IOSR Journal of Research & Method in Education (IOSR-JRME)
e- ISSN: 2320–7388,p-ISSN: 2320–737X Volume 8, Issue 4 Ver. III (Jul. – Aug. 2018), PP 27-32
www.iosrjournals.org

Influence Of Selected Test Biases On Computerized Adaptive Testing On ICT For Senior Secondary School Two Students In Rivers State, Nigeria.

Wokoma, Tamuno-Olo Abbott, PROF. ANDY I. JOE
Department of Educational Psychology, Guidance and Counseling, Faculty of Education, University of Port-Harcourt Rivers State, Nigeria.
Corresponding Author: Wokoma

Abstract: *This study investigated influence of selected test biases on computerized adaptive testing on ICT for senior secondary school two (SS2) students in Rivers State of Nigeria. Three research questions and three hypotheses guided the study. The hypotheses were tested at 0.05 level of significance. This research adopted ex-post facto research design. A sample of 100 SS2 students was drawn from the population (all the SS2 students in Rivers State) through multistage sampling technique from the three senatorial zones of the state. Six instruments were designed by the researchers and used for data collection. These are; content bias-free CAT; content biased CAT; construct bias-free CAT; construct biased CAT; predictive bias-free CAT; and predictive biased CAT. Each test has 100 items in its item pool. Items of these CATs were validated in terms of face, content and construct by test experts. The items were also calibrated using threeparameter logistic model of item response theory in RStudio. All items selected for the CAT had reliability coefficient falling within 0.78 to 0.94. Data were analyzed using mean, standard deviation and paired samples t-test. The results showed that content bias, construct bias, and predictive bias influence computerized adaptive testing. Based on the findings, the researchers recommended that appropriately constructed representative items should be used to design computerized adaptive tests for assessment.*

Keywords: *Biases, Computerized, Adaptive, Testing,Content, Construct, Predictive,*

Date of Subissions: 21-08-2018 Date of acceptante: 04-09-2018

I. Introduction

Tests are very important instruments for measuring attributes. Tests are assessment devices used for various purposes such as for determining if learning has taken place in learners in school setting, ascertaining the presence, strength and degree of ailment in patients in hospitals; predicting the aptitude, attitude and productivity of employees in firms; investigating the presence and quality of trait and personality in individuals; and even determining the probability of certain phenomenal occurrences in both animate and inanimate samples.Confirming these uses of tests, Iweka (2014) noted that tests are instruments that are used to measure as accurately as possible, the trait, characteristic, personality or behavior for which it is designed.

Items in different test (if well designed) uniquely differentiate tests according to their purposes and uses of intention. Designers of such tests rigorously ascertain these items in terms of their reliability, validity, dimensionality, homogeneity and other psychometric properties that conform to their intention of the designed test. The design, arrangement, and method of administering these items classify these tests into the different modes of testing. Typical examples are pen/pencil-on-paper (POP) mode, oral testing mode, Computer based testing (CBT) mode, Computer adaptive testing (CAT) mode, multiple-choice, and so forth. These are all alternatives testing mode. Test experts designing items for testing put into consideration salient factors that align such test items to their intended construct to measure. In supporting this,Strecher, Rahn, Ruby, Alt, Robyn and Ward (1997) stated that alternative assessments range from written essay to hands-on performance tasks to cumulative portfolios of diverse work products. In this light, tests containing biased items, when used for testing, may give different interpretations to performances of test-takers belonging to different sub-groups having the same ability.

Items are sub-units of a testing instrument. They are the individual questions that make-up the test. Their individual characteristics sum up to give the test its characteristics. Items having biased characteristics will also contribute same to its test and this will cause deviation from the test objective and result precision if not eliminated or properly edited. Crane, Belle and Larson (2004) stated that "if tests are free of bias, responses to items will be related only to the level of the underlying trait that the items in the test are trying to measure;

and if item bias is present, responses to items will be related to some other factors as well as the level of the underlying trait that items in the test are also trying to measure".

