ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2023; 2; 38-48

Comparative Psychometric Properties of Electronically-Tested and Conventionally-Tested Items of a Selected University Course in Nigeria

[a]Olumide Dorcas Olagundoye* & [b]Eyitayo Rufus Ifedayo Afolabi

[ab]Department of Educational Foundations and Coumselling, Obafemi Awolowo University, Nigeria

Abstract

The study assessed the comparative psychometric properties of electronically and conventionally tested items of a selected university course in Nigeria using classical test theory and item response theory models. These were with a view to providing information on the overall quality of electronic mode of testing (E-Testing). The results showed that the electronic mode of testing was moderately reliable at KR20 value of 0.73 with an acceptable level of content validity. Also, the conventional mode of item testing was equally reliable with KR value of 0.70. This indicated that the two modes of testing were moderately reliable with a slight difference of 0.03. Item analysis of the testing modes showed that both theories produced similar results on the difficulty and discrimination strengths of the items. Comparing assessment modes, result revealed that they were significantly equivalent and homogenous at Z2 = 0.75. The SDs of 4.72 and 4.44 for the testing modes indicated that for C-testing, majority of the examinees scored between 15 and 25 marks while for E-testing some of the examinees scored between 14 and 23 marks. Item analysis showed that the models harmoniously supported the acceptability and otherwise of 17 and 18 items for the testing modes respectively in terms of the difficulty and discrimination strengths of the items. In conclusion, the psychometric properties of the assessment modes were comparable in terms of relationship, reliability, difficulty, and discrimination strengths of the items with E-Testing being more effective in terms of the reliability of test mode effects.

Keywords: psychometric properties, electronically-tested, conventionally-tested

* Corresponding author.

Department of Educational Foundations and Coumselling, Obafemi Awolowo University, Nigeria, e-mail: olumideolagundoye08@gmail.com

To cite this article:

Olagundoye, O. D. & Afolabi, E. R. I. (2023). Comparative Psychometric Properties of Electronically-Tested and Conventionally-Tested Items of a Selected University Course in Nigeria. *Journal of Computerized Adaptive Testing in Africa, 2*, 38-48.

Received: 3 November 2023
Accepted: 30 November 2023
Published: 23 December 2023

INTRODUCTION

It has been observed that attention on assessment procedure in recent times is been shifted from the generally recognized traditional (C-testing) procedure to electronic (E-testing) procedure. Assessment items play a vital role in the cognitive assessment of examinee, it is an instrument designed to obtain information about examinee's progress or otherwise in an instructional process, Jacqueline, Mark, and Stephen, (2004). Assessment items and procedures are generally expected to have acceptable psychometric properties (validity, reliability, item difficulty and discrimination indices). For an item to be generally acceptable, it must successfully measure the real characteristics of an examinee in terms of the content it is designed to measure.

Modern psychometrics originated from two major sources, namely: the classical test theory (CTT), and the item response theory (IRT), Furr and Bacharach (2008). The classical test theory is aimed at explaining the total final score of examinee, that is the sum total of responses provided to a series of items by examinee, while the item response theory is concerned with developing quality test items, that is, determining the difficulty and discrimination level of individual item in a given task, Eleje, Chidiebere, and Fredrick (2018), Carlo, (2009). The combination of both classical test theory (CTT) and item response theory (IRT), therefore, leads to the development of reliable and valid assessment instrument(s).

The two modes of assessment (electronic and conventional) involves the activities of test experts in developing a pool of items for assessment but differs in assessment mode in the sense that, conventionally, test takers are subjected to assessment items on paper following the sequence in which the items are being arranged but electronically, they (tests takers) are subjected to assessment items as displayed on the screen, Folk and Smith (2002), Wending and Noriko, (2019). This study was therefore, designed to bring into focus the Comparative Psychometric Properties of Electronically and Conventionally Administered (E-testing and C-testing) Items of a course (Philosophical Foundation of Education) in the selected University. This title is developed by the reason that, attention on higher educational institution assessments of larger population courses is being shifted from the commonly accepted conventional-based (C-testing) to electronic-based (E-testing), this is in-line with the assertion of Ojerinde and Wieger, (2015) and with the aim to establishing the statistical qualities of reliability, validity, item difficulty, and discrimination indices of assessment item(s) used for students' assessment.

