© JSS 2020
PRINT: ISSN 0970-9274 ONLINE: ISSN 2456-6608
J Soc Sci, 63(1-3): 9-20 (2020)
DOI: 10.31901/24566608.2020/63.1-3.2264

# Paradigm Shifts in Computer Adaptive Testing in Nigeria in Terms of Simulated Evidences

**Jumoke Iyabode Oladele[1], Musa Adekunle Ayanwale[2] and Henry Olumuyiwa Owolabi[3]**

*[1,3]Faculty of Education, University of Ilorin, Ilorin, Nigeria*
*E-mail: [1]<oladele.ji@unilorin.edu.ng>, [3]<henryowo@unilorin.edu.ng*
*[2]Educational Foundations, Kampala International University, Kampala, Uganda*
*E-mail: adekunle.ayanwale@kiu.ac.ug*

**KEYWORDS** Computer Adaptive Testing (CAT). Computer-Based Tests. Educational Assessment. Item Response Theory (IRT). Simulation

**ABSTRACT** Computer and Information Technology have permeated all areas of students' assessment with the aid of Computer-Based Test (CBT). Despite the laudable progress made with CBT, Computerised Adaptive Tests (CAT) is an emerging paradigm in educational assessment with the potentials for greater precision in determining examinees ability level. This study is a simulated CAT assessment with a focus on item selection criteria as a core function. Finding of the study revealed that a-Stratification with b-Blocking item selection method was a preferred method for CAT with a higher SEE, optimal item usage and lesser item exposure rates. Adopting CAT was recommended to guarantee accurate ability placement required for high-stakes testing and leading to improvement in educational assessments. This strengthens the need for high-stakes assessments paradigm shifts from CBT to CAT.

## INTRODUCTION

There is a positive correlation between national and educational development and all educational enterprises are built on outlined objectives. The extent to which the outlined educational objectives are achieved is gauged through educational assessments being an integral part of the educational processes. Technology has impacted educational assessments with the use of computers referred to as Computer Based Testing (CBT). The history of the use of computers in testing dated back to the early 1930s, with IBM model 805 machine used in 1935 as the first attempt to use computers in testing domain. It aimed to score multiple-choice tests of millions of American examinees yearly (Khoshsima and Toroujeni 2017). Moncaleano and Russell (2018) asserted that 2017 marked a century since the first large-scale standardised test was developed and administered in the United States from when standardised testing has developed into a well-established corporate enterprise with giant testing services such as Education Testing Service (ETS), Graduate Record Examinations (GRE), Pearson VUE among others.

CBT is a method of administering tests in which the responses are recorded, assessed, or both electronically (Alabi et al. 2012). With CBT, the instrumentation for testing has become more technological; the most visible technological advancement being the transition from paper-based test administration to computer-based delivery (ETS 2014). Therefore, CBT is advancement in testing not just for transferring a paper-based exam onto a computer screen but encompasses developing a complete end-to-end assessment service to develop, manage, deliver and grow the assessment programme. According to Moncaleano and Russell (2018), the shift to digital delivery of educational tests has ignited interest in developing novel approaches to collecting evidence of student learning through embedded assessments.

Redecker and Johannessen (2013) gave the four generations of computerised educational assessment from linear computerised testing which involved administering conventional tests by computer (*Generation 1*), through computerised adaptive testing which tailor the difficulty or contents or an aspect of the timing based on examinees' responses (*Generation 2*) and continuous measurement (using calibrated measures to continuously and unobtrusively esti-

**Address for correspondence:*
Dr. Jumoke Iyabode Oladele (PhD)
Faculty of Education,
University of Ilorin,
Ilorin, Nigeria
*Telephone:* +234-8060226110
*E-mail:* oladele.ji@unilorin.edu.ng

mate dynamic changes in the student's achievement trajectory *(Generation 3*), to intelligent measurement aimed at producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers through knowledge bases and inference procedures *(Generation 4*). While the first two generations have now become main-stream in developed nations such as America and Europe, with a focus on moving to the third generation, the Sub- Saharan African continent remains in the first generation.

It is noteworthy to say that the current practices in Nigeria showed that high-stake testing such as Joint Admission and Matriculation Board (JAMB) entrance examination and Teachers' Registration Council of Nigeria (TRCN) examination for licensing professionally trained teachers still operating within the confine of linear CBT. The Joint Admission and Matriculation Board (JAMB) is Nigeria's official entrance examination board for candidates seeking admission to the nation's higher institutions. The board was established by Decree No 2 of 1978 (amended by Section 5 of Decree 33 of 1989) with the responsibility of conducting the Universities Matriculation Examination (UME) after which results are sent to higher institutions chosen by the candidates, so that each institution selects and recommends candidates to JAMB for admission (Federal Government of Nigeria 1989). As a way of validating scores obtained by candidates in the UME, Post-UME are conducted by various universities mostly deployed using the linear CBT (Alabi et al. 2012). The linear CBT mode of examination are also notable in some tertiary institutions for taking semester examinations with the University of Ilorin being one of the institutions championing this course (Alabi et al. 2012; Olafare et al. 2017). Similarly, teacher licensure examinations currently use linear CBT in all the accredited centres to administer their examinations twice in a year in all the thirty-one states (TRCN 2014).

