SIMULATED EVIDENCE OF COMPUTER ADAPTIVE TESTING: IMPLICATIONS FOR HIGHSTAKES ASSESSMENT IN NIGERIA

By

[1]Mayowa O. **OGUNJIMI (Ph.D.),** [2]Musa A. **AYANWALE** (Ph.D.), [3]Jumoke, I. **OLADELE** (Ph.D.), [4]Dorcas S. **DARAMOLA** (Ph.D.), [5]Idris, M. **JIMOH** (Ph.D.) & [6]Prof. Henry, O. **OWOLABI**

[16]Department of Adult and Primary Education, Faculty of Education University of Ilorin, Nigeria
[2]Educational Foundations, Kampala International University, Kampala, Uganda
[345]Department of Social Sciences Education, Faculty of Education, University of Ilorin, Nigeria

A paper for 2020 Assessment Institute Conference- October 25-27, 2020 (Holding Virtually)

Hosted by
Indiana University-Purdue University (IUPUI)
Indianapolis Marriott Downtown
350 West Maryland Street
Indianapolis, Indiana

Presenting/corresponding authors: [2]adekunle.ayanwale@kiu.ac.ug; [3]oladele.ji@unilorin.edu.ng

Abstract

High stakes testing in Nigeria like other African countries has suffered remarkable 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. This study employed a survey design for describing the psychometric characteristics of a simulated 3-parameter logistic IRT model designs to support off-site assessments. This simulation study was designed to include three stages of 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, 3-PL Model, Fixed/Variable Test Length, Item exposure, Simulation

Introduction

- Covid-19 pandemic is ravaging the world in unimaginable ways with a death toll of over nine hundred thousand recorded globally (Worldometers.info, 2020).
- Statistics in Africa, revealed 874,036 cases; 18,498 deaths; 524,557 recoveries
- In Nigeria, 59,287 Infections; 1,113 Mortality (NCDC Nigeria, 2020) leading to total school closure in the public sector with a few private schools struggling with online teaching and learning handling assessment in the most unprofessional manner. Also, standardised examinations (WAEC and NECO) are still largely conducted as paper pencil and which had to be postponed due to the need to adhere to Covid-19 rules.
- These show the that the Covid-19 Pandemic has impacted the educational sector in no small measures which calls for technology led assessment solutions
- Information and Technology (IT) has become a way of life shaping the way people work, play, live, think, and this holds for EDUCATIONAL ASSESSMENTS often implemented as Computer Based Test (CBT).
- CBT makes it possible for responses to be electronically assessed, collated, recorded and reported and no doubts has increase efficiency and reduces overhead costs while offering students an equal chance of success and as such has become attractive for both school-based and standardized testing (Pearson VUE, 2017).
- The shift to digital delivery of educational tests has ignited interest in developing new approaches to collecting evidence of students' learning with efforts geared towards Computer Adaptive Testing where questions are presented to candidates based on their ability level (Kimura, 2017). Some benefits with CAT are ability level items, shorter tests, reduction in test anxiety, test efficiency, and higher measurement precision.
- CAT is premised on Item Response Theory (IRT) which explains examinees' responses to test items via a mathematical function modelled using either 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).
- Hinged on any of these models in designing CAT, the decision as to when to stop a CAT test is crucial (Tian et al., 2007) based on predefined termination criteria which could be when the item bank is exhausted which occurs generally with small-item banks, when every item has been administered to the test-taker with maximum test length is reached when the ability measure is estimated with sufficient precision; when the ability measure is far enough away from the pass-fail criterion or when the test-taker is exhibiting off-test behaviour. Also related is the exposure of items as affected by any of these stopping criterion

Objective of the Study

Feasibility studies through simulation research become necessary for making important decisions for CAT (Thompson & Weiss, 2011).

- The objective of this study is therefore to simulate three-parameter logistics 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.

The study was designed to examine the:

- measurement precision of the fixed-length simulated adaptive test;
- measurement precision of the variable-length simulated adaptive test;
- item exposure profile of the fixed-length simulated adaptive test;
- item exposure profile of the variable-length simulated adaptive test

Method

This simulated study explored the equal and variable-length test designs implemented following the three components of the conventional CAT item selection algorithm which are test content balancing, the item selection criterion, and item exposure control; deployed, using SimulCAT (Han, 2012). SimulCAT is deemed appropriate being a specialized 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 5 best items)
	Test length	Fixed length= 30 (a terminate when 30 items are attempted)
		Variable length (a terminate when 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)
		Initial score randomly chosen between −0.5 and 0.5

Using the parameters specified in Table 1, 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 equal 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

As shown in Table 2, 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 could be as a result of the same examinee data and item characteristics used for the simulation. It therefore connotes that using 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.

Results

Research Question 1: The simulation results showed 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) which is a summary of the overall measurement error (both systematic and non-systematic 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 Question 2: 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) which is a summary of the overall measurement error displayed diagrammatically in Fig. 5 for the variable test length ranged from 0.2 to 0.7. This is a higher range compared to those obtained with fixed-length test. As shown in Figure 6, the RMSE was consistent around 0.3 for the variable-length test. This shows that the fixed test length has a higher precision than the variable test length. For typical examinees, the variable-length test is longer but provides only a small increase (and sometimes a decrease) in precision.

Research Question 3: 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 over exposed. Also, for the fixed length, 115 of 300 items (38.3% of the item pool) were not put to use. There seem 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: 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 over exposed. Also, for the variable test length, 149 of the 300 items (49.7% of the item pool) were not used at all. This shows that items were better maximized with fixed than variable length tests. This is not without implication for the item election method used; in this case, being the randomesque method and its setting (1 of the best 5 items) which points to ineffectiveness. According Han (2018a), changing the item selection method from randomesque to the b-matching method may reduce the item exposure rates since it doesn't rely on the a-parameter.