Statement of the Problem

The inclusion of Information and Communication Technologies (ICTs) in education for cognitive assessment due to the demands which society places upon its citizens to find and produce knowledge, has over time led to the need for the consideration, modification and /or change of the traditional (C-Testing) methods of assessment both at the primary, secondary, and tertiary levels of education to modern (E-Testing) methods of examination. Over time, literature evidence abounds on the psychometric characteristics of different types of conventionally – Tested items in terms of their mean scores with little or no attention on the criteria of adequate reliability, validity, as well as the difficulty and discrimination indices of assessment items but information relating to how they compare with those tested electronically is sparse in literatures. It is therefore necessary to estimate the psychometric properties (reliability, validity, item difficulty, and item discrimination indices) of these modes of item testing (electronic and conventional) to ascertain their psychometric equivalence or otherwise.

The specific objectives of this study were to;

1. examine the reliability and validity of electronically administered test items;
2. assess the reliability of conventionally administered test items;
3. determine the level of agreement of the psychometric properties of electronically and conventionally administered test items.

Research Questions

1. Are the test items administered using electronics (E-Testing) acceptable enough to meet the required index for reliability and validity?
2. Will the reliability and content validity of conventionally administered test items meet required index?
3. What will be the level of agreement of the psychometric properties of electronically and conventionally administered test items using CTT and IRT?

Research Hypothesis

H01: There is no significant relationship between the score of students in conventional (C-Testing) and electronic (E-Testing) testing modes

METHODOLOGY

The study adopted survey research design; it involves obtaining data (eliciting information) from a purposively selected sample of a compulsory year II undergraduate course, in the Department of Educational Foundations and Counselling of the selected University. The selected course being one of the courses in the Department with larger population and where students were assessed both on computer (electronic) and on paper (conventional) for the semester the study was conducted. The selected course was registered for by an intact class (population without manipulation) of year II undergraduates from eight departments in the chosen Faculty for the study in the selected University. The study was interested in comparing the psychometric properties of the 30 electronically and conventionally administered test items as the same set of items used for the conventional aspect of the assessment were been fixed for the electronic aspect of the assessment, which made the examinees to be subjected to a section of homogenous thirty items for all and another section of heterogeneous items for all. The psychometric properties of the items considered included; reliability (internal consistency of the items), content validity, item difficulty and discrimination indices.

The population of the study comprised all undergraduate courses in the Faculty of Education of the selected institution for the study in Osun State. The study sample consisted of a purposively selected year II undergraduate course registered for by an intact class of 637 year II candidates in the Department. The study covers a total of 628 students that participated both in the traditional (C-Testing) and electronic (E-Testing) testing modes. The study sample, a compulsory year two undergraduate course, being one of the courses with large population and where students were assessed both on computer and paper in the sampled department was drawn using a purposive sampling technique. An intact class of 628 Year II undergraduate students who participated both in the conventional test (C-Testing) and the electronic test (E-Testing) out of 637 students who registered for the compulsory course were selected from eight departments in the Faculty. The 628 study participants were subjected to the same set of test items generated for the study sample using the two modes of testing (electronic and conventional) at an interval of one month

In carrying out the study, 30 multiple-choice test items developed by the course lecturer(s) on the selected course for students' assessment for the semester, was used to obtain information from the participants based on their (participants) responses to the test items both on computer (electronic-based) and on paper (conventional-based) at different interval. The research instrument, a 30 item multiple-choice test developed by course lecturer(s) for students' assessment on the selected course for semester assessment of students. These items were school-based assessment instruments that were assumed to have been previously validated by the course coordinator and other course lecturer(s) in the selected Department before they were administered to the study participants. However, the course coordinator together with the course lecturer through thorough checking and comparison of the items with the course content for the semester, confirmed that the items were constructed within the speculation of the course content, this then confirmed the content validity of the items.