Items in test can be presented in different formats such as multiple choice, matching, dichotomous, open-ended, essay, and so forth. A format chosen for testing must give equal probability to all examinees during testing. Both item and its format during testing must be reliable, valid, unambiguous, practical, fair, socially sensitive and friendly to all test-takers. Items must measure the same trait in all individuals taking the same test without deviation from its (test) objective. Each item behavior during testing in all test-takers must be the same. Items should not function differently nor drift in its characteristics from one test-taker to another. The mode of item administration adopted should not give unfair advantage to a subgroup over others, but to ensure effective measurement of intended construct in all test-takers (Lord, 1971).

Computerized adaptive test (CAT) mode administers items using a computer system. It is a tailored test in which items administered on test-takers(s) are in regards to the test-takers' ability (Wainer&Dorans, 2000). Items are calibrated and stored in an item bank from where the CAT item selection algorithm picks each item presented to the test-taker(s) according to success on previously presented item (Van der Linden &Veldkamp, 2004). This is the reason why computerized adaptive tests are called tailored test. Although CAT mode is seen as the best for educational testing but the fact remains that no measurement is error-free.

Test biases are systematic errors. These can be in the form of content bias, construct bias, predictive bias, situation bias, method bias, and so forth. Biases in tests misrepresent examinees' ability. This is the case when some examinees do not have equal probability of choosing the right responses to items among their peers of the same ability taking the same test. A test would be biased if its items have biased characteristics. This is because a test characteristic is the summation of its entire items characteristics. Such test(s), if administered and its result used for decision making will be very erroneous. In placement situations, such tests can affect students'standing, and can be disproportionate. For predicting future performance,forecasts and evaluation of test-takers to cope and perform task ahead, could also be misleading.

In psychometrics, bias is a systematic error in testing that wrongly estimates measurement outcome. A test is biased if it systematically underestimates or overestimates the ability of an examinee. This error is more conspicuous between the determined means of affected and unaffected subgroups of examinees. The affected subgroups of examinees are sometimes the minority of a given population. In SIOP (N.D.) standards noted that bias refers to construct-irrelevant sources of variances that result in systematically higher or lower indexes of performances for identifiable subgroups of test takers.

Aguinis and Smith (2007) adopted the consensual operationalization of test bias as differences in regression lines of performances across groups of examinees. Jensen (1980) puts it forward that "tests are in various ways culturally biased as to discriminate unfairly against racial and ethnic minorities or persons of low socioeconomic status. Do these biases also influence computerized adaptive tests? The algorithms of CAT are automated and precise.Test items are selected and presented according to the items properties and an examinee's response to previously answered items during test session.The system scores an examinee at the end based on the examinee's performance. From these activities CAT has become an instrument for quality testing and this is why modern testing organizations are opining its use for assessment. CAT operation is anchored on item response theory (IRT). This is a modern test theory that deviates from the commonly used classical test theory (CTT) that has an error component in its response outcome. IRT is a modern latent trait theory (LTT) that ascertains the true ability of a test taker. It is a paradigm for designing, analyzing and scoring instrument and examinees' performances for measuring traits (Wiki, 2018). Do all these attributes make CATsfree from bias?

Although CAT is more effective for use in determining dispositions of trait levels in individuals when compared to traditional tests, but it also has its falls and imbalances as found by some researchers. CATs have been under study for its merits and disadvantages for the past thirty years. Weiss (2004) wrote that a CAT terminated on fixed number of items or by imposing a time limit will cause an operational issue. He further stated that if the CAT's termination criterion is a specified minimum standard error of measurement (SEM), CAT ofsuchterminated will not give resultsofequiprecised measurement.

