Simulated Evidence of Computer Adaptive Test Length: Implications for High Stakes Assessment in Nigeria

Mayowa O. Ogunjimi
University of Ilorin

Musa A. Ayanwale
University of Johannesburg

Jumoke, I. Oladele
University of Johannesburg

Dorcas S. Daramola
University of Ilorin

Idris M. Jimoh
University of Ilorin

Henry, O. Owolabi
University of Ilorin

Like other African countries, high-stake testing in Nigeria has suffered significant setbacks due to the Covid-19 pandemic. Computerised Adaptive Tests (CAT) is a paradigm shift in the educational assessment that ensures accuracy in ability placements. A survey design was employed to describe the psychometric characteristics of a simulated 3-parameter logistic IRT model designs to support off-site assessments. This simulation protocol involved generating examinee and item pool data, specifying the item selection algorithm and specifying CAT administration rules for execution with SimulCAT. Findings revealed that the fixed-length test guarantees a higher testing precision with an observed systematic error less than zero, a CMAE ranging from 0.2 to 0.3 and RMSE being consistent around 0.2. Findings also revealed that the fixed-length test had a higher item exposure rate which can be handled by falling back on the item selection methods that rely less on the a-parameter. Also, item redundancy was lesser for the fixed-length test compared to the variable-length test. Conclusions are for the fixed-length test option for high-stakes assessment in Nigeria.

Keywords: CAT, IRT, 3PL models, test length, measurement precision, item exposure, simulation

INTRODUCTION

The ongoing Covid-19 pandemic is ravaging the world in unimaginable ways with a death toll of over nine hundred thousand recorded globally (Worldometers.info, 2020). The virus's rapid spread across continents has not left Africa behind, as discovered in all 54 African countries (Xinhu, 2020). Profoundly affected countries include; South Africa, Egypt, Djibouti, Morocco, and Nigeria. Nigeria has a total number of fifty-five thousand, six hundred and thirty-two active cases and over a thousand deaths (Worldometers.info, 2020). With six months of shutting down schools as a way of coping with the pandemic, it is evident that we may have to live with the virus for some time owing to the length of time required to proffer a permanent solution in terms of drugs and vaccines (Lowe, 2020). Teaching and learning have gone virtual worldwide. While this has been easier for developed countries, educational institutions in developing countries are still struggling to keep above the pandemic's high waters. As a case study, in Nigeria, educational activities in the public sector have been on total lockdown with only the private sector operating skeletally. The lockdown has resulted in a distortion of the regular school calendar from basic to tertiary education levels.

Furthermore, students in examination classes are impacted as external examinations conducted by the National Examinations Council (NECO), and West African Examinations Council (WAEC) were postponed due to the Covid-school shutdown. Only recently were students in examination classes called back to school to write their examination even though the world is yet to be rid of the Covid-19 virus that led to the initial school closure. The need for technology-led solutions must be devised for off-site assessments to mitigate current and future occurrences.

LITERATURE REVIEW

Information and Technology (IT) has become a way of life because it shapes the way people work, plays, live, think, and interact with one another. IT has also impacted educational assessment as computers are used for ascertaining the extent to which instructional objectives are achieved. Oladapo (2013) concluded that IT had been the primary building block of the present society quickly. ICT gadgets are available and used in all spheres of life, strengthening the conclusion that the world has reached the digital age (Ando et al. 2016). With over a century of large-scale standardised testing and administration in the United States, standardised testing has developed into a well-established corporate enterprise with giant testing services such as Education Testing Service (ETS), Graduate Record Examination (GRE) and Pearson VUE (Moncaleano & Russell, 2018). Computer Based Test (CBT) makes it possible for responses to be electronically assessed, collated, recorded and reported (Alabi, Issa & Oyekunle, 2012). CBT for educational assessments has transformed testing and made it more technological, owing to a transition from paper-based to computer-based test administration (ETS, 2014). Therefore, CBT is an advancement in testing for transferring a paper-based exam into a computer screen and developing a complete end-to-end assessment service to develop, manage, deliver, and grow the assessment programmes. Pearson VUE (2017) stressed that CBT increases efficiency and reduces overhead costs while offering students an equal chance of success and as such, has become attractive for both school-based and standardised testing. According to Moncaleano and Russell (2018), the shift to digital delivery of educational tests has ignited interest in developing new approaches to collecting evidence of students' learning through embedded assessments.