The linear CBT adopted for high stakes assessment in Nigeria is one of the two types of CBTs. A linear test is similar to a full-length paper-pencil examination administered through a computer without considering examinees' ability level and scored in the same way as a paper-based test (Georgiadou et al. 2007; Alabi et al. 2012). Therefore, the linear tests are at par with the paper test forms as precisely the same set of test items is administered to all examinees taking a given test form. Both paper and CBT linear forms have a limit number of parallel forms containing no or partially overlapping item sets (Becker and Bergstrom 2013). Psychometric properties of linear CBT are hinged on the Classical Test Theory (CTT) also known as the true score theory. In CTT, test accuracy is expressed as a reliability coefficient, which is used to create confidence intervals around test scores. A confidence interval is defined by the lower and upper limits on a score scale between which it is assumed that the true score of a candidate lies with a certain probability (mostly 90 or 95%). The range of a confidence interval remains the same for every test score on a particular test, suggesting that measurement accuracy is the same for all persons tested.

However, this not always true in real-life situation and perhaps the source of the lacuna between certification and actual performance. For example, a set of very difficult mental arithmetic items administered to an examinee who is fairly able in mental arithmetic will provide more information about the examinee's ability than the same set of items administered to a one who is very poor in mental arithmetic. The reason is that the latter examinee is not likely to respond correctly to any of the test items. Such an examinee's test score will be low and not provide useful information about his or her ability. The only safe conclusion is that the test in question was too hard for him or her. The former examinee will have many of the items right but also get a few items wrong. From this response pattern and test score, one can form a fairly clear picture of the examinee's ability (Straetmans and Eggen 1998).

The aforementioned weaknesses of linear CBT are circumvented by test adaptiveness, also known as Computer Adaptive Testing (CAT). Reckase (2010) refers to CAT as a testing procedure that uses on-the-fly techniques to align to students' ability levels to improve precision while reducing test length, also known as CAT. CAT is one in which the computer selects the range of questions-based student's ability level estimated on their performance on a test (Kimura 2017). Items are taken from a reasonably large pool of possible test items 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 CAT, each examinee receives items at the matching level of difficulty to their ability. CAT begins with an item 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 the next question, which is one that best follows the previous performance. Adaptive tests select test items based on the candidates' last response, allow for a more efficient administration mode while keeping measurement precision (Martin 2008; Redecker and Johannessen 2013). Khoshsima and Toroujeni (2017) also explained that in a CAT program; test items are selected based on the relative ability of the examinee according to their correct or incorrect answers given to the items. Still, they are not precisely targeted to the exact ability estimate.

CAT depends on the quality of the items and item selection procedures; popularly approached with Item Response Theory (IRT). In IRT, the characteristic measured by an item is conceived as an underlying continuum, often referred to as a latent trait. This latent trait is represented 'by a numerical scale, upon which a person's standing can be estimated based on responses to precalibrated test items. Items measuring the trait are seen as being on the same scale. Unlike classical test theory, IRT is an 'itemised' theory (Straetmans and Eggen 1998). With IRT, the focus is not on the test but the individual items. The implication is that the probability that an examinee correctly answers a particular item is specified. An item provides information about the ability of an examinee when the probability of correctly answering it is about fifty percent. In which case, the difficulty level of the item matches the ability level of the examinee. Thus, the amount of information provided by a particular test item depends on the examinee's position on the ability scale, which can give much information about the ability level of a high-ability person. In contrast, it provides little information about the ability level of a low-ability person or just the opposite. Remembering the relationship between item information and measurement accuracy, in IRT, the measurement accuracy of a test varies across ability levels and thus across examinees (Straetmans and Eggen 1998).

IRT explains an examinee's response to test items via a mathematical function based on their ability (Al-A'ali 2006). The theory establishes the level of interaction of the examinees with the items in the test, based on the probability of correct response to an item (Magno 2009). In IRT, the 3-Parameter Logistic (PL) model of item response theory has three parameter estimates which are difficult, discrimination and guessing. The first parameter is used to find out the difficulty level of an item concerning the examinee's ability, which is denoted by 'b' in the IRT equation. Baker (2001) remarks that on the metric scale, the difficulty ranges from -" to +", but the typical range for the difficulty index is between -3 and +3, especially for a non-reference test. Test items with values greater than 3 can be regarded as bad items and extremely difficult. The second parameter in the model is the discrimination index denoted by 'a' which expresses how well an item discriminates (differentiates) from one examinee to another with different ability levels (Obinne 2012; Adedoyin and Mokobi 2013). The typical range of discrimination is between 0 and 2. The higher the discrimination index, the steeper the slope of the item characteristics curve, and the better information provided about the test items. When discrimination becomes very high, it means the items are malfunctioning. The third is the three-parameter logistic model which tells the probability of an individual guessing a multiple-choice item correctly with known examinee ability level, after identifying difficulty and discrimination indices. It is denoted by "c" in the IRT equation. The ideal range for guessing is between 0 and 0.35, which is considered acceptable and if otherwise unacceptable (Baker 2001). Therefore, the chosen range for simulation falls within the cut-off points.

Adaptive testing requires a calibrated item bank. The item bank is a sizeable collection of accessible test items. As explained by Straetmans and Eggen (1998), accessibility means that the items are classified or organised in such a way that they can be retrieved easily for test assembly. Items in an item bank are usually classified according to content, question type, performance type, cross-reference to other items or common stimulus material, author, testing history, and psychometric characteristics, including the difficulty level. A basic CAT procedure is when an examinee is evaluated to have a particular ability. An item of a similar difficulty level is presented. If the examinee succeeds in the item, the ability estimate is raised. If the examinee fails in the item, the ability estimate is reduced. Another item is presented to the examinee based on the revised ability estimate, and the cycle is repeated. Each change in the ability estimate is recorded until the estimate is hardly changing at all to provide the final ability estimate (Tian et al. 2007). Thus, the computer terminates testing when some stopping rule is satisfied (Straetmans and Eggen 1998; Anatchkova et al. 2009). The most well-known algorithm designed to provide an accurate point estimation of individual achievement is achieved through the item selection criterion, a vital function of CAT (Thompson and Weiss 2009; Chang 2015; Han 2018b).

