Journal of Pedagogical Research
Onlinefirst Article, 2024
https://doi.org/10.33902/JPR.202425210

*Research Article*

# The feasibility of computerized adaptive testing of the national benchmark test: A simulation study

**Musa Adekunle Ayanwale[1] and Mdutshekelwa Ndlovu[2]**

[1]*Faculty of Education, University of Johannesburg, South Africa (ORCID: 0000-0003-4023-8219)*
[2]*Faculty of Education, University of Johannesburg, South Africa (ORCID: 0000-0001-8318-6987)*

The COVID-19 pandemic has had a significant impact on high-stakes testing, including the national benchmark tests in South Africa. Current linear testing formats have been criticized for their limitations, leading to a shift towards Computerized Adaptive Testing [CAT]. Assessments with CAT are more precise and take less time. Evaluation of CAT programs requires simulation studies. To assess the feasibility of implementing CAT in NBTs, SimulCAT, a simulation tool, was utilized. The SimulCAT simulation involved creating 10,000 examinees with a normal distribution characterized by a mean of 0 and a standard deviation of 1. A pool of 500 test items was employed, and specific parameters were established for the item selection algorithm, CAT administration rules, item exposure control, and termination criteria. The termination criteria required a standard error of less than 0.35 to ensure accurate abilities estimation. The findings from the simulation study demonstrated that fixed-length tests provided higher testing precision without any systematic error, as indicated by measurement statistics like CBIAS, CMAE, and CRMSE. However, fixed-length tests exhibited a higher item exposure rate, which could be mitigated by selecting items with fewer dependencies on specific item parameters (a-parameters). On the other hand, variable-length tests demonstrated increased redundancy. Based on these results, CAT is recommended as an alternative approach for conducting NBTs due to its capability to accurately measure individual abilities and reduce the testing duration. For high-stakes assessments like the NBTs, fixed-length tests are preferred as they offer superior testing precision while minimizing item exposure rates.

Keywords: Computerized adaptive testing; Fixed-length; Item exposure; Item response theory; Monte Carlo simulation; Measurement precision; SimulCAT software; Variable-length

Article History: Submitted 5 November 2023; Revised 7 Janurary 2024; Published online 9 March 2024

## 1. Introduction

Many parts of the world, including South Africa, had all physical gatherings, contacts, and educational activities suspended or severely curtailed due to the upsurge of the COVID-19 pandemic worldwide. Consequently, the Centre for Educational Testing for Access and Placement [CETAP] had to scrap the 2020 National Benchmark Tests [NBTs] in South Africa during the lockdown that began on 27th March 2020. NBTs candidates were not allowed to access their examination venues across all provinces in the COVID-19 risk-adjusted strategy. Therefore, it was necessary to shift the assessment paradigm to off-site technologies. In response to these challenges, CETAP announced on 25th July 2020 a transition to a highly secure computer-based assessment

**Address of Corresponding Author**

Musa Adekunle Ayanwale, PhD, Department of Science and Technology Education, Faculty of Education, University of Johannesburg, Johannesburg, Auckland Park, 2006, South Africa.

✉ ayanwalea@uj.ac.za

**How to cite:** Ayanwale, M. A. & Ndlovu, M. (2024). The feasibility of computerized adaptive testing of the national benchmark test: A simulation study. *Journal of Pedagogical Research.* Advance online publication. https://doi.org/10.33902/JPR.202425210

(CETAP, 2020). Evidence showing the shift to computer-based tests came from a recent study by Sango et al. (2022), who demonstrated that NBTs have many restrictions that prevent them from being administered in a paper-based format, as they were before the COVID-19 pandemic. The study also submits that researchers do not anticipate ever returning to exclusively paper-based delivery, even if circumstances return to making paper delivery easier (Sango et al., 2022).