The data collection process lasted for a period of one year, the researcher visited the course coordinator, the lecturer-in-charge of the course, the Dean of Computer Science, the HOD of Information Communication Technology (ICT), and the CBT coordinator as well as other concerned lecturer(s) to seek their consent. The researcher was present during the period of first administration of the items to the participants using paper-and-pencil (conventional-based) for their (study participants) mid-semester test, the researcher also went to the university ICT (information, communication and technology center) to request for individual participants' aggregate score as well as their individual responses to each of the items in the second stage of the assessment (Computer-based). The two data (results on conventional and electronic based) together with the test items used as well as the keys (correct options) were collected by the researcher after the computer-based test (E-Testing) stage to confirm each candidate's responses to each of the test items, which is used to determine the difficulty and discrimination indices of the test items as well as the comparability (validity and reliability) of the participants' scores in the two testing modes. The data collected from the study were analyzed in line with the research questions and hypothesis as follows.

Research questions one and two were subjected to reliability of examinees' scores in electronic and conventional based testings using internal consistency method of Kuder and Richardson (KR 20) developed by Katrina Korb , while the content validity of the items was conducted in line with the Table of Specification Guidelines developed by Charles Kilbler. Research Question 3 was subjected to item analysis of difficulty and discrimination indices using Psych statistics. The research hypothesis was analyzed using Pearson Correlation Coefficient (R) developed by Karl Pearson. Analysis was done with Validate R (R Package).

RESULT

Research Question One: Are the test items administered using electronics (E-Testing) acceptable enough to meet the required index for reliability and validity?

Table 1: Reliability Coefficient Table for C-Testing and E-Testing in EFC 202 Assessment Items

Examination mode	N	Reliability Coefficient KR20	Remark
C-Testing	30	0.70	Reliable
E-Testing	30	0.73	Reliable

The reliability coefficient of 0.70 and 0.73 showed that the test items for the two test modes (C-Testing and E-Testing) are moderately reliable.

Table 2: Content Validity for 30 Multiple-Choice Items for E-Testing and C-Testing in Philosophical Foundation of Education

S/N	Topic	Objective Remembering	Understanding	Thinkin g	Total
1.	Philosophical School of Thought	4	6	1	11
2.	Branch of Philosophy	1	1	1	3
3.	Methods of Teaching Philosophy	1	1	1	3
4.	Philosophical Knowledge	1	1	1	3
5.	Meaning of Philosophy	1	1	1	3
6.	Philosophical Exponents	1	1	2	4
7.	Philosophical Proponents	1	1	1	3
Total		10	12	8	30

From the table above, it was indicated that there were seven (7) objective matter areas of instruction with philosophical school of thought attracting the highest number of items (11) at the three level of instructional objectives, for the objectives, understanding the instructional objectives attracted the highest number of items at all areas of instruction, followed by understanding and lastly thinking.

Table 3: Comparative Psychometric Properties (Difficulty and Discrimination) of C-Tested Items using Classical Test Theory (CTT) and Item Response theory (IRT)

C-Tested Items							
Item	**CTT**			**IRT**			**Inference**
	Difficulty	**Discrim.**	**Comment**	**Difficulty**	**Discrim.**	**Comment**	
1.	0.74	0.32	NA	-1.51	0.80	A	Contrast
2.	0.74	0.29	NA	-2.02	0.56	A	Contrast
3.	0.83	0.29	NA	-1.68	1.16	NA	Harmony
4.	0.60	0.35	A	-0.84	0.49	A	Harmony
5.	0.95	0.04	NA	-10.52	0.28	NA	Harmony
6.	0.66	0.33	A	-1.30	0.54	A	Harmony
7.	0.72	0.38	NA	-1.34	0.80	A	Contrast
8.	0.73	0.36	NA	-1.37	0.85	A	Contrast
9.	0.28	0.09	NA	917.23	0.00	NA	Harmony
10.	0.42	0.07	NA	5.49	0.06	NA	Harmony
11.	0.35	0.18	NA	4.86	0.13	NA	Harmony
12.	0.58	0.33	A	-0.88	0.40	A	Harmony
13.	0.51	0.39	A	-0.06	0.63	A	Harmony
14.	0.47	0.19	NA	3.77	0.03	NA	Harmony
15.	0.68	0.13	NA	10.25	-0.07	NA	Harmony
16.	0.67	0.14	NA	-372.57	0.00	NA	Harmony
17.	0.73	0.43	NA	-1.17	1.04	NA	Harmony
18.	0.30	0.23	NA	1.84	0.50	A	Contrast
19.	0.68	0.36	A	-1.54	0.50	A	Harmony
20.	0.81	0.33	NA	-1.50	1.25	NA	Harmony
21.	0.77	0.41	NA	-1.22	1.25	NA	Harmony
22.	0.52	0.46	NA	-0.12	0.85	A	Contrast
23.	0.53	0.36	A	0.20	0.56	NA	Contrast
24.	0.67	0.43	A	-0.81	1.08	NA	Contrast
25.	0.66	0.46	A	-0.76	1.15	NA	Contrast
26.	0.65	0.51	A	-0.64	1.26	NA	Contrast
27.	0.67	0.48	A	-0.79	1.19	NA	Contrast
28.	0.62	0.55	A	-0.51	1.32	NA	Contrast
29.	0.76	0.45	NA	-1.04	1.59	NA	Harmony
30.	0.56	0.50	A	-0.29	1.23	NA	Contrast