Several item selection criteria have been developed over the years. Kingsbury and Zara (1989) broadly categorised item selection criteria into pre-structured and unstructured methods. Pre-structured methods used in the earliest attempt at adaptive testing were the two-stage, pyramidal, Flexi-level, and stradaptive pre-structured procedures appropriate for the era when computers were considerably slower and scare (Kingsbury and Zara 1989). With the emergence of more powerful, readily available computers and software, most adaptive testing applications have used the unstructured methods for item selection, which are more sophisticated. Item selection is also approached using Bayesian methods. It functions by using prior information about the students' ability level. Some conventional Bayesian methods are maximum posterior weighted information (MPWI), and the minimum expected posterior variance (MEPV) (Murphy et al. 2010; Nandakumar and Viswanandhne 2018). More recent methods centre on item information for Item selection, which requires less computer time (Yao 2019). Some of these methods are maximised Fisher information criterion, the b-matching method, a-stratification method with or without b-blocking, Kullback-Leibler information criterion, the weighted likelihood information criterion, the efficiency balanced information criterion (Barrada et al. 2010; Han 2018b).

A major concern with the choice of item selection criteria with CAT is the measurement precision of ability estimates. Therefore, standard errors of ability score estimates are crucial and efforts should be geared in the direction of improving the accuracy of ability estimates (Matteucci and Veldkamp 2009). According to Han (2012), standard error of estimation (SEE) is used to evaluate item efficiency. It should be noted that when SEE is large particularly at the early stage of CAT administration, an item with a lower a-parameter will result in a larger item efficiency value if all other conditions are the same among items. Items with lower a-parameter tend to show greater efficiency at a wider range of ability levels. With assessment data, SEE is also useful for estimating the accuracy of a prediction that is made for ascertaining ability estimation (Thompson 2018). Another concern stressed by Barrada et al. (2010) on the choice of item selection criterion is on either accuracy or security for a reasonable assessment of CAT efficiency. Therefore, there is a positive relationship between item security and measurement accuracy. Having a for-knowledge of items will naturally result in a correct response. Therefore, the likelihood of answering the item correctly no longer depends on the examinee's trait level, and item parameters and test validity are compromised. Therefore, test security is of utmost importance with CATs for high-stakes testing. A wide range of item selection rules are available with various CAT software. Some of these methods priorities' accuracy, others focus on reducing item overexposure and showing a negative relationship between the two variables. Therefore, the assumption is that this trade-off holds both within and between rules (Chang and Ansley 2003).

Maximum Fisher Information (MFI) has been reported in the literature as the oldest item information related item selection criterion with CAT (van der Linden and Pashley 2000; Murphy et al. 2010; Han 2018b). Sulak and Kelecioðlu (2019) examined CAT item selection methods with regards to ability estimation and test stopping rules. Though their study revealed that, the SE values that were obtained by using the MFI method were found to be higher than that obtained by using the Expected a Posteriori Distribution ability estimation method, MFI methods has been reported to have a greedy tendency for selecting items that display a maximum Test

Information Function (TIF) at particular ability levels and so rarely used in actual operations of CAT applications. According to Han (2018b), the high dependence of MFI item selection on a-parameter creates issues with item pool utilisation with implications on test items security which have necessitated the development of other item selection criteria. One of such methods is *a*-Stratification with *b*-Blocking criterion. This method is an improvement on a-stratification without b-blocking, which addresses the high positive correlation between a-parameter and b-parameter as experienced with proper uses. The method functions by hoarding items with high *a*-parameter values for use at the advanced stages of CAT (Chang 2015). This method yields a stable performance and a striking equilibrium between the measurement efficiency of CAT in relation to an overall item pool usage. Another method is the Match b-Value. This approach to item selection uses the item difficulty match-ing parameter to assess the difference between the interim theta and the b-parameter of all eligible items. It, therefore, also eases the challenges of the MFI criterion, which rely heavily on a-parameter (Han 2018b).

With CATs, item selection cannot be treated in isolation from item usage and exposure rates with implications on item bank security. Barrada et al. (2010) explained that the use of methods that restrict the maximum exposure rate of the items has been the most common solution to this problem, as they are effective in reducing the overlap rate of the rule with a lower item bank security. Improvement in item security is related to measurement accuracy. If an examinee receives an item that is known beforehand, a correct response may be expected. As the probability of a correct answer no longer depends on the examinee’s trait level and item parameters, test validity is compromised. The relevance of test security will vary between CATs as the quality of the assessment is maintained with the controlled revealing of questions. Items getting exposed over a while and security is of significant concern in managing/operating the test. Also, there is a possibility of some questions being presented to students more often than others leading to overexposure of items (Nandakumar and Viswanandhne 2018). As practiced with high-stakes licensure examinations, it is essential to assure the utilisation of many paths through the test to guide against the overuse of critical items and to enhance the security of the testing procedure. One procedure to meet this need is to select an item for administration at random from a group of several items that would provide acceptable measurement (Kingsbury and Zara 1989, as cited in van der Linden 2005).