In NBTs, students are tested on their ability to integrate academic and quantitative literacy [AQL] into tertiary courses (Frith & Prince, 2018; Prince et al., 2021; Sebolai, 2014). South African universities administer their NBTs in English and Afrikaans, and these include three multiple-choice tests, including one that combines academic with quantitative literacy. Three hours are allotted to the AQL test, and scored separately (NBT, 2022; Prince et al., 2018). Furthermore, the mathematics [MAT] test is multiple-choice and takes three hours (NBT, 2022). The academic literacy [AL] test measures the ability of a candidate to communicate effectively in a medium of instruction conducive to academic study (Cliff & Yeld, 2006). Quantitative Literacy [QL] tests measure a candidate's ability to solve issues using fundamental quantitative knowledge presented vocally, visually, tabularly, or symbolically in a natural setting relevant to higher education (Frith & Prince, 2006). The NBT Mathematics test assesses a candidate's writing ability in the context of secondary school mathematics ideas relevant to higher education studies (CETAP, 2019).

CETAP's swift action to continue to administer NBTs through technology-led solutions of computer-based tests [CBTs] should be applauded during the pandemic. Assessment of instructional effectiveness has been positively impacted by technology as computers are used to measure whether educational objectives have been met. Technology has been the primary building block of the 21st century (Asiyai, 2014). Technology gadgets are used in all areas of life, proving the rise of the digital age (Ando et al., 2016). It has been more than a century since large-scale standardized testing in the United States was implemented with giant testing services like Education Testing Service, Graduate Record Examination, and Pearson VUE (Moncaleano & Russell, 2018). The Computer Based Test [CBT] allows the outcome/s to be gathered, collated, recorded, and reported electronically (Alabi et al., 2012). Due to the conversion from paper to computer-based testing, CBT for educational assessments has transformed and made this process more technological (Educational Testing Services, 2014). As a result, the CBT can transfer paper-based exams onto a computer screen and provide complete end-to-end assessment services for developing, managing, delivering, and growing programmes (Ogunjimi et al., 2021).

### 1.1. Need to Shift from CBT to CAT Assessment

Researchers are asking themselves - regarding African countries deploying this approach to educational assessment - to what extent they have been successful. Africa has seen an increase in CBT, but it is still not ubiquitous. In South Africa, candidates across provinces are required to take paper-pencil tests; the NBT continues to deploy both online and paper-pencil tests (Sango et al., 2022). In contrast, many other large-scale testing programs in developed countries (Kimura, 2017; Veldkamp & Verschoor, 2019) have adopted computer-based tests. CBTs fall into two categories; one is known as conventional, linear, traditional tests, or computerized fixed-form tests [CFTs], which provide candidates with a pre-determined set of items. Candidates receive test items electronically instead of paper-and-pencil testing [PPT]. The second category is computer-based variable-form testing, which allows administrators to assign items at examination time rather than relying on items that are predetermined. A computer-adaptive test [CAT] (Weiss & Kingsbury, 1984) and linear-on-the-fly testing [LOFT] (Luecht, 2005, 2016; Luecht & Sireci, 2011) are two widely-used variable-form approaches.

Even though CFTs is the least sophisticated of the computer-based tests, it offers several advantages over PPT. Evidently, printing, storing, and distributing booklets, as well as collecting and scanning answer sheets, are no longer required. Consequently, CFT offers the advantage of facilitating the use of item formats not available with PPT, such as multimedia stimuli, and it can record information that is not available in PPT, such as the time it takes for a user to respond to an

item. Additionally, continuous access to tests is possible as opposed to routine administration restricted by logistical issues typically encountered with printed forms (Nandakumar & Viswanandhne, 2018). The item sequences can also be reshuffled, increasing security for each candidate. However, it is worth noting that one of the most recognizable advantages of the CBT is the availability of results immediately following the test (Oladele et al., 2020; Thompson & Weiss, 2009).