Alabi, Issa and Oyekunle (2012) stated that CBTs are of two types; linear and adaptive. A linear test is a full-length examination in which the computer selects different questions for individuals without considering their performance level. It consists of a full range of test questions-from the easiest to the most difficult-but not always presented to testees in that order. The linear test is scored in the same way as a paper-based test. On the other hand, CAT is a type of CBT in which the computer selects the range of questions based on the individual's ability level (Alabi, Issa & Oyekunle 2012; Kimura, 2017). Therefore, CAT is a testing procedure that can improve precision for a specified test length or reduced test length with no measurement precision loss (Reckase, 2010; Kimura, 2017). Test items in CAT are taken from a vast

pool of possible questions categorised by content and difficulty. When a paper-based test is taken, students are asked to answer questions ranging from easy to hard. In a computer-based adaptive test, each test-taker receives questions that are at the right level of difficulty for their ability. These tests begin with a question that is of medium level of difficulty for most test-takers. After each question is answered, the computer uses the answer and all previous answers to determine which question will be answered next. The next question is one that best follows the previous performance. As such, adaptive tests select test items based on the candidates' previous response, allow for a more efficient administration mode with fewer items and less testing time, while keeping measurement precision (Martin, 2008; Alabi, Issa & Oyekunle, 2012; Redecker & Johannessen, 2013).

CAT has been an operational option for assessing examinees' ability levels which largely depends on Item Response Theory (IRT). It promotes measurement accuracy, improves the quality of the items and adopts superior item selection procedures. (Reckase, 2010). IRT explains examinees' responses to test items with a mathematical function based on ability (Al-A'ali, 2006); and can be modelled using one (difficulty-*b*), two (discrimination-*a* of an item after item difficulty parameter has been computed) or three (guessing-*c* in addition to *b* and *a*) parameter to explains the level of interaction of the examinees with test items based on the probability of correct response (Baker, 2001; Magno, 2009; Oladele, Ayanwale and Owolabi, 2020). Tian, Miao, Zhu and Gong (2007), noted that deciding when to stop a CAT test is crucial. The stoppage point's importance is reflected in test length because the test is too short, then the ability estimate may be inaccurate.

Conversely, if the test is too long, then time and resources are wasted, and the items could be unnecessarily over-exposed. The examinees also may be tired, and drop in performance level, leading to invalid test results. Chae, Kang, Jeon and Linacre (2000) explained some criteria for termination of CAT included when the item bank is exhausted which generally occurs with small-item banks when every item has been administered to the test-taker, maximum test length is reached when the ability measure is estimated with sufficient precision and when the ability measure is far enough away from the pass-fail criterion or when the test-taker is exhibiting off-test behaviour.

CAT has been widely implemented for a variety of licensing and certification examinations administered to health professionals in the United States of America since the 1990s such as the National Council Licensure Examination for Registered Nurses, the American Society for Clinical Pathology Board of Certification, the National Registry of Emergency Medical Technicians, and the National Association of Boards of Pharmacy (Seo, 2017; Han, 2018a). The Test of English as a Foreign Language (TOEFL) (Mojarrad et al., 2013; Hosseini et al. 2014; Khoshsima & Toroujeni, 2017) and Graduate Record Examinations (GRE) (Rezaie & Golshan, 2015) also began to use the computer-adaptive testing format in the 1990s. Other testing programmes such as the Graduate Management Admission Test (GMAT), Scholastic Aptitude Test (SAT), and Microsoft's qualifications. Also employ the use of adaptive testing for their assessments (Giouroglou & Economides, 2004). CAT has been widely applied in clinical psychology and medicine (Anatchkova, Saris-Baglama, Mark Kosinski & Bjorner, 2009; Kimura, 2017; Seo (2017). The popularity with CAT for high stakes testing is gradually increasing as a result of its advantages including test efficiency leading to greater precision of measurement (Han, 2018b), reliability of assessment and improved selection procedures for large-item banks (El-Alfy & Abdel-Aal, 2008; Han. 2018b), automated scoring for free-text answers as well as written text assignments (Noorbehbahani & Kardan, 2011), transformative testing using complex simulations for sampling students' performances repeatedly over time and integration of assessment with instruction, and the measurement of new skills in more sophisticated ways (Bennett, 2010).