A = Acceptable items, NA = Unacceptable items

The comparison of the psychometric indices for confessional based testing on table 3 above indicated that the analysis of the difficulty index of the items under the two theories (CTT and IRT) revealed that items 4, 6, 12, 13, 18, 19, 22, 23, 24, 25, 26, 27, 28, and 30 were unanimously agreed by the theories to be moderately difficult, CTT identified item 9 to be too difficult while IRT identified items 5 and 9 to be too difficult.

Table 4: Comparative Psychometric Properties (Difficulty and Discrimination Indexes) Table of E-Tested Items using Classical Test Theory (CTT) and Item Response Theory (IRT)

E-Tested Items

Item		**CTT**			**IRT**		**Inference**
	Difficulty	**Discrim.**	**Comment**	**Difficulty**	**Discrim.**	**Comment**	
1.	0.66	0.23	A	-2.08	0.32	A	Contrast
2.	0.64	0.20	NA	-2.68	0.21	A	Contrast
3.	0.66	0.36	A	-1.17	0.59	A	Harmony
4.	0.67	0.38	A	-1.09	0.71	A	Harmony
5.	0.74	0.32	NA	-1.45	0.81	A	Contrast
6.	0.67	0.38	A	-1.06	0.76	A	Harmony
7.	0.69	0.36	A	-1.18	0.79	A	Harmony
8.	0.72	0.45	NA	-0.99	1.21	NA	Harmony
9.	0.61	0.41	A	-0.62	0.78	A	Harmony
10.	0.57	0.34	A	-0.50	0.57	A	Harmony
11.	0.56	0.20	NA	-1.23	0.19	NA	Harmony
12.	0.58	0.38	A	-0.57	0.64	A	Harmony
1 3.	0.62	0.44	A	-0.70	0.82	A	Harmony
14.	0.68	0.38	A	-0.99	0.80	A	Harmony
15.	0.65	0.28	NA	-1.46	0.45	A	Contrast
16.	0.70	0.32	A	-1.49	0.61	A	Harmony
17.	0.66	0.44	A	-0.79	1.02	NA	Contrast
18.	0.70	0.37	A	-1.20	0.79	A	Harmony
19.	0.70	0.35	A	-1.09	0.90	NA	Contrast
20.	0.70	0.48	A	-0.82	1.48	NA	Contrast
21.	0.69	0.44	A	-1.01	0.92	NA	Contrast
22.	0.66	0.43	A	-0.79	1.07	NA	Contrast
23.	0.73	0.28	NA	-1.57	0.70	A	Contrast
24.	0.67	0.41	A	-1.03	0.81	NA	Contrast
25.	0.67	0.38	A	-1.06	0.78	A	Harmony
26.	0.70	0.13	NA	-4.91	0.17	NA	Harmony
27.	0.70	0.43	A	-1.10	0.84	A	Harmony
28.	0.66	0.34	A	-1.33	0.52	A	Harmony
29.	0.64	0.29	NA	-1.34	0.46	A	Contrast
30.	0.64	0.19	A	-5.09	0.11	NA	Harmony

A = Acceptable items, NA = Unacceptable items

From the table above, the difficulty indices derived for each of the items revealed that 25 items (1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 27, 28, and 29) were agreed by the two theories to be good with moderate difficulty strength. Items 5, 8, and 23 were found to be too difficult by CTT while items 26 and 30 were identified to be too easy by IRT. Therefore, 5, 8, 23, 26, and 30 needs to be revisited or discarded.