Linacre (2000) stated that CATs constrain test-takers compared to paper and pencil test. Items on computer screens are found to take longer time to read than printed items. It is also more difficult to identify mistakes on computer screens (Bugbee&Bernt, 1990). Tian, Miao,Zhu, and Gong (2007) found that hardware limitations restrict the types of items that can be admitted by a CAT. Items involving extensive passage, thorough and particularized art work are almost impossible to present in CAT. Stone and Davey (2011) from their review put forward that adaptive tests are more efficient than linear tests, requiring fewer items to be administered to reach a particular measurement precision but in turn decrease its face validity. From this background, the researchers considered it necessary to investigate the influence of some selected test biases on CAT.

The aim of this study was to ascertain the extent to which some test biases (content, construct and predictive bias) can influence computerized adaptive testing. Based on the aim of this study, the researchers formulated the following research questions

1. To what extent does the mean of content bias-free test scores differ from the mean of content biased test scores of CAT on ICT?
2. To what extent does the mean of construct bias-free test scores differ from the mean of construct biased test scores of CAT on ICT?
3. To what extent does the mean of predictive bias-free test scores differ from the mean of predictive biased test scores of CAT on ICT?

Three hypotheses were also formulated to guide this study and were tested at 0.05 alpha level.

1. There is no significant difference between the means of content bias-free test scores and content biased test scores of CAT on ICT.
2. There is no significant difference between the means of construct bias-free test scores and constructbiased test scores of CAT on ICT.
3. There is no significant difference between the means of predictive bias-free test scores and predictive biased test scores of CAT on ICT

II. Methods

The design adopted for this study is ex-post facto research design. Karlingar (1970) defined ex-post facto research as the design in which the independent variable or variables have already occurred and which the researcher starts with the observation of a dependent variable or variables. The population of the study comprised of all senior secondary schools (SS2) students of 2017/2018 session in public (government owned) schools in Rivers State of Nigeria. A sample of 100 SS2 students was drawn from the population of 48, 753 (Rivers State source senior Secondary Schools Board) through multistage sampling technique. Six computerized adaptive tests on ICT designed on Concerto platform were used for the study; namely content bias-free CAT, Content-biased CAT, construct bias-free CAT, constructbiased CAT, predictivebias-free CAT, and predictivebiased CAT. The bias-free CATs do not have biased items while the content-, construct-, and predictive-biased CATs have 50% biased ICT items. Each version of the instrument consists of 100 items in its item pool. These instruments were designed by the researchers. Items of the CATs were written in conformity with SS2 school curriculum on ICT. Face and content validity were ascertained by psychometricians. Each items factor loading were determined by the researchers and the appropriate items were selected. Items for each pool were also calibrated using RStudio in IRT three parameter logistic model (3PLM) before including it in the item pool being developed. All the items for the item pool were trial tested using SS2 students from private schools and the items selected for the different versions were analyzed individual item reliability. Guttman lambda analysis was used in RStudio and all the items finally chosen had reliability coefficients fallingbetween 0.84 and 0.98 respectively. The researchers personally administered the online tests on the subjects and directly collected the data for the study. Data collected were subjected to paired sample t-test analysis for result.

III. Results

The results of the data analysis are presented below.

Research question 1: To what extentdoes the mean of content bias-free test scores differ from the mean of content biased test scores of CAT on ICT? Descriptive (measures of central tendency) statistics was used to answer this research question as shown in table 1.

Table 1:Mean and standard deviation of content bias-free test scores and content biased test scores

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ContentBias-Free TestScores	.1932	100	1.11431	.11143
	ContentBiasedtestScores	-.5725	100	1.16021	.11602

Table 1 shows that 100 students took the two test. The mean of content bias free test scores was 0.1932 and the mean of content biased test scores was-0.05725. Therefore, the researchers concludedthat there is a difference between the means of content biased test scores and content bias-free testscores.

Hypothesis 1: There is no significant difference between the means of content bias-free test scores and content biased test scores of CAT on ICT. Paired samples test was used to test this hypothesis.