Despite the benefits, CAT requires a large item pool of not less than 1000 (Thompson & Weiss, 2011; Han, 2018a). This requirement has implications on the length of time required for item pool development, expertise and requires resources for implementing CAT. Research continues to provide opportunities to review prior efforts and consider the subsequent development and innovation (Moncaleano & Russell, 2018). Feasibility studies through simulation research for CAT have become necessary. Simulation is one of the recommended procedures for carrying out feasibility, applicability, and planning studies (Thompson & Weiss, 2011). Simulations are carried out under varying conditions for many imaginary examinees to

enable informed decision making. According to van der Linden & Glas (2010), Monte Carlo simulation studies enables the estimation of not only the test length and score precision that CAT would produce, but also to evaluate issues such as item exposure and the size of item bank necessary to produce the desired precision of examinee scores. According to Thompson and Weiss (2011), the two most important dependent variables to consider in Monte Carlo simulations are measurement precision of the test and item exposure. Investigating these variables with implications for off-site testing becomes necessary with schools' shutdown due to the Covid-19 pandemic. Therefore, the purpose of this study is to simulate a three-parameter IRT model for designing a fixed and variable-length CAT programme with implications on measurement precision and item exposure for high-stake testing in Nigeria.

RESEARCH QUESTIONS

1. What is the measurement precision of the fixed-length simulated adaptive test?
2. What is the measurement precision of the variable-length simulated adaptive test?
3. What is the item exposure profile of the fixed-length simulated adaptive test?
4. What is the item exposure profile of the variable-length simulated adaptive test?

METHOD

This simulated study exploring the equal and variable-length test designs implemented following the conventional CAT item selection algorithm's three components are test content balancing, the item selection criterion, and item exposure control; deployed, using SimulCAT (Han, 2012). SimulCAT is deemed appropriate being a specialised Monte-Carlo based simulation software. The simulated study design is explicitly described in Table 1.

TABLE 1
COMPUTERIZED ADAPTIVE TESTING SIMULATION DESIGN USING SIMULCAT

Step	Activity	Activity Description
1 (Simulee and item data)	Simulees	5,000 Simulees with θ–N (mean = 0; SD=1)
	Item pool	300 Items based on a 3-parameter logistic item response theory model a–U (0.5, 1.2); b–U (-3, 3); c– (0,0)
2 (Item Selection)	Item selection criterion	Maximum Fisher information
	Item exposure control	Randomesque (randomly select an item from among the five best items)
	Test length	Equal length= 30 aTerminate when 30 items are attempted Variable-length • terminate when the standard error of estimation becomes smaller than 0.35
3 (Test Administration)	Score estimation	maximum likelihood estimation with fences (lower and upper fences at −3.5 and 3.5, respectively) The initial score was randomly chosen between −0.5 and 0.5

Table 1 shows the adaptive testing algorithm simulated right/wrong item responses for the 5000 simulees "taking" the adaptive test at time slot 1. The Descriptive statistics for the item parameter estimates for an items pool of 300 for both equally of 30 items and variable lengths are given in Table 2.

TABLE 2
DESCRIPTIVE STATISTICS FOR ITEM POOL, N=300

Parameters	Minimum	Maximum	Mean	Std. Deviation
a	.50	1.20	.8496	.20089
b	-3.00	2.96	-.0128	1.73604
c	0	0	0	0

Table 2 shows the mean of parameters a, b, and c to be 0.85, -0.01 and 0, respectively for both the fixed and variable-length simulated tests. This result could be due to the same examinee data and item characteristics used for the simulation. Therefore, it connotes that using the same data sets yields the same parameters be it fixed or variable length tests. The result also shows that the b-parameter is lesser than the a-parameter while the c-parameters for both fixed and variable length tests was 0.

Answering Research Questions

Research Questions 1: What is the measurement precision for the fixed-length simulated adaptive tests?