Advantages of CAT include flexible test management, immediate feedback, and the motivation of examinees, ability level items, reduction in test anxiety, test efficiency and higher precision of measurement. Using CATs can be very beneficial, where many learners should be placed into different classes immediately (Rezaie and Golshan 2015). Other advantages are shorter tests with research showing that CAT can reduce testing time by 50 percent or more with obvious financial benefits. CATs can be designed so that examinees are all measured with the same level of precision, even though they all potentially see unique items. Test precision makes the test fair from a psychometric perspective while still using fewer items. CAT provides an appropriate challenge for each examinee where low examinees are not discouraged or intimidated and high examinees enjoy receiving difficult items. Therefore, students are motivated; there is greater test security which holds the promise of curbing examination malpractices while being open to enjoy the same advantages of linear CBT. For example, tests delivered by the computer, if adaptive, can efficiently use multimedia such as audio and video files (Thompson 2011). These advantages strengthen Reigeluth’s (2012) position that technology plays a crucial role in the success of the post-industrial paradigm of education while enabling a quantum improvement in educational assessments at a lower cost per student per year than in the current industrial-age paradigm.

## Statement of the Problem

Although, CAT has been proved to have a number of attractive advantages, switching to CAT calls for feasibility studies for ascertaining the practicability and applicability of CAT to a testing programme. CAT requires a large pool of items with time, resources and cost implications which may be laborious at the planning stage.

The required item pool can be generated using appropriate software tailored for simulations research for determining the feasibility of CAT. Thompson and Weiss (2011) outline a framework for CAT development with feasibility study being the first of five stages outlined. Feasibility studies are carried out to answer salient questions such as psychometric expertise, item banking capacity, availability of an affordable CAT delivery engine, and translational benefits of reduction in test length among others. These questions are not answered by mere conjecture but through simulation research. Monte Carlo simulation studies allows researchers to estimate not only the test length and score precision that CAT would produce but also to examine issues such as item exposure and the size of item bank necessary to produce the desired precision of examinee scores (van der Linden and Glas 2010, as cited in Thompson and Weiss 2011). Simulation research helps to proffer answers to important questions before the development of an item bank or even a delivery platform before the test development process (Thompson and Weiss 2011). Therefore, feasibility studies through simulation research are important not only from a practical viewpoint but also serve validation purposes. Thompson and Weiss (2011) further stressed that a CAT developed without adequate research and documentation in each of these stages runs the danger of being inefficient at the least and legally indefensible at the worst. For example, arbitrarily setting specifications for a live CAT to start, item selection algorithm, scoring algorithm and termination criterion, without empirical evidence for the choices could cause examinee scores that are not as accurate as claimed, providing some subtraction from the validity of their interpretations. Being the core function of CAT, this study aimed at pre-determining appropriate item selection criteria for high stakes testing using CAT.

### Objective of the Study

The general objective of the study was to provide simulated evidence from three item selection criteria MFI, *a*-Stratification with *b*-Blocking and Matching b-Value while evaluating standard error estimation as well as usage and exposure of the item pool. Stemming from this objective, the following research questions were considered for the study:

1. How does MFI, *a*-Stratification with *b*-Blocking and Matching b-Value item selection criteria impact Standard Error of Estimation?
2. How does MFI, *a*-Stratification with *b*-Blocking and Matching b-Value item selection criteria impact item usage?
3. How does MFI, *a*-Stratification with *b*-Blocking and Matching b-Value item selection criteria impact item exposure?

## MATERIAL AND METHODS

The study adopted the Monte-Carlo simulation method for carrying out CAT feasibility studies. The simulation used a three-parameter logistic model of data generated using SimulCAT; a specialised Monte-Carlo based simulation software (Han 2018a). An item pool with 100 dichotomously scored items was created using the three-parameter logistic (3PL) item response model with item discrimination *(a)*, the difficulty *(b)*, and the guessing *(c)* drawn from a uniform distribution using the following minimum and maximum parameters of *a* (0.5, 1.2), *b* (-3, 3) and *c* (0.15,0.30) respectively. The Descriptive statistics for the item parameter estimate for a pool of 100 items used for the simulated CAT are presented in Table 1.

**Table 1: Descriptive statistics for item pool, n=100**

| *Parameters* | *Minimum* | *Maximum* | *Mean* | *Std. deviation* |
|---|---|---|---|---|
| A | 0.50 | 1.20 | 0.84 | 0.22 |
| B | -2.93 | 2.79 | -0.06 | 1.56 |
| C | 0.15 | 0.30 | 0.22 | 0.04 |

As shown in Table 1, the mean of *a*, *b*, and *c* parameters are 0.84, -0.06 and 0.22, respectively for the fixed-length simulated computer-adaptive test. Parameter a being 0.84 implies the simulated items could discriminate adequately between the low and high-ability students. Also, it provides a reasonable amount of information about the ability of the students on the test items. Also, for b parameter, it implies the simulated test items were not too difficult while c parameter depicts that majority of the test items were within the pseudo guessing. It connotes that the distracters were very plausible and is difficult for students to guess. Choosing these ranges for simulation was with the benchmark for determining items that function maximally (Han 2018a).