On the other hand, the computer-adaptive tests are the one in which the difficulty of subsequent questions changes as a result of how an examinee answers the previous question, as well as tests that use computers to administer and score non-branching linear tests (Scheuermann & Björnsson, 2009). This form of CBT is growing in popularity because it confers several advantages over linear or CFT forms, such as improving measurement efficiency, reducing administration time, and improving a person's estimated accuracy (Kantrowitz et al., 2011, as cited in Tsaousis et al., 2021). The number of items used in CAT systems is reduced by 50% versus linear or CFT evaluation methods (Flens et al., 2016). For Linacre (2000), a CAT system will enhance the validity of the assessment process, and its technique will reduce undesirable assessment phenomena, such as floor and ceiling effects, by reducing problems such as boredom, lack of motivation, attention deficits, and fatigue, which are sources of measurement errors (Seo, 2017; Tsaousis et al., 2021). In high–stakes tests such as NBTs, CAT systems may contribute to increased credibility by reducing measurement error and bias and improving the accuracy of the assessment (Oladele et al., 2023; Mills & Steffen, 2016). For instance, by tailoring tests to candidates' abilities, an irrelevant variance may be reduced. Ludlow and O'Leary (2019) found in PPT, respondents omitting responses for being too difficult resulted in higher measurement error levels. These individuals are called upon to respond only to items relevant to their levels of proficiency, adaptive testing is less likely to skip items. Kantrowitz et al. (2011) submit that the challenge of high-stakes tests, particularly those designed to assess cognitive abilities, is maintaining the quality of security necessary to ensure those test items are protected from being compromised. In numerous studies, it has been shown that two factors could diminish the validity of test scores by affecting the reliability and fairness of trait estimation (Chen & Lei, 2015): item exposure (i.e., how many times this item is used during the assessment process) and item overlap (i.e., how many items are used by more than one examinee at a time). In addition, one of the main components of testing security in the CAT system is item exposure, due to the fact that over time, items are reused, and items with high levels of exposure are more likely to be known than those with low exposure frequency.

Compromising test security poses a serious threat to the validity and fairness of tests, as individuals with prior knowledge may handle items differently (Han, 2018b). Computerized Adaptive Testing systems employ various statistical indices and methods to monitor and control both phenomena, ensuring equitable testing conditions for all candidates (Oladele et al., 2020; Thompson & Weiss, 2011; Tsaousis et al., 2021). Additionally, to guarantee security benefits in CAT, a substantial item bank was initially thought necessary, with a suggested minimum of 1,000 items. However, Monte Carlo simulations conducted by Thompson (2009) and expanded upon by Thompson and Weiss (2011) revealed that a more modest item bank of 500 items is effective. This discovery challenges the initial assumption, emphasizing the practicality of adopting CAT without the requirement for an extensive item bank. Although the development of the item pool may be time-consuming, and CAT implementation may demand more expertise, resources, and research (Moncaleano & Russell, 2018), Thompson and Weiss (2011) recommend simulation research for feasibility, applicability, and planning studies.

### 1.2. Current Study

Frequently used as an evaluative tool, NBTs assess students' academic preparedness and potential to enter higher education. However, due to its traditional fixed-form format, the NBT may have limitations in terms of measurement precision, efficiency, and fairness. To address these concerns,

our study explores the potential implementation of CAT for the NBT. The main objective is to improve the efficiency of the NBT through CAT by selecting test items based on each test-taker's estimated ability level. This involves evaluating the average number of administered items, reducing test duration, and enhancing the overall testing experience. Compared to fixed-form tests, we question whether CAT is better for test-taker abilities. To evaluate this, we conduct a simulation that compares the measurement precision of CAT with the current precision of the NBT. We investigate CAT's ability to provide reliable and valid estimates of test-taker abilities by adapting items at appropriate difficulty levels. In the simulation process, it is essential to examine item difficulty distribution, evaluate item discrimination, analyze fitting statistics, and assess the overall psychometric properties of the NBT's item bank. This critical step is necessary to assess the quality and effectiveness of the item bank within the CAT framework. This research aims to provide insights into the feasibility and potential advantages of implementing CAT for the NBT through a simulation study incorporating an item bank, ability estimation models, item selection algorithms, and evaluation metrics. The expected outcome of this study is that adopting CAT in the NBT may lead to improvements in assessment practices and higher education admissions outcomes.