Thirty (30) items were stipulated for the fixed-length test. Measurement precision was determined based on the conditional estimated BIAS (-0.00302), Mean Absolute Error (MAE-0.17887) and Root Mean Squared Errors (RMSE-0.22449) statistics as shown in Table 3 and figures 1 to 3 respectively.

TABLE 3
MEASUREMENT PRECISION OF VARIABLE-LENGTH AND FIXED-LENGTH SIMULATED ADAPTIVE TESTS

Theta Area	Number of cases	Test Length	CBIAS	CMAE	CRMES
-3.0	17	30	-0.080	0.273	0.345
-2.5	100	30	-0.029	0.180	0.230
-2.0	229	30	0.001	0.182	0.221
-1.5	482	30	0.016	0.175	0.222
-1.0	739	30	-0.008	0.181	0.225
-0.5	952	30	-0.003	0.180	0.225
0.0	940	30	-0.011	0.174	0.222
0.5	702	30	0.009	0.174	0.219
1.0	489	30	-0.002	0.183	0.226
1.5	215	30	-0.010	0.184	0.232
2.0	85	30	-0.029	0.186	0.233
2.5	34	30	0.054	0.188	0.238
3.0	8	30	-0.124	0.185	0.273

FIGURE 1
CBIAS ACROSS THETA AREA

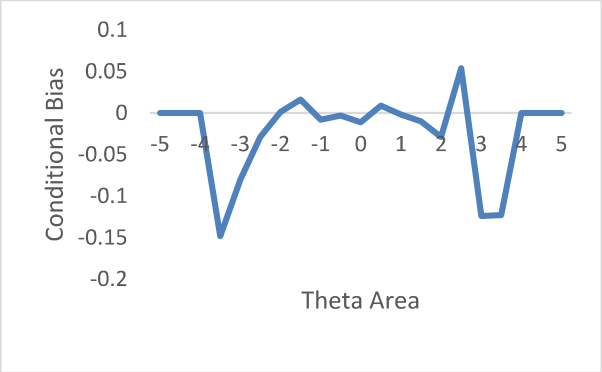


FIGURE 2
CMAE ACROSS THETA AREA

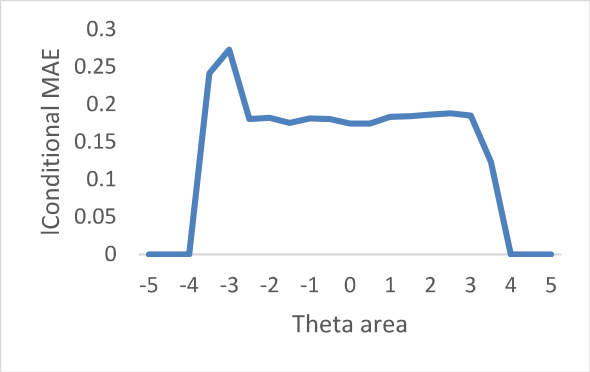
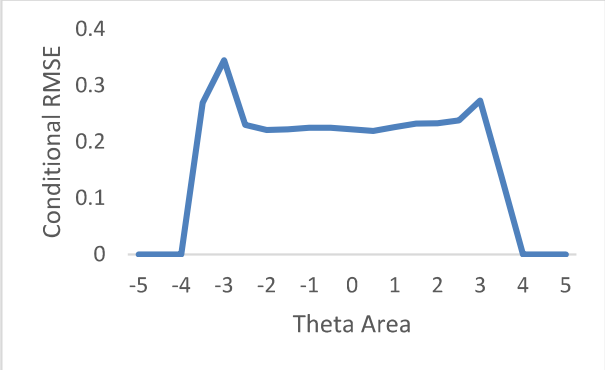


FIGURE 3
CRMSE ACROSS ALL THETA AREA



The simulation results show that the Conditional BIAS (CBIAS) (Fig. 1) indicated that the observed systematic error was less than zero for the fixed test length. With the goal of measurement precision being aimed at attaining zero observed systematic error or as much as possible, the fixed test length could be termed as having an adequate measurement precision. The simulation results also show that the Conditional Mean Absolute Error (CMAE) is a summary of the overall measurement error (both systematic and random error) displayed diagrammatically in Fig. 2 for the fixed test length ranged from 0.2 to 0.3. The lesser the CMAE values, the better the measurement precision of a CAT test. As shown in Figure 3, the RMSE was consistent around 0.2 for fixed test length. The observed consistency with the fixed test length is indicative of its precision.

