of upper and 27% of lower group were selected for the purpose of test analysis. Many researchers used total 54% of the students for this purpose (Higrorjo & Jaleel, 2012; Backhoff et al., 2000; Mitra et al, 2009; Sim & Rasiah, 2006) used 27% upper and 27% lower group of the students for item analysis. Items with values ranging from 0.20 to 0.80 were selected while rest of the items were discarded. The following criterion was followed by the different studies (Shah, 2005; Naseen, 2011).

Results

Table1

Item Parameters for CTT and IRT – Comparison

Item ID	Classical Test Theory		Item Response Theory		
	Difficulty (P)	Discrimination (R)	Discrimination (a)	Difficulty (b)	Pseudo guessing (c)
1[K]	0.42	0.191	0.582	2.618	0.330
2	0.42	0.376	0.639	1.728	0.315
3	0.64	0.378	0.628	0.226	0.398
4	0.50	0.251	0.540	0.882	0.249
5	The item was removed for its low proportion of correct response				
6	0.42	0.280	0.599	1.412	0.248
7	0.48	0.369	0.662	0.777	0.246
8	0.32	0.179	0.654	2.517	0.248
9	0.50	0.502	0.733	0.452	0.243
10	0.56	0.265	0.561	0.376	0.249
11	0.46	0.205	0.588	1.252	0.251
12	0.30	0.134	0.653	2.907	0.249
13	0.44	0.119	0.563	1.692	0.254
14	0.62	0.329	0.561	-0.043	0.249
15	0.46	0.323	0.589	1.129	0.249
16	0.54	0.258	0.569	0.549	0.249
17	0.68	0.191	0.515	-0.347	0.252
18	The item was removed for its low proportion of correct response				
19	0.34	0.372	0.633	1.738	0.242
20	0.38	0.244	0.587	1.893	0.249
21	0.32	0.247	0.643	2.237	0.246
22	0.54	0.365	0.586	0.449	0.248
23 [K]	0.28	0.107	0.672	2.721	0.245
24	0.44	0.242	0.595	1.311	0.249
25 [K]	The item was removed for its low proportion of correct response				
26	0.32	0.135	0.650	2.643	0.250
27	The item was removed for its low proportion of correct response				
28	0.36	0.170	0.599	2.206	0.250
29	0.30	0.305	0.644	2.349	0.244
30	0.38	0.335	0.657	1.496	0.245
31	0.42	0.298	0.590	1.534	0.250
32 [K]	0.46	0.032	0.561	1.900	0.260
33	0.58	0.280	0.568	0.297	0.250
34	0.44	0.206	0.563	1.521	0.252
35	0.32	0.285	0.629	2.305	0.246
36	0.38	0.335	0.606	1.702	0.247
37	0.38	0.342	0.642	1.433	0.244
38	0.32	0.496	0.679	1.566	0.238
39	The item was removed for its low proportion of correct response				
40 [K]	0.36	0.015	0.636	2.805	0.256
41	0.52	0.251	0.577	0.732	0.250
42	0.44	0.289	0.603	1.302	0.249

43 [K]	The item was removed for its low proportion of correct response				
44	0.58	0.328	0.602	0.275	0.250
45	0.42	0.173	0.563	1.599	0.250
46	0.56	0.373	0.637	0.246	0.247
47	0.38	0.397	0.622	1.556	0.245
48	0.58	0.400	0.643	0.124	0.247
49	0.44	0.500	0.705	0.732	0.240

Teaching, learning and assessment are three identifiable entities that come together to make a single monolithic whole, may be referred to as education. These three entities are indeed inseparable especially at higher education level. From a classroom to the standardized context, assessment is of critical importance for assuring the success as well as quality of the education process.

Classroom assessment essentially may yield fallacious output for the students if devoid of certain theoretical framework i.e., classical test theory (CTT) and/or item response theory (IRT). Present analyses of test of MCQs offered a vivid example of the case. The raw score of the students showed that there were only 7.5% of students who crossed the cut score of 50%. If employed CTT, which suggested deleting 16 items from the test, this percentage will rise to surprising level of 50%. Going further, use of IRT removed 6 items besides ascertaining the intrinsic hardness in the test items through their parameters i.e., difficulty, discrimination and pseudo guessing; therefore, suggesting to revise student scores/grades or to propose relative grades as is the case in norm referenced assessment.