Using these benchmarks, the researchers simulated a computer adaptive testing process for three item selection criteria which were Maximum Fisher information (MFI), a-Stratification with b-Blocking and Matching b-Value. The simulation design stipulated a fixed-length test of thirty (30) items. A simulated CAT was specified for 1000 simulees "taking" the adaptive test at time slot 1. Measures used to compare the performance of the item selection criteria were the Standard Error of Estimation, item information function using a-parameter and item usage/exposure patterns. Examinee characteristics were drawn from a normal distribution with a mean of 1 and Standard Deviation of 0. Score estimation was determined using Maximum Likelihood Estimation with Fences (MLEF) with a fixed-length test of 30 items. Also, item exposure controls were set to select an item among the five best randomly. The initial theta value was determined based on actual score data generation for the 1000 examinees. The researchers evaluated the performance of the item selection criteria in relation to the item usage/exposure patterns.

## RESULTS

### How does MFI, a-Stratification with b-Blocking and Matching b-Value impact Standard Error of Estimate?

The Standard Error of Estimation (SEE) using the MFI criterion for 1000 simulees was investigated using the item usage output .sca file from SimulCAT. The Mean SEE and the standard deviation were obtained using MFI, a-Stratification with b-Blocking and Matching b-Value item selection methods (See Table 2).

**Table 2: The mean SEE using MFI, a-Stratification with b-Blocking and Matching b-Value**

| *SEE* | *N* | *Mean* | *Std. deviation* |
|---|---|---|---|
| Maximum Fisher information (MFI) | 1000 | 0.27 | 0.01 |
| *a*-Stratification with *b*-Blocking | 1000 | 0.29 | 0.02 |
| Matching b-Value | 1000 | 0.29 | 0.02 |

As shown in Table 2, the mean SEE using the MFI method was 0.27 while a mean of 0.29 were obtained for both a-Stratification with b-Blocking and b-Matching Value methods. This reveals that the SEE for a-Stratification with b-Blocking and Matching b-Value item selection methods were larger than that of MFI method.

### How does MFI, a-Stratification with b-Blocking and Matching b-Value Impact Item Usage Patterns?

Analysis of item usage using the MFI, a-Stratification with b-Blocking and Matching b-Value item selection criteria was carried out using percentages (See Table 3).

**Table 3: Item usage statistics**

| | *MFI* | *Methods a-Stratification with b-Blocking* | *Matching b-Value* |
|---|---|---|---|
| Total Number of items | 30 | 30 | 30 |
| Percentage of item used | 28 | 30 | 30 |
| Percentage of item used | 93.33% | 100% | 100% |

Table 3 revealed that for MFI method, 93 percent of the items were used leaving 7 percent unused. With a-Stratification with b-Blocking and Matching b-Value item selection criterion, 100 percent were used. This connotes that MFI method did not guarantee an optimal item usage while a-Stratification with b-Blocking and Matching b-Value item selection criteria guaranteed optimal item usage.

To further answer this question, item usage was plotted against a-parameter for each of the three methods (MFI, a-Stratification with b-Blocking and Matching b-Value) as shown in Figures 1, 2 and 3 respectively.

Figure 1 shows a pattern where items with low a-parameters were initially at the commencement of the test, after which items with high a-parameters were used predominantly. This pattern further strengthens the reason for the item redundancy experienced with the MFI method.

Figure 2 shows a pattern where the CAT started with items with high a-parameters, followed by items with low a-parameters, and ended using items in the middle of the continuum. This

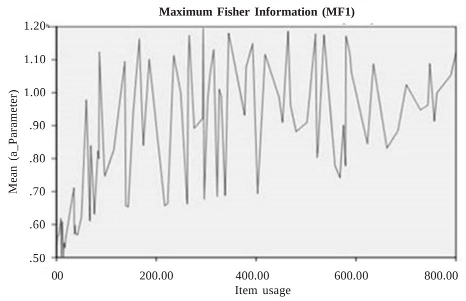

**Fig.1. Item usage pattern using MFI Criterion based on a-parameter**
*Source:* Oladele et al. 2020

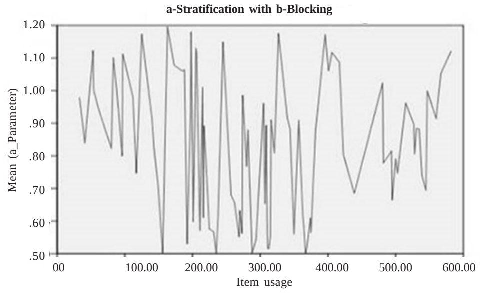

**Fig. 2. Item exposure pattern using *a*-Stratification with the *b*-Blocking criterion based on a-parameter**
*Source:* Oladele et al. 2020

result shows a flexible approach to item selection which lessens the number of unused items while maximising item pool usage.

Figure 3 shows a pattern where the CAT started with items with high a-parameters and continued with items with low, middle and high a-parameters; ended using items with low a-parameters. This pattern shows an unstable approach to item selection which tilts towards low ability examinee while also maximising item pool.

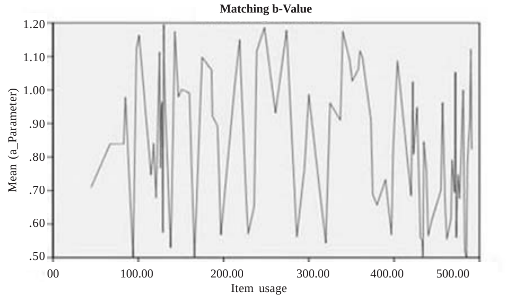

**Fig. 3. Item exposure pattern using matching b-Value criterion based on a-parameters**
*Source:* Oladele et al. 2020

## How does MFI, a-Stratification with b-Blocking and Matching b-Value Impact Item Exposure Patterns?

Analysis of item exposure statistics for the three methods (MFI, a-Stratification with b-Blocking and Matching b-Value) was carried using the Mean and Standard Deviation (See Table 4) and skewness statistics with corresponding Figures 4, 5 and 6 respectively.