Research Questions 2: What is the measurement precision for the variable-length simulated adaptive tests?

The simulation results tightly controlled the Conditional Standard Error of Estimation (CSEE) to be lower than 0.35 across all θ areas for the variable-length test. Measurement precision was determined based on the conditional estimated BIAS (0.00216), Mean Absolute Error (MAE-0.28116) and Root Mean Squared Errors (RMSE-0.35489) statistics as shown in Table 4 and used for plotting Figures 4 to 6 respectively.

TABLE 4
MEASUREMENT PRECISION OF VARIABLE-LENGTH SIMULATED ADAPTIVE TESTS

Theta Area	Number of cases	Test Length	CBIAS	CMAE	CRMES
-3.0	17	14.353	-0.007	0.262	0.299
-2.5	100	13.400	0.054	0.298	0.358
-2.0	229	12.856	-0.002	0.265	0.343
-1.5	482	12.886	0.007	0.286	0.361
-1.0	739	12.675	0.008	0.284	0.361
-0.5	952	12.355	0.008	0.291	0.364
0.0	940	11.824	-0.001	0.272	0.344
0.5	702	11.838	-0.019	0.283	0.356
1.0	489	12.376	0.004	0.287	0.363
1.5	215	13.070	-0.009	0.261	0.331
2.0	85	13.576	0.010	0.260	0.325
2.5	34	14.265	-0.030	0.237	0.324
3.0	8	24.500	0.248	0.343	0.400

FIGURE 4
CONDITIONAL BIAS ACROSS THETA AREA

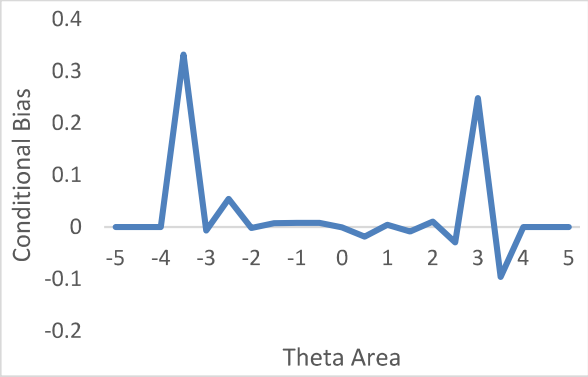


FIGURE 5
CONDITIONAL MAE ACROSS THETA AREA

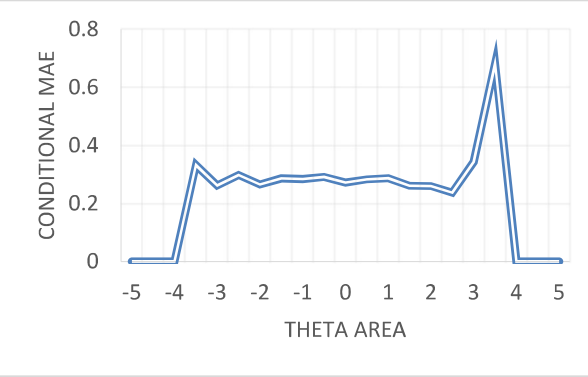
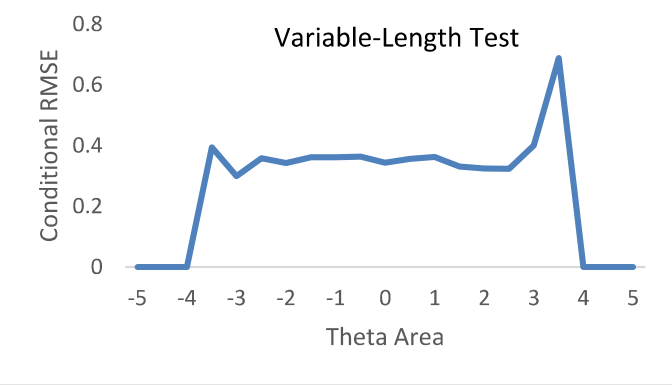