Table 2:**Paired samples testanalysis of content bias-free test scores and content biased test scores.**

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	ContentBias-FreeTestScores – ContentBiasedtest Scores	.76574	.85804	.08580	.59549	.93599	8.924	99	.000

In table 2, the statistics hold that there is a significant difference between content bias-free test scores mean and content biased test scores mean. The data from the obtained table values (t = 8.924 at p-value = .000 and 99 at the degrees of freedom) clearly showed this. The mean decrease shows a value of 0.76574 with a 95% confidence interval stretching from a lower bound of 0.59549 to an upper of 0.93599. The hypothesis is therefore rejected.

Research question 2: To what extentdoes the mean of construct bias-free test scores differ from the mean of construct biased test scores of CAT on ICT?The descriptive paired samples statistics was also used to answer research question 2 as shown in table 3.

Table 3:Mean and standard deviation of construct bias-free test scores and construct biased test scores.

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ConstructBias-FreeTest Scores	.2096	100	1.13607	.11361
	ConstructBiasedTestScores	-.4224	100	1.12169	.11217

In table 3, the 100 samples that took the two tests showed mean of 0.2096 for construct bias-free test scores and mean of -0.4224 for construct biased test scores. With this, the conclusion was drawn that there was a decrease (difference) between the means of construct biased test scores and construct bias-free test scores.

Hypothesis 2: There is no significant difference between the means of construct bias-free test scores and construct biased test scores of CAT on ICT. The statistics obtained from the analysis in table 5 that were used to test hypothesis informed the decision of rejecting the hypothesis.

Table 4: **Paired samples test analysis of construct bias-free test scores and construct biased test scores.**

		Paired Differences					t	df	Sig.(2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	ConstructBias-FreeTestScores – ConstructBiasedTestScores	.63200	.81223	.08122	.47084	.79316	7.781	99	.000

In table 4, the t-value obtained is 7.781, at .000 level of significance (2-Tailed) and 99 degrees of freedom. The mean decrease obtained was 0.63200 with a 95% confidence interval boundary of 0.47084 and 0.79316 respectively. This clearly revealed a significant difference between the construct bias-free testscores and construct biased test scores.

Research question 3: To what extent does the mean of predictive bias-free test scores differ from the mean of predictive biased test scores of CAT on ICT? Again the descriptive statistics of paired samples statistics in table 7 were used to answer this research question.

Table5:Mean and standard deviation of predictive bias-free test scores and predictive biased test scores.

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PredictiveBias-FreeTestScores	.1732	100	1.08526	.10853
	PredictiveBiasedTestScores	-.5183	100	1.14960	.11496

In table 5, it shows that the mean score of predictive bias-free test scores was 0.1732 and that of predictive biased test scores was -0.5183 for the 100 samples that took the tests. With these values, the researchers concluded that there is decrease in the predictive bias-free test scores from predictive biased test scores; hence there is a difference.

Hypothesis 3: There is no significant difference between the means of predictive bias-free test scores and predictive biased test scores of CAT on ICT. The outcome of analysis for hypothesis 3 statistics contained in table 8 revealed that there is a significant difference between predictive bias-free test scores and predictive biased test scores of CAT on ICT.

Table 6:Paired samples test analysis of predictive bias-free test scores predictive biased test scores.

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PredictiveBias-FreeTestScores – PredictiveBiased TestScores	.69154	.76764	.07676	.53922	.84386	9.009	99	.000

At 99 degrees of freedom, the obtained t-value is 9.009at .000 alpha level (2-tailed). A mean decrease of 0.69154 was also obtained from the sample test analysis, with boundaries from 0.53922 (as the lower boundary) to 0.84386 (as the upper boundary) within 95% confidence interval. With this, the hypothesis is not retained.