**Table 4: Item exposure statistics**

| | *MFI* | *Methods a-Stratification with b-Blocking* | *Matching b-Value* |
|---|---|---|---|
| Mean | 300.00 | 300.00 | 300.00 |
| Std. Deviation | 247.36 | 150.11 | 142.87 |
| Minimum | 0.00 | 33.00 | 45.00 |
| Maximum | 800.00 | 582.00 | 491.00 |

As shown in Table 4, for MFI method, the item exposure mean was 300.00 with a Standard Deviation of 247.36 and a maximum observed item exposure rate was 800 out of 1,000 simulees. This result revealed that more than half of the simulee saw the items which connote that the item was overexposed. This outcome was further verified using the skewness statistics as presented in Figure 4.

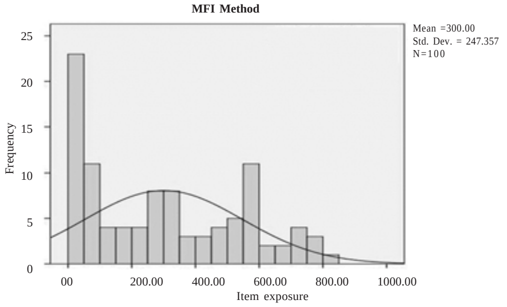

**Fig.4. Skewness for item exposure using MFI Method**
*Source:* Oladele et al. 2020

As shown in Figure 4, a positive skewness was observed for item exposure using the MFI method as the item selection criteria. This outcome points to the fact that using the MFI item selection criteria requires that items should be of top quality for it to function maximally owing to the method's dependence on high b-parameters.

Results in Table 3, reveal that *a*-Stratification with *b*-Blocking method yielded an item exposure means of 300.00 and a Standard Deviation of 150.11. Minimum and maximum observed item exposure rate was 33 and 582 out of 1,000 simulees, respectively. This outcome revealed that more than half of the simulee saw the item which connotes that the item was overexposed. This result was further verified using the skewness statistics (See Fig.5).

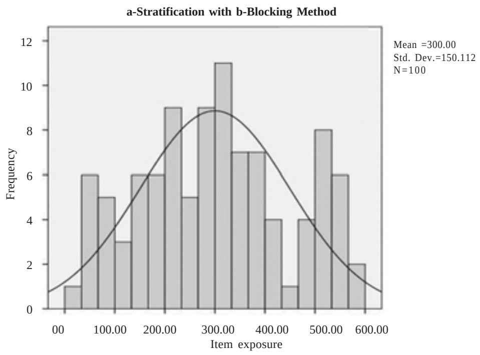

**Fig. 5. Normal curve for item exposure using *a*-Stratification with the *b*-Blocking method**
*Source:* Oladele et al. 2020

As shown in Figure 5, a relatively normal distribution was observed for item exposure using *a*-Stratification with the *b*-Blocking method as the item selection criterion. This outcome connotes that this method is most suitable for candidates' ability placement. Moreover, Table 3 showed the Matching b-Value method, the item exposure mean was 300.00 with a Standard Deviation of 150.11 and a minimum and maximum observed item exposure rate was 15 and 491 (out of 1,000 simulees) respectively. This result revealed that less than half of the simulees saw the item which connotes that the item was not overexposed. This outcome was further verified using the skewness statistics (See Fig. 6).

As shown in Figure 6, a negative skewness was observed for item exposure using the Matching b-Value method as the item selection criteria. This result connotes that this method is most suitable for high ability candidates.

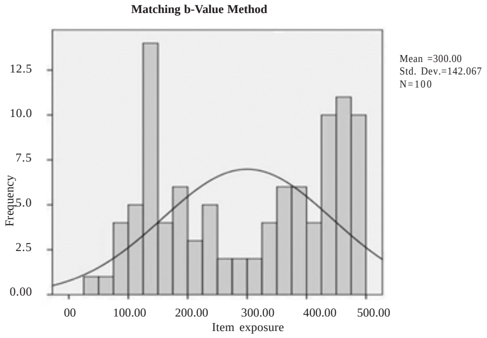

**Fig. 6. Skewness for item exposure using matching b-Value method**
*Source:* Oladele et al. 2020

## DISCUSSION

Results in this study revealed that the Maximum Fisher Information (MFI) criterion yielded the least Standard Error of Estimation (SEE) with a non-optimal item usage and the highest item exposure with more than half of the simulees seeing each of the item. With the goal of measurement precision, attained by minimising measurement error to the barest minimum, the MFI criterion may be seen to be appropriate. However, a significant setback for the MFI criterion was its high dependence on the b-parameter. With an unintended effect on overexposure of items with high a-parameters and non-usage of items, the method does not meet the set standards leading to item pool redundancy with evidence of 6 percent of the items not used. This result is in line with the findings of Han (2018b), which showed that MFI had been reported to have a greedy tendency for selecting items that display a maximum test information function at a particular ability level. The tendency, as mentioned above with MFI criterion foists a risk to test security while creating issues with item pool usage and so rarely used in actual operations of CAT applications.

Similarly, Sulak and Kelecioðlu (2019) reported that Maximum Likelihood Estimation method exhibited higher SE values than were obtained by using the Expected a Posteriori Distribution ability estimation method. Chang (2015) further reiterated that MFI tries to find an item whose difficulty is close to the examinee's estimated proficiency and has a steep item characteristic curve. This poses a limitation to this method in that sharply discriminating items were always chosen first and left many items of appropriate difficulty but lesser discriminating ability only rarely, if ever, used which constitute a waste and while making it easier for candidates to effectively cheat on examinations because of the reduced size of the item pool. Another limitation is that the algorithm performed poorly at the beginning of the exam, with the examinee's proficiency badly estimated.