FIGURE 6
CRMSE ACROSS THETA AREA



Conditional BIAS (CBIAS) (Fig. 4) indicated that the observed systematic error was greater than zero for the variable test length. With the goal of measurement precision being aimed at attaining zero observed systematic error or as much as possible, this is an undesirable outcome with variable test length. The simulation results also show that the Conditional Mean Absolute Error (CMAE) is a summary of the overall measurement error (both systematic and randoms error) displayed diagrammatically in Fig. 5 for the variable test length ranged from 0.2 to 0.7; a higher range compared to those obtained with the fixed-length test. As shown in Figure 6, the RMSE was consistent at around 0.3 for the variable-length test. This result shows that the fixed test length has higher precision than the variable test length. The variable-length test is longer for typical examinees but provides only a small increase (and sometimes a decrease) in precision.

Research Question 3: What is the item exposure profile for fixed-length simulated adaptive tests?

The item exposure profile was investigated using the item usage output file from SimulCAT, which revealed that the maximum observed item exposure rate for fixed-length was 3946 (out of 5,000 test administrations/simulees). With more than half of the simulated examinees seeing the item connotes that the item was overexposed. Also, for the fixed length, 115 of 300 items (38.3% of the item pool) were not used. There seems to be an inverse relationship between item exposure and item redundancy. Conclusions would be based on the outcomes with the variable-length test.

Research Question 4: What is the item exposure profile for the variable-length simulated adaptive tests?

The item exposure profile was investigated using the item usage output file from SimulCAT, which revealed that maximum observed item exposure rate for variable-length test was 3707 (out of 5,000 test administrations/simulees). Similar to the fixed-length test, more than half of the simulated examinees seeing the item connotes that the item was overexposed. Also, for the variable test length, 149 of the 300 items (49.7% of the item pool) were not used at all. This result shows that items were better maximised with fixed than variable-length tests. This is not without implication for the item selection method used; in this case, it is the randomesque method and its setting (1 of the best five items) which points), ineffectiveness. According to Han (2018a), changing the item selection method from randomesque to the b-matching method may reduce the item exposure it does not rely on the a-parameter.

DISCUSSIONS AND CONCLUSION

Findings from this simulated study revealed that the fixed-length test guarantees a higher testing precision but with a higher item exposure rate which can be handled by falling back on the item selection methods that rely less on the a-parameter. Also, item redundancy was lesser for the fixed-length test compared to the variable-length test. Conclusions for the option of fixed-length test can therefore be made. Furthermore, test precision aids maximizing score reliability across a wide-ranging score scale which is usually the goal with high stakes testing. As such, the fixed-length test would go a long way to meeting this goal. CAT is usually best suited for large-scale assessments with huge test volumes and continuous or multiple test windows. Because the development of an adaptive-form test involves more cost than a linear fixed-form test, a large population is necessary for a CAT testing program to be financially fruitful.

Another consideration for adopting CAT for high stakes assessment is the cost of technology, access to computing facilities and technological literacy also being a significant factor relevant to test performance. As such, test administrators will need to guard against any potential advantage experienced by tech-literate and tech-advantaged students. Another set of critical issues concerns the need to acknowledge and consider threats to the validity of scores and interpretations that may affect adaptive testing. A common criticism of item-level adaptive testing is that all students are not completing the same items, and even if the same items are answered, they may not be in the same order for all students. Although tailoring is fundamental to adaptive testing and is an essential means by which items presented to students can be reduced (whereas fixed order tests must present all items to cover all levels of ability), the possibility that different item and

order may impact the score for an individual test-taker cannot be discounted. Thus, test administrators are to be appropriately vigilant and nuanced when interpreting the score.

Implications for High-stakes Assessment in Nigeria

Since Nigeria is characterised with a large number of testees with assessments spanning for weeks, the financial requirements of CAT may be overlooked for measurement precision, with a lesser number of items as well as testing time. Accurate ability estimates would be guaranteed, and the length of testing time would be reduced with ripple effects on reduced incurred costs during examinations. This outcome implies that the CAT is appropriate for high stakes testing in Nigeria. Simultaneously, the accuracy of ability estimation can be leveraged on for off-site assessments for mitigating the handicapped situations examination bodies found themselves in the face of the Covid-19 pandemic. This study opens the need for further research on off-site assessment security.