DISCUSSION

Based on the findings of the study, items tested using the two testing modes were reliable and relatively equivalent with a difference .03, this showed that the differences in the reliability of the two testing modes is not significant at .03, while table 2 confirmed the content validity of the test items used for both conventional and electronics testing modes. This is in line with the assertion of Kibler (1999) that, "the table of specification is used to ensure that the subject matter content and the course objectives are adequately sampled by the test items". This then confirmed the content validity of the items.as same set of items were used for students' assessment on the two modes of assessment. Furthermore, the findings of this study showed the comparability of the psychometric property of the items used for the two test modes. Comparing the psychometric decision of the theories (CTT and IRT) on the items, it was discovered that agreeably, sixteen items needs to be revisited or discarded. This finding attested to the assertion of the study conducted by Jennifer and Jerry (2015) on the use of CTT, IRT, and Rash Model theories to evaluate patient-reported outcome measures, which revealed the result of the study to be similar across the three theories. The findings of the study also revealed similar and slight difference in the opinion of the two theories on the discrimination and difficulty strength of the items across test modes. The differences in examinees' performance in the two testing modes is assumed to be as a result of some eternal factors that came into play most especially before and during the assessment processes.

CONCLUSION

In view of the findings, it was concluded that there appeared to be slight difference in the mean score of students in the two modes of assessment with relatively equivalent and contradictory opinions of the two models (CTT and IRT) on the comparative psychometric properties of the items in each of the assessment modes. Likewise, there was a significant correlation between the two modes of testing at 0.05 level of significance.

RECOMMENDATION

Based on the findings and conclusion of the study, the following recommendations were made:

1. Test developers should seek to improve the psychometric qualities of assessment items used in determining students' academic achievement at all levels of education especially with the use of table of specification (TOS) for generating dichotomously scored items.
2. Classical Test Theory and Item Response Theory (CTT and IRT) should be employed while determining the psychometric qualities of items to have items that are reliable and valid enough to determine student academic performance.
3. The development and administration of items as well as the scoring of examinees' responses to items in courses where attention are being shifted from conventional (C-Testng) to electronics (E-Testing) should be given proper and detailed attention to ensure the effectiveness of assessment outcomes.

REFERENCES

Carlo, M. (2009). Demonstrating the Difference Between Classical Test Theory and Item Response Theory Using Derived Theory. De La Salle University, Manila. *International Journal of Educational Psychological Assessment, 1*(1).

Eleje, I. Chidiebere, C. A. & Fredrick, E. O. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Result. *European Journal of Educational and Social Sciences, 1*(3), 27-45.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTS. Cited in Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2000.). *Computer-Based Testing: Building the Foundation for Future Assessments*. Lawrence Erlbaum Associates. Mahwah, NJ.

Furr, R. M., & Bacharach, V. R. (2008). Psychometrics: An introduction, Sage publication Inc. Lagos.

Jacqueline, P. L., Mark, J. G., & Stephen, M. H. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Ruler-Space Approach. *Journal of Educational Measurement.*

Jennifer, p. & Chery, D. C. (2015). Using CTT, IRT and Rash Model Theory to Evaluate patient-Reported Outcome Measures. A Comparison of Worked Example. *Value in health bulletin, 8*(1), 11-45.

Kilber, M. E. (1999). Classroom Assessment and Learning. New York: Longman, an in-print of Addison Wesley Longman. Inc.

Ojerinde, D., & Wiegers, J. (2015). Strategic Planning and policy implementation in the introduction of large – scale computer – based test: Perspectives on Cito and JAMB Experiences. 33rd AEAAConference, Accra, Ghana. JAMB 2016 Publication.

Wenjing Y. U. & Noriko I. (2021). Comparison of Test Performance on Paper-Based Testing (PBT) and Computer-Based Testing (CBT) By English Majored Undergraduate Students in China.