With a-Stratification with the b-Blocking criterion, findings showed a higher SEE. Items usage was optimal with a 100 percent item usage and lesser item exposure having about half of the simulees seeing the item. Han (2012) stressed that large SEE are desirable particularly at the early stage of CAT administration, as item with a lower a-parameter will result in a larger item efficiency especially when all other conditions are the same among items. Also with a higher SEE, the optimal item usage is a selling point for using *a*-Stratification with *b*-Blocking criterion owing to a balanced pattern where the CAT started with items with high a-parameters after which CAT adjusts itself to simulees' ability level accordingly. This result shows a flexible approach to item selection for accurate ability placement while lessening the number of unused items and maximising item pool usage. This finding is like that of Barrada et al. (2010) who reported the method that yielded an optimal item pool for CAT. It was also reported to outperform other methods, both security and accuracy (Barrada et al. 2006).

Similarly, Nandakumar and Viswanandhne (2018) explained that with the stratification method, the quality of the assessment is maintained with the controlled revealing of questions. This is also in line with Chang (2015) who explained that a-stratified method was proposed for its use of less discriminating items at the beginning of the test and saves highly discriminating items until later stages, when finer gradations of estimation are required. He also stresses that using a-stratified method attempts to match item exposure rates, which has resulted to positive remarks from many researchers on the method. The study by Sulak and Kelecioðlu (2019) revealed a high SE value using a-stratification item selection method with fixed test lengths of 30 items.

Results on the Matching b-Value Criterion also revealed an SEE same as that of a-Stratification with the b-Blocking criterion and also an optimal item usage at 100 percent and the least item exposure having less than half of the simulees seeing the item. However, the item pattern showed a heavy reliance on items with low a-parameters consistently throughout the CAT session which makes the method inappropriate for high ability estimation. This result is a significant setback for this criterion for item selection for CAT. According to Han (2018), this method would be more appropriate with the Rash 1-parameter logistic Model showing the most information when difficulty is matched with appropriate ability level. However, a silver line for this method is that it does not display the greedy item selection pattern with a preference for items with higher a-parameter values as MFI.

## CONCLUSION

Based on the findings of this study, *a*-Stratification with *b*-Blocking is a preferred method for CAT with a flexibility item selection method leading to accurate ability placement. Using this item selection criterion for CAT will enhance placing examinees appropriately on the ability scale while allowing high performers to be distinguished. The major advantage of CAT is of providing more efficient latent trait estimates with 30 items as shown in the simulated study as against 100 items required for the linear computer based testing strengthens the arguments that high-stakes assessments in Nigeria can be efficiently deployed using CAT in terms of ability estimation and cost saving benefits.

## RECOMMENDATIONS

Based on the simulated evidence of CAT being efficient with placing examinees appropriately on the ability spectrum, the authors' recommended that in Nigeria, high-stakes assessments paradigm should be budged from computer-based tests to computer adaptive testing. Policy makers should endeavour to formulate policy statements that would encourage the adoption of CAT for high-stakes assessment for greater reliability of results and satisfying placement of the examinees. Also, assessment experts should embrace CAT and mount rigorous campaign on its usefulness and statistical accuracy within the assessment community. Standardised assessment organizations should organize time to time training for their staff on how to develop CATs.

## ACKNOWLEDGEMENT

The authors give due acknowledgement to Dr. Kyung (Chris) T. Han of the Graduate Management Admission Council who authored SimulCAT, a free software package used to generate the data for this study.

## REFERENCES

Adedoyin OO, Mokobi T 2013. Using IRT psychometric analysis in examining the quality of junior certificate mathematics, multiple-choice examination test items. *International Journal of Asian Social Science*, 3(4): 992-1011.

Alabi AT, Issa AO, Oyekunle RA 2012. The use of computer-based testing method for the conduct of examinations at the University of Ilorin. *International Journal of Learning and Development*, 2(3): 68-80.

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, pp. 518–522.

Anatchkova MD, Saris-Baglama RN, Mark Kosinski MA et al. 2009. Development and preliminary testing of a computerised adaptive assessment of chronic pain. *J Pain,* 10(9): 932–943.

Baker FB 2001. *The Basic of Item Response Theory. Test Calibration*. University of Maryland, College Park, MD: ERIC Clearing House on Assessment and Evaluation.

Barrada JR, Mazuela, P, Olea J 2006. Maximum information stratification method for controlling item exposure in computerised adaptive testing. *Psicothema,* 18: 156-159.

Barrada JR, Olea J, Ponsoda V et al. 2010. A method for the comparison of item selection rules in computerised adaptive testing. *Applied Psychological Measurement,* 34(6): 438-452.

Becker KA, Bergstrom BA 2013. Test administration models. *Practical Assessment, Research, and Evaluation*, 18(1): 14.

Chang SW, Ansley TN 2003. A comparative study of item exposure control methods in computerised adaptive testing. *Journal of Educational Measurement*, 40(1): 71-103.

Chang HH 2015. Psychometrics behind computerised adaptive testing. *Psychometrika,* 80(1): 1-20. https://doi.org/10.1007/s11336-014-9401-5

Georgiadou E, Triantafillou E, Economides AA 2007. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment,* 5(8): n8. https://eric.ed.gov/?id=EJ838610

Educational Testing Service (ETS) 2014. A Snapshot of the Individuals Who Took the GRE Revised General Test. From <https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf> (Retrieved on 12 Frbruary 2020).