In this analysis of multiple choice items, the researcher used MS Excel, Iteman and X-Calibre for CTT as well as IRT to understand the comparative statistics of the two theories. CTT model is relatively simple at item level. It does not make a complex theoretical model to relate student ability to success on an item. In fact, it takes into account a group of students and empirically examines their achievement rate on an item scored dichotomously. On the other hand, the IRT model emphasizes on both item and person statistics; and this is what makes IRT a better option in giving adequate information on the behaviour of item as well as student. The test was found to have high internal consistency (IRT Alpha = 0.82). IRT calibrated 43 items (using X-Calibre); whereas total items were 49. The test was identified as fairly difficult having a modest level of item discrimination index. Following conclusions were made on the basis of findings. Student raw score ranged from 7 (student 13 and 164) to 49 marks

(student 153). Only 15 students (7.5%) got 50% marks if looked at raw scores.

CTT (through MS Excel and Iteman) proposed to reject seven items (5, 12, 23, 29, 35, 38, 39); whereas, IRT (through X-Calibre) removed six (5, 18, 25, 27, 39, 43) on the basis of item difficulty index. CTT (through MS Excel and Iteman) proposed to reject 16 items and to revisit other 14 items on the basis of item discrimination index. CTT (through X-Calibre) proposed to reject18 items (Item IDs 5, 7, 9, 10, 12, 18, 22, 23, 25, 29, 31, 36, 40, 41, 42, 43, 47 and 48) were bad items as they have no ability to differentiate between high and low achievers. Six items (Item IDs 1, 23, 25, 32, 40 and 43) were flagged with K (which indicated that the keyed alternative (true option) did not have the highest correlation with total score i.e., T-Rpbis)

Pseudo guessing was found ranged from 0.238 to 0.398 with a mean value of 0.255. Seven items (Item IDs 44, 33, 41, 31, 45, 28, and 26) were identified with exactly 0.25 value of 'c' parameter. Nine items (Item IDs 11, 17, 34, 13, 40, 32, 2, 1, and 3) had 'c' value ranged from 0.251 to 0.398. Rest of the (27) items had 'c' value ranged from 0.238 to 0.249. Under the p value in CTT six items were rejected and according to IRT's parameter a there were also six items, these were same in number but not the same items which were rejected except Item IDs 5 and 39. Under the S-pbis in CTT 18 items were rejected and according to IRT's parameter 'b' there were 6 items which were rejected. Only the items that were rejected by the application of both theories were Item IDs 5, 18, 25 and 43.

Our analyses yielded value of alpha 0.28 for CTT and 0.82 for IRT. Ojerinde (2013) stressed that IRT provides better reliability. Present research traced local independence among the test items through values of "Alpha w/o". Joshua, Ubi &Abang (2011), as cited by Ojerinde (2013), carried out a large study and found that the (Mathematics Test) items were locally independent i.e., the test items did not clue to one another during the testing session.

Discussion

Present analyses of the test items found that IRT may be better way to student assessment of ability. Supporting the argument of the current study Tang et al., (2023) established that using IRT scores to calculate significant individual improvements and identify treatment responders is a better option because they perform better, or least comparably, in the majority of

circumstances. Similarly, it is evident that weighing the total scores using IRT parameters can increase the score when utilizing the 9Q to assess the severity of depressive symptoms in Thai people (Kawilapat et al., 2022).
Holding the stance of the present study, it is discussed that IRT can be used to enhance measurement in the field and has a number of advantages over CTT. For instance, compared to scales made using CTT, measures made using the IRT technique are substantially more exact and require fewer elements
(Anderson et al., 2020). It is established that IRT-based ratings were more responsive than CTT-based scores in the early stages of the disease, demonstrating the IRT-based scores' improved applicability for use in preclinical settings (DubbeIman et al., 2023).
Similarly, Idowu et al., (2011) established excellence of IRT over CTT in determining the assessment quality of Mathematics items. Ojerinde (2013), highlighted that CTT did not rely on total score for signifying the measured ability; therefore, researchers (De Ayala, 2009; Welch & Hoover, 1993; Idowu et al., 2011), got attracted towards IRT for its potential to make such allowances. Ojerinde et al., (2012) found that there was a prominent progress in the item statistics using IRT compared to CTT. Our item analyses described that many items were rejected by CTT.
On contrary, a study conducted by Bichi et al., (2019) discussed that CTT and IRT frameworks seems to be more useful and reliable in assessing items of test because two frameworks provide same and comparable results. It is established that approaches of items analysis work more effectively with integration in aspects of item development and evaluation so that measurement errors can be reduced. Azevedo et al., (2019), discussed that study revealed internal consistency and, as a result, some reliability in the question bank. In order to create assessments that are as fair as possible, it is crucial that the teacher get a set of questions having similar characteristics. Adedoyin (2010) observed the invariance of individual parameter estimates based on classic test and item response theory and came to the conclusion that the examinee's ability or score was testing dependent that is influenced by the specific collection of items used. Literature verified that estimated individual trajectories using item response theory, as opposed to classical test theory, provide a more thorough description of individual change over time because patterns of item response theory are more informative than classical test theory (Gorter

et al., 2019). Yasar (2019) claimed that IRT has many favourable aspects and it is proved better in various aspects than CTT because it covers the limitations caused by CTT. It is argued that CTT and IRT may be used as complimentary techniques in developing national exams since they were equivalent in evaluating item characteristics of statistical and psychometric tests (Awopeju & Afolabi, 2016). Oyiborhoro (2023), discussed that based on the results of the study, IRT and CTT played a complementary role in development of test; parameters of both approaches contribute significantly in test development.