ACKNOWLEDGEMENTS

The authors give due acknowledgement to the following:

1. Indiana University-Purdue University (IUPUI); Indianapolis, Indiana, for providing complimentary registration for the 2020 Assessment Institute Conference held virtually between October 25 to 27; where this paper (Track number 10R); https://assessmentinstitute.iupui.edu/program/2020-important-links.html; and
2. Dr Kyung (Chris) T. Han of the Graduate Management Admission Council authored SimulCAT, a free simulation software package used to generate the data for this study.

REFERENCES

Al-A'ali, M. (2006). IRT-item response theory assessment for an adaptive teaching assessment system. *Proceedings of the 10th WSEAS international conference on applied mathematics* (pp. 518-522). Dallas, Texas, USA. Retrieved from https://www.researchgate.net

Alabi, A.T., Issa, A.O., & Oyekunle, R.A. (2012). The Use of Computer-Based Testing Method for the Conduct of Examinations at the University of Ilorin. *International Journal of Learning & Development*, *2*(3), 68-80. Retrieved from www.macrothink.org/ijld

Anatchkova, M.D., Saris-Baglama, R.N., Mark Kosinski, M.A, & Bjorner, J. (2009). Development and Preliminary Testing of a Computerised Adaptive Assessment of Chronic Pain. *J Pain*, *10*(9), 932–943. https://doi.org/10.1016/j.jpain.2009.03.007

Ando, T., Yamamoto-Hanada, K., Nagao, M., Fujisawa, T., &Ohya, Y. (2016). Combined program with computer-based learning and peer education in early adolescents with asthma: A pilot study. *Journal of Allergy and Clinical Immunology*, *137*(2),18-24. https://doi.org/10.1080/09751122.2017.1346563

Baker, F.B. (2001). The basics of item response theory. *United States of America: ERIC Clearinghouse on Assessment and Evaluation* (2nd Ed.). Retrieved from http://ericae.net/irt

Bennett, R.E. (2010) Technology for large-scale assessment, In P. Peterson, E. Baker & B. Mcgaw (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 8, pp. 48–55). Oxford, Elsevier.

Chae, S., Kang, U., Jeon, E., & Linacre, J.M. (2000). *Development of Computerised Middle School Achievement Test* [in- Korean]. Komesa Press.

Educational Testing Service [ETS]. (2014). A snapshot of the individuals who took the GRE revised general test. Retrieved from https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf

El-Alfy, E.S.M., & Abdel-Aal, R.E. (2008) Construction and analysis of educational tests using abductive machine learning. *Computers and Education*, *51*, 1–16.

Giouroglou, H., & Economides, A. (2004). State-of-the-Art and Adaptive Open-Closed Items in Adaptive Foreign Language Assessment. In *Proceedings 4th Hellenic Conference with International*