Federal Government of Nigeria 1989. Joint Admissions and Matriculation Board Act (Chapter 193). From <http://www.nigerialaw.org/Joint%20 Admissions% 20and%20Matriculation% 20Board% 20Act.m> (Retrieved on 22 March 2020).

Han KT 2012. An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement,* 49(3): 225-246. https://doi.org/10.2307/41653611

Han KCT 2018a. Conducting simulation studies for computerised adaptive testing using SimulCAT: An instructional piece. *Journal of Educational Evaluation for Health Professions,* 15(20). https://doi.org/10.3352/jeehp.2018.15.7

Han KCT 2018b. Components of the item selection algorithm in computerised adaptive testing. *Journal of Educational Evaluation for Health Professions,* 15(7): 1-13. https://doi.org/10.3352/jeehp.2018.15.20

Khoshsima H, Toroujeni SMH 2017. Computer Adaptive Testing (Cat) Design: Testing algorithm and administration mode investigation. *European Journal of Education Studies*, 3(5): 764-794.

Kingsbury GG, Zara AR 1989. Procedures for selecting items for computerised adaptive tests. *Applied Measurement in Education,* 2(4): 359-375. https://doi.org/10.1207/s15324818ame0204_6

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.

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.

Martin R, 2008. New possibilities and challenges for assessment through the use of technology. *Towards a Research Agenda on Computer-based Assessment,* 6-9.

Matteucci M, Veldkamp BP 2009. Computer Adaptive Testing With Empirical Prior Information: A Gibbs Sampler Approach For Ability Estimation. From <http://amsacta.unibo.it/2659/1/matteucci_veldkamp_CAT_2009.pdf> (Retrieved on 28 July 2020).

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 TestingTechnology*, 19(1): 1-19.

Murphy DL, Dodd BG, Vaughn BK 2010. A comparison of item selection techniques for testlets. *Applied Psychological Measurement,* 34(6): 424–437.

Nandakumar GS, Viswanandhne S 2018. A survey on item selection approaches for computer based adaptive testing. *International Journal of Recent Technology and Engineering,* 7(4): 417-419.

Obinne ADE 2012. Using IRT in determining test items prone to guessing. *World Journal of Education,* 2(1): 91-95.

Olafare FO, Akinoso SO, Omotunde C et al. 2017. Students' perceptions of computer-based test in Nigerian universities. *Nigerian Journal of Educational Technology,* 1(2): 117-129.

Reckase MD 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 O 2013. Changing assessment-Towards a new assessment paradigm using ICT. *European Journal of Education*, 48(1): 79-95.

Reigeluth CM 2012. Instructional theory and technology for the new paradigm of education. *RED, Revista de Educación a Distancia,* 32: 1-18

Rezaie M, Golshan M 2015. Computer-Adaptive Test (CAT): Advantages and limitations. *International Journal of Educational Investigations,* 2(5): 128-137.

Straetmans GJJM, Eggen TJHM 1998. Computerised adaptive testing: What it is and how it works. *Educational Technology,* 38(1): 45-52.

Sulak S, Kelecioglu H 2019. Investigation of item selection methods according to test termination rules in CAt applications. *Egitimde ve Psikolojide Ölçme ve Degerlendirme Dergisi,* 10(3): 315-326. https://dergipark.org.tr/en/download/article-file/797961

Teachers' Registration Council of Nigeria 2014. *Professional Qualifying Examination National Benchmarks.* 1st Edition. Abuja, Nigeria: Federal Ministry of Education.

Thompson NA, Weiss DJ 2009. Computerised and adaptive testing in educational assessment. In: F Scheuermann, J Björnsson (Eds.): *The Transition to Computer-Based Assessment.* Italy: Office for Official Publications of the European Communities.

Thompson NA 2011. Advantages of Computerised Adaptive Testing (CAT). Assessment Systems (White Paper). From <https://assess.com/docs/Advantages-of-CAT-Testing.pdf> (Retrieved on 17 September 2019).

Thompson NA 2018. Psychometrics, Test Development: The Story of the Three Standard Errors. Assess Systems Corporation, 5 January 2018. From <https://assess.com/2018/01/05/the-story-of-the-three-standard-errors/> (Retrieved on 28 July 2020).

Tian JQ, Miao DM, Zhu X, Gong JJ 2007. An Introduction to the Computerised Adaptive Testing. Online Submission, 4(1): 72-81. From <https://files.eric.ed.gov/fulltext/ED497385.pdf> (Retrieved on 15 May 2020).

Van der Linden WJ, Pashley P J 2000. Item Selection and Ability Estimation in Adaptive Testing. In: WJ van der Linden, CAW Glas (Eds.): Computerised Adaptive Testing: Theory and Practice. New York, NY: Springer, pp. 1–25. DOI 10.1007/978-0-387-85461-81.

Van Der Linden WJ 2005. A comparison of item selection methods for adaptive tests with content constraints. *Journal of Educational Measurement,* 42(3): 283-302. https://doi.org/10.1111/j.1745-3984.2005.00015.x

Yao L 2019. Item selection methods for computer adaptive testing with passages. *Frontiers in Psychology,* 10: 240. https://doi.org/10.3389/fpsyg.2019.00240.

**Paper received for publication in June, 2020**
**Paper accepted for publication in July, 2020**