According to Fan (1998), sample dependence has restricted CTT to be used in certain specific measurement situations like test score equating, item banking and computer adaptive testing. Ojerinde (2013) concluded that CTT has an advantage of simplicity, the sample dependency of item and test statistics limit their usefulness and utility in psychometric analysis; however, IRT has been used effectively in test score equating and item banking.

Conclusion

We conclude with Thorndike's remarks. For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. (Thorndike, 1982). This study showed that using IRT for item analysis may be useful in determining the course grades and the number of students passing the cuts-core. Lack of quality assurance process may pose a significant reputational risk to the institution (Brown & Abdulnabi, 2017), therefore, such kind of item analyses are found beneficial for fairer assessment, particularly, where probably the test does not conform to the normal requirements.

Recommendations

It was recommended to verify before applying IRT if the test items were local independence (no clue to other items), one-dimensional and the ICCs fit the model. Moreover, it was recommended that CTT and IRT both be used together; the defects of CTT could easily be compensated for by that of IRT and made to be complementary to it. It is recommended that IRT and CTT may be made core modules in undergraduate programmes,

graduate research and retraining of academics through workshops, lectures and seminars. Current study was delimited to the students of department of Education, University of Sargodha; similar study may be conducted by using a different and larger sample size for more generalizability of the results.

Author's Contribution

IB conducted the study and prepared the manuscript. AAS conceptualize the idea and incorporated the requisite changes subsequently. SN refined the final version of manuscript and incorporated suggested changes. All authors contributed to the article's planning, reading of draft and approved the submitted version.

Funding

The author(s) received no specific funding for this study.

Conflict of Interest

Researchers declares no conflict of interest.

Ethics Statement

According to local legislation and institutional requirements, ethical review and approval was not required for the study on human participants. while preparing this manuscript, all other ethical considerations were fulfilled and consent of participants was taken.

Acknowledgements

Researchers pays heartiest gratitude to all those who has contributed to this study and Respected faculty members of the Department of Education, University of Sargodha for their valuable suggestions during the completion of research; they extended all possible and available facilities for this research work. We are extremely grateful to our family members for their love, prayers and sacrifices.

References

Adedoyin, O. (2010). An Investigation of the effects of teachers' classroom questions on the achievements of students in mathematics: Case study of Botswana community junior secondary schools. *European Journal of Educational Studies, 2(3).*

Algina, J. and Penfield, R. D. (2009). *Classical test theory. In R. Millsap & A. Maydeu-Olivares*

(Eds.). The Sage handbook of quantitative methods in psychology (pp. 93-122). Thousand Oaks, CA: Sage. Retrieved from csus-dspace.calstate.edu.

Anderson, S. R., & Miller, R. B. (2020). Improving measurement in couple and Family Therapy: An item response Theory Primer. *Journal of Marital and Family Therapy*, 46(4), 603-619.

Awopeju, O. A., & Afolabi, E. R. I. (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), 263-284.

Azevedo, J. M., Oliveira, E. P., & Beites, P. D. (2019). Using learning analytics to evaluate the quality of multiple-choice questions: A perspective with classical test theory and item response theory. *The International Journal of Information and Learning Technology*, 36(4), 322-341.

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Bin Ibrahim, A. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. *International Journal of Engineering and Advanced Technology*, 8(5), 1260-1266.

Dubbelman, M. A., Postema, M. C., Jutten, R. J., Harrison, J. E., Ritchie, C. W., Aleman, A., ... & Sikkes, S. A. (2023). What's in a score: A longitudinal investigation of scores based on item response theory and classical test theory for the Amsterdam Instrumental Activities of Daily Living Questionnaire in cognitively normal and impaired older adults. *American Psychological Association.*

Embretson, S. E. and Reise, S. P. (2000). Item Response Theory for Psychologists. *Mahwah, NJ: Lawrence Erlbaum Associates*. Retrieved from https://support.sas.com.