Discussion and Conclusions

The study reveal that fixed-length test guarantees a higher testing precision but with a higher item exposure rate which can 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.

A consideration for adopting CAT for high stakes assessment is the cost of technology, access to computing facilities and technological literacy also being an important factor relevant to test performance. As such, test administrators will need to guard against any potential advantage experienced by tech-literate students. 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 important 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 ordering may impact the score for an individual test-taker cannot be discounted. Thus, test administrators are to be appropriately vigilant and nuanced when interpreting score.

Conclusions for the fixed-length test can therefore be made which guarantees a higher testing precision. Furthermore, test precision aids maximizing score reliability across a wide-ranging score scale which is usually the goal with high stakes testing. As such, 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.

Implications for High-stakes Assessment in Nigeria

Since Nigeria is characterised with large number of testees with assessments spanning for weeks, the financial requirements of CAT may be overlooked for measurement precision, with lesser number of items as well as testing time. Accurate ability estimates would be guaranteed and length of testing time would be reduced with ripple effects on reduced incurred costs during examinations. This implies that CAT is appropriate for use with high stakes testing in Nigeria while 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.

References

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. www.macrothink.org/ijld

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, Dallas, Texas, USA, 518–522. https://www.researchgate.net

Anatchkova, M. D., Saris-Baglama, R. N., Mark Kosinski, M.A, & Bjorner, J., (2009). Development and Preliminary Testing of a Computerized 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 computerbased 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.). 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 Computerized 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. 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, 747-756. Athens, 27 September – 3 October. New Technologies: Publ. ISBN 960-88359-1-7.

Han, K. T. (2012). User’s Manual for *SimulCAT*: Windows Software for Simulating Computerized Adaptive Test Administration. https://www.umass.edu/remp/software/simcata/simulcat/

Han, K. C. T. (2018a). Conducting simulation studies for computerized 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 computerized adaptive testing. *Journal of Educational Evaluation for Health Professions, 15(7),* 1-13.

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.

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Profesionals*, *14(12*), 1-5.

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.

Lowe, D. (2020).Coronavirus Vaccine Prospects. 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. 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. www.sciedu.ca/we

Oladapo, C. O. (2013). 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,5-8 August, 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/10.31901/24566756.2020/63.1-3.2264

Pearson VUE (2017). A Guide to e-testing Excellence. https://www.pearsonvue.co.uk/Documents/Market- expertise/Africa.aspx

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized 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. http://www.ijeionline.com/attachments/article/42/IJEI_Vol.2_No.5_2015-5-11.pdf

Seo, D. (2017). Overview and current management of computerized 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. https://teacherhead.com/2017/07/16/towards-an-assessment-paradigm-shift/

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

Thompson, Nathan A., & Weiss, David A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation,* 16(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1

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

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

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

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

Figures

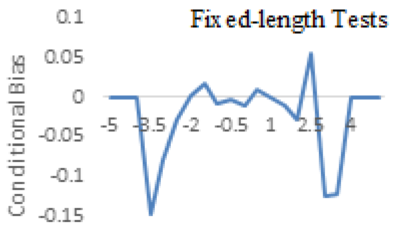

Figure 1: CBIAS across theta area

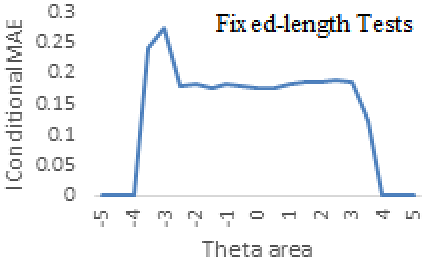

Figure 2: CMAE across theta area

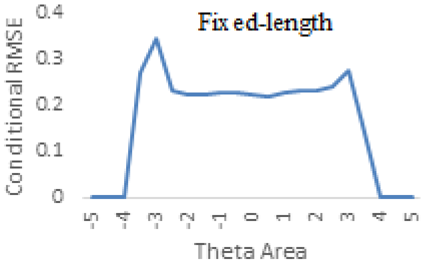

Figure 3: CRMSE across all theta

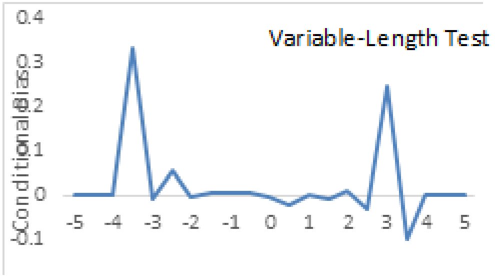

Figure 4: Conditional BIAS across theta area

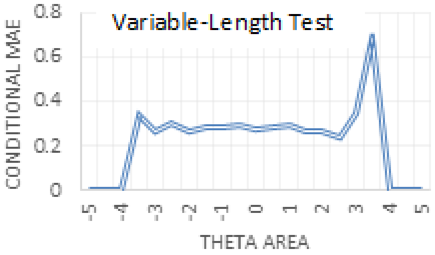

Figure 5: Conditional MAE across theta area

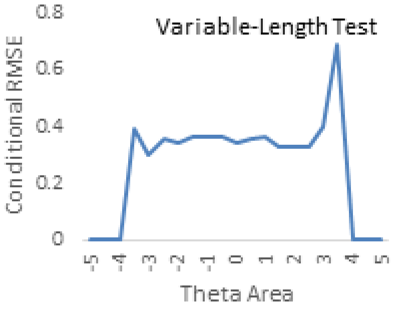

Figure 6: CRMSE across theta area