Participation: Information and Communication Technologies in Education, A (pp. 747-756). Athens, 27 September - 3 October. New Technologies: Publ. ISBN 960-88359-1-7.
Han, K.C.T. (2018a). Conducting simulation studies for computerised adaptive testing using SimulCAT: an instructional piece. *Journal of Educational Evaluation for Health Professions*, *15*(20), 1-11.
Han, K.C.T. (2018b). Components of the item selection algorithm in computerised adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*(7), 1-13.
Han, K.T. (2012). User's Manual for *SimulCAT*: Windows Software for Simulating Computerized Adaptive Test Administration. Retrieved from https://www.umass.edu/remp/software/simcata/simulcat/
Hosseini, M., Zainol Abidin, M., & Baghdarnia, M. (2014). Computer-based tests (CBT) and paper and pencil tests (PPT) among English Language Learners in Iran. *Procedia-Social and Behavioral Sciences*, *98*, 659 – 667.
Khoshsima, H., & Toroujeni, S.M.H. (2017). Computer Adaptive Testing (CAT) Design; Testing Algorithm and Administration Mode Investigation. *European Journal of Education Studies*, *3*(5), 764-794.
Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professionals*, *14*(12), 1-5.
Lowe, D. (2020). Coronavirus Vaccine Prospects. Retrieved from https://blogs.sciencemag.org/pipeline/archives/2020/04/15/coronavirus-vaccine-prospects
Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11. Retrieved from https://files.eric.ed.gov
Martin, R. (2008). New possibilities and challenges for assessment through the use of technology. In F. Scheuermann & A.G. Pereira (Eds.), *Towards a Research Agenda on Computer-Based Assessment.* Office for Official Publications of the European Communities.
Mojarrad, H., Hemmati, F., JafariGohar, M., & Sadeghi, A. (2013). Computer-based assessment (CBA) Vs. Paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension. *International Journal of Language Learning and Applied Linguistics World*, *4*(4), 418-428.
Moncaleano, S., & Russell, M. (2018). A Historical Analysis of Technological Advances to Educational Testing: A Drive for Efficiency and the Interplay with Validity. *Journal of Applied Testing Technology*, *19*(1), 1-19.
Noorbehbahani, F., & Kardan, A.A. (2011). The automatic assessment of free-text answers using a modified Bleu algorithm. *Computers and Education*, *56*, 337–345.
Obinne, A.D.E. (2012). Using IRT in determining test item prone to guessing. *World Journal of Education*, *2*(1), 91-95. Retrieved from www.sciedu.ca/we
Oladapo, C.O. (2013, August 5-8). Promoting Quality Education in Nigeria: The Role of the Stakeholders. Lead Paper Presented at the *3rd National Conference of the School of Education, Federal College of Education (Technical) Akoka,* Yaba, Lagos.
Oladele, J.I., Ayanwale, M.A., & Owolabi, H.O. (2020). Paradigm Shifts in Computer Adaptive Testing in Nigeria in Terms of Simulated Evidences. *Journal of Social Sciences*, *63*(1-3), 9-20. Publication of Kamla-Raj Enterprises (KRE) Publishers. https://doi.org/10.31901/24566608.2020/63.1-3.2264
Pearson, V.U.E. (2017). A Guide to e-testing Excellence. Retrieved from https://www.pearsonvue.co.uk/Documents/Market-expertise/Africa.aspx
Reckase, M.D. (2010). Designing item pools to optimise the functioning of a computerised adaptive test. *Psychological Test and Assessment Modeling*, *52*(2), 127-141
Redecker, C., & Johannessen, Ø. (2013). Changing Assessment —Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, *48*(1), 79-95.
Rezaie, M., & Golshan, M. (2015). Computer Adaptive Test (CAT): Advantages and Limitations. *International Journal of Educational Investigations*, *2*(5), 128-137. Retrieved from http://www.ijeionline.com/attachments/article/42/IJEI_Vol.2_No.5_2015-5-11.pdf

Seo, D. (2017). Overview and current management of computerised adaptive testing in licensing/ certification examinations. *Journal of Educational Evaluation for Health Professionals*, *14*(17), 1-9. https://doi.org/10.3352/jeehp.2017.14.17

Sherrington, T. (2017). Towards an Assessment Paradigm Shift. Retrieved from https://teacherhead.com/2017/07/16/towards-an-assessment-paradigm-shift/

Study.com. (2020). The importance of assessment in education. Retrieved from https://study.com/academy/lesson/the-importance-of-assessment-in-education.html

Thompson, N.A. (2011). Advantages of Computerized Adaptive Testing (CAT). Assessment Systems (White Paper). Retrieved from https://assess.com/docs/Advantages-of-CAT-Testing.pdf

Thompson, N.A., & Weiss, D.A. (2011). A Framework for the Development of Computerised Adaptive Tests. *Practical Assessment, Research & Evaluation*, *16*(1). Retrieved from http://pareonline.net/getvn.asp?v=16&n=1

Tian, J., Miao, D., Zhu, X., & Gong, J. (2007). An Introduction to the Computerized Adaptive Testing. *US-China Education Review*, *4*(1), 72-81.

Worldometers.info. (2020, September 10). COVID-19 Coronavirus Pandemic. Retrieved from https://www.worldometers.info/coronavirus/

Xinhu. (2020, June 16). Covid-19 UN Chief calls for unity of Security Council. The East African. Retrieved from https://www.theeastafrican.co.ke/scienceandhealth/Africa-virus-cases-pass-240000/3073694-5577250-tqrar/index.html