Fan, X (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurem... June 1998 v58 n3 p357(25).* Retrieved from https://www.ncbi.nlm.nih.gov. http://www.assess.com/item-response-theory/

Gorter, R., Fox, J. P., Riet, G. T., Heymans, M. W., & Twisk, J. W. R. (2020). Latent growth modeling of IRT versus CTT measured longitudinal latent variables. Statistical methods in medical research, 29(4), 962-986.

Idowu, E. O .Eluwa, A.N. and Abang, B.K. (2011). Evaluation of Mathematics Achievement Test: A Comparison Between Classical Test Theory (CTT) and Item Response Theory (IRT). *Journal of Educational and Social Research,1(4):99-106.*

Joshua, Ubi and Abang, (2011). Classical Test Theory (CTT) VS Item Response Theory (IRT) an evaluation of the comparability of item analysis results by prof. Retrieved from ui.edu.ng.

Kawilapat, S., Maneeton, B., Maneeton, N., Prasitwattanaseree, S., Kongsuk, T., Arunpongpaisal, S.,& Traisathit, P. (2022). Comparison of unweighted and item response theory-based weighted sum scoring for the Nine-Questions Depression-Rating Scale in the Northern Thai Dialect. BMC Medical Research Methodology, 22(1), 1-15.

Le, Dai-Trang, (2013). Applying item response theory modeling in educational research . *Graduate Theses and Dissertations. 13410*. Retrieved from http://lib.dr.iastate.edu.

Linn, R. L. and Gronlund, N. E. (2000). *Measurement and Assessment in Teaching. Eighth Edition.* Retrieved from eric.ed.gov/?id=ED435651.

Molenaar, I. W. and Sijtsma, K. (2002). Non parametric item response theory *International educational and professional publisher thousand oaks London.* Retrieved from https://books.google.com.

Morizot, J. Ainsworth, A. T. and Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. *In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology*

Nasir, M. (2014). *Application of classical test theory and item response theory to analyze multiple choice questions*. Retrieved from theses.ucalgary.ca.

Ojerinde, D. Onoja, G. O and Ifewulu, B. C. (2014). A Comparative Analysis of Candidate's Performances in the Pre and Post IRT Eras in JAMB on the Use of English language for the 2012 and 2013 UTME. *A paper presented at the 39th IAEA Annual Conference in Tel Aviv in Israel.* Retrieved from www.iaea.info.

OYIBORHORO, A. V. (2023). Application of Item Response Theory in the Validation of Basic Science Test of Delta State Basic

Education Certificate Examination. International Journal of Research in Education and Sustainable Development, 3(7), 1-13.

Reise, S.P. and Waller, N.G. (2003) *How many IRT parameters does it take to model psychopathology items? Psychol.Meth. 8:164–84* . Retrieved from www.psycholosphere.com.

Tayn, K.S. (2010). *An evaluation of multiple-choice test questions deliberately designed to include multiple correct*. Retrieved from scholarsarchive.byu.edu.

Tang, X., Schalet, B. D., Peipert, J. D., & Cella, D. (2023). Does scoring method impact estimation of significant individual changes assessed by patient-reported outcome measures? Comparing Classical Test Theory versus Item Response Theory. Value in Health.

Thissen, D., and Orlando, M. (2001). Item response theory for items scored in two categories. *In D. Thissen & H. Wainer (Eds.), Test Scoring. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.* Retrieved from faculty.psy.ohio-state.edu.

Thorndike, R. L. (1982). Educational measurement: Theory and practice. In D. Spearritt (Ed.) The improvement of measurement in education and psychology: *Contributions of latent trait theory. Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations. (ERIC Document Reproduction Service No. ED 222 545).*

Topczewski, A. M., Kapoor, S. and Cunningham, P. (2013). Examining the parameter recovery of BILOG-MG 3 and WinBUGS 1.4.3. *Poster presented at the Annual Meeting of the National Council on Measurement in Education San Francisco California.* Retrieved from ir.uiowa.edu.

Waller, N. G. and Reise, S. P. (2010). Measuring psychopathology with non-standard IRT models: Fitting the four-parameter model to the MMPI. In S. E. Embretson (Ed.), *Measuring psychological constructs with model-based approaches,* pp. 147-173.

Williamson, L.M. (2010). *An item response theory revision of the internal control index.* Diss. California State University, Sacramento, 2012. Retrieved from csus-dspace.calstate.edu.

Yaşar, M. (2019). Development of a “Perceived Stress Scale" based on Classical Test Theory and graded response model. *International Journal of Assessment Tools in Education*, 6(3), 522-538.

Zaman, A. Kashmiri, A. R. Mubarak, M. and Ali, Arshad. (2008). *Students Ranking, Based on their Abilities on Objective Type Test: Comparison of CTT and IRT*. *Research online Institutional Repository*.