IV. Discussion

Finding from the data analysis to answer research question one and also test of hypothesis one revealed that there is significant difference between the means of content bias-free test scores and content biased test scores, with the mean score of content biased test scores (-0.5752) being lower than the that of content bias-free test scores mean (0.1932). Hypothesis one is not retained because the t-value was significant at a probability value of less than 0.05 at which it was tested.Hence, content bias influence computerized adaptive testing. There is no literature similar to this investigation on CAT but from traditional testing, Wechsler used his abbreviated scale of intelligence scales to determine verbal, performance, and full scale intelligence quotient (IQ) scores of examinee between 6 – 89 years. The test was found to underestimate IQ scores in the minorities' sub-group who are not acculturated together with the majorities' sub-group. The test content was more familiar to the majority groups than the minorities. (Wechsler, 2008).

The result of the study also reveals that construct biased test influence computerized adaptive testing on ICT. The construct biased test scores mean from the respondents is lower (-0.4224) than the construct bias-free test scores mean. The null hypothesis is also not retained because they obtained t-value of 7.781 had a p-value of 0.000.At this P-value the researchers concluded that construct biased test influence computerized adaptive test on ICT. In Thaler, Thames, Cagigas, and Norman (2015),reported a study on Stanford-Binet intelligence scale for measuring traits, having a test structure based on Cattell-Horn-Carrol (CHC) model; having factors of crystallized intelligence, fluid reasoning, visual processing, short term memory and quantitative knowledge; designed for both children and adult test takers. Their result revealed that African American examinees underperformed approximately one standard deviation from white examinees.

The result of this study again showed that predictive-biased test scores mean (-0.5183) is lower than the predictive bias-free test scores mean (0.1732). This decrease in mean answers research question three. The mean of the paired difference in table 8 gave 0.6915 within the bounds of 0.5392 (lower) and 0.8439(upper).The t-value of 9.01 at 99 degrees of freedom showed a p-value of 000. Based on this result the hypothesis is also not retained. From similar work, Naglieri and Ronning (2000) in their study using Naglieri Non-verbal ability test for children between 5-19 years which relied on progressive matrices that produced a single IQ score, was reported to show minimal racial score differences. Further study on the same subject conducted by Lohman (2005) disagreed with this, obtaining a resultthat deviated from this report. Lohman's study revealed that minorities with higher socioeconomic status were preferentially predicted and selected.

V. Conclusion

From the result of the study, the following conclusions were drawn.

1. Content biased test significantly influence computerized adaptive testing on ICT.
2. Construct biased test significantly influence computerized adaptive testing on ICT.

3. Predictive biased test significantly influence computerized adaptive testing on ICT.

Recommendations

The researchers made the following recommendations, which were based on the findings of the study.

1. Computerized adaptive test developers should write items that are representatives of the content domain to be examined. They should ensure that the domain from which items are selected to constitute item pool is well exposed to all groups of examineesduring instruction.
2. CAT developers should carry-out proper item analyses to identify the items with the appropriate construct/factor loading in items. Items without such attribute shouldn't be selected to make-up the test pool.
3. Test developers and administrators should guide against use of any CAT differentially predicting different subgroups of test takers in an examination. Any computerized adaptive test that tends to over predict or under predict an examinee's ability should be edited to predict accurately.

References

[1]. Aguinis, H.&Smith. M.A. (2007), Understanding the impact of test validity and bias onselection errors and adverse impact in human resource selection. Retrieved from https://onlne library.wiley.com>doi>abs. on 23rd July, 2018.

[2]. Bugbee, A.C, &Bernt, F.M. (1990). Testing by Computer: Findings in Six Years of Use 1982-1988.Journal of Research on Computing in Education.23, (1) 87-100, 1990.

[3]. Crane, P. K., Belle G. V.,& Larson E. B. (2004). Test Bias in a Cognitive Test: Differential Item Functioning in the CAST.Statist Med 23(3)241-256 .Retrieved from faculty.washingtonedu>pcrane>mypubson 7th April, 2018.

[4]. Iweka, F.O.E. (2014). Comprehensive Guide to Test Constitution and Administration.Rivers State, Nigeria. ChifasNigeria.

[5]. Jenson, A.R. (1980) Bias in mental testing. Retrieved from emilkirkegaard.dk>wp-content.uploads on 17th June, 2018.

[6]. Kerlinger, F.N. (1986). Foundation of Behavioral Research.New York, Holt, Rinehart &Winston.

[7]. Linacre, J.M. (2000). Computer-Adaptive Testing: A Methodology Whose Time HasCome.Retrieved fromhttp://www.iacat.org>computer-adaptive-teston 8th June, 2018.

[8]. Lohman, D. F. (2005).Review of Naglier& Ford (2003). Does the Naglier Nonverbal Ability TestIdentify Equal Proportions of High-Scoring white, Black, and Hispanic students?Gift Child Quarterly:49(1) 19-28.

[9]. Lord, F.M. (1971).The Self-Scoring Flexilevel Test. Journal of Educational Measurement, 8, 147-151. Retrieve from Onlineibrary.wiley.com>doi>pdf on 21st June, 2018.

[10]. Naligeri, J.A., &Ronning, M.E. (2000). Comparison of white, African American, Hispanic, and Asian children on the Naglier; Nonverbal Ability Test. Psychological assessment, 12(3), 328-334.

[11]. SIOP (n.d.).Fairness and Bias-SIOP.Retrieved fromwww.siop.org>principlesreview >pages.pdf on 12th May, 2018.

[12]. Stone, E. & Davey, T. (2011).Computer-Adaptive Testing for Students with Disabilities.A Review ofthe Literature Research Report ETS-11-32.Retrieved fromhttp://www.ets.org/research/contact.htmlon 18th April, 2018.

[13]. Strecher, B. M., RahN M. L., Ruby A., Alt M., Robyin A.&Ward B (1997). UsingAlternative Assessments in Vocational Education Reports.Retrieved fromhttps://www.rand.org>monograph. on 9th June, 2018.

[14]. Tensen, A.R. (1980). Bias in Mental Testing. The Free Press, New York A Divisionof Macmillan Publishing Co. Inc.

[15]. Thaler, N.S., Thames, A.D, Cagigas, X. & Norman, M.A. (2015). IQ testing and theAfrican American Client.APA PsycNET. Retrieved from Psycnet.apa.ord>2014-54243-005 on 28th July, 2018.

[16]. Tiam, J,Mailo D., Zhu Xia, and Gong J. (2004). An Introduction to the Computerized Adaptive Testing.US-China Educational Review. 4(1) 154-213).

[17]. Van der, Linden W.J.&Veldkamp(2004) Constraining (Item Exposure inComputerized Adaptive Testing with Shadow Tests .Journal of Educational and Behavioral Statistics.29(3). 273- 391.

[18]. Vijver, F. &Tanzer, N.K. (2004) Bias and Equivalence in cross-cultural Assessment: an Overview.Revenue europeenne de psychologieappliquee 5(4) 119-135.

[19]. Wainer, H.,&Dorans H. J. (2000).Computerized Adaptive Testing: A Primer Hillsdale: NJ: Lawrence ErIbaum Associates, Inc.

[20]. Wechsler, D. (2008). Wechsler Adult Intelligence Scale – Fourth Edition. San Antonio, TX Pearson Retrieved from https://www.researchgate.net>publication on 16th July, 2018.

[21]. Weiss, D.J (2004) Computerized Adaptive Testing for Effective and Efficient Measurementin Counseling and Education.Journal of Measurement and Evaluation in counseling and Development, 37(2), 70-84.

[22]. Wikipedia,com (2018). Construct Validity. Retrieved on 14/05/2018from https://en.wikipledia.org/wiki/conruct-validity.

Wokoma "¬¬¬¬¬¬Influence Of Selected Test Biases On Computerized Adaptive Testing On Ict For Senior Secondary School Two Students In Rivers State, Nigeria. "IOSR Journal of Research & Method in Education (IOSR-JRME) , vol. 8, no. 4, 2018, pp. 27-32.