This paper follows the following structure: In Section 2, we reviewed existing literature on CAT systems and the underpinning theory. In section 3, we presented the methodology, including the context, the design (the Monte-Carlo simulation), and the data analysis method used. The simulation findings are reported in section 4. Section 5 discusses the implications of our findings, including identifying limitations and suggestions for future research. This paper concludes with section 6.

## 2. Literature Review

The item response theory [IRT] framework is currently used by most CAT systems. IRT is a set of algorithms that measure each item's characteristic on a scale, which corresponds to an individual's traits (Ayanwale, 2019; Ayanwale et al., 2022; van der Linden & Glas, 2010). In IRT models such as one-parameter (which looks at the difficulty -*b* of the item), two-parameter (which looks at the discrimination-*a* after item difficulty parameter is computed), three-parameter (which looks at the guessing-*c* in addition to *b* and *a*), and four-parameters (which looks at the carelessness-*d* in addition to *b*, *a* and *c*) the examinee's behaviour is taken into account at the item level (Ayanwale et al., 2018, 2019; Baker, 2004). By modelling at the item level, scores can be reported, and CAT can be developed more efficiently. The item statistics generated by IRT models do not depend with candidate samples and their statistics do not have to do with the items administered (Ayanwale & Adeleke, 2020; Baker, 2001; Baker & Kim, 2017). The assumptions allow test-taker results to be compared even if candidates took different test versions (Zanon et al., 2016). Central assumptions of IRT models are unidimensionality and local independence. An item set and/or an assessment are assumed to be unidimensional when a single latent trait ($\Theta$) is measured, and local independence refers to when pairs of items are taken in a test, there is no statistical relationship between primary trait measured by the test conditioned (Ayanwale et al., 2020; Aybek, 2021; Aybek & Demirtasli, 2017). Both assumptions refer to the same data, while the third assumption involves a model of the relationship between item responses and the trait measured that implies the CAT system.

### 2.1. CAT System Requirements

There are five requirements in every CAT system (Thompson, 2007). These components can be implemented for specific purposes by choosing different options. The first component of the assessment is the item bank, which contains many calibrated items which cover a broad range of difficulties (i.e. $-\infty$ to $+\infty$) regarding the level of difficulty of the ability (or attribute) being measured (Thompson, 2011). The decision rule in the second component determines which item comes first. The starting items are usually based on an ability level $\Theta = 0$ (i.e., the average ability

level) or a random item from the range between - 0.50 and +0.50 if the participant's previous ability level is unknown. Based on the item selection algorithm implemented by the computer system, the system selects items for the third component from a list of items suited for the participant's ability level. Accordingly, the algorithm selects the next item based on the test-takers' correct or incorrect response to the previous item. Maximum Fisher Information [MFI], *a*-stratification (with/without *b*-blocking), the interval information criterion [IIC], Kullback-Leibler information [KLI], Best matching *b*-value, Randomisation, Likelihood Information Criterion, Efficiency Balanced Information, and Gradual Maximum Information Ratio, are methods considered for item selection (Han, 2018a). Ogunjimi et al. (2012) describes the MFI method as the most popular and the most effective criterion for selecting items, whereas the MFI method begins with the item information function. To increase the chances of an item being selected, the information function should be as high as possible (Magis et al., 2017). A simple way to predict ability is by using the MFI criterion, which is very effective and easy to use. It is also noted that the standard error of this criterion is lower than those of any of the other criteria used in the study such as a-stratification, interval information criteria, likelihood weighted information criteria, Kullback–Leibler information, and the gradual information ratio (Deng et al., 2010).

The fourth component describes the approach used to estimate ability levels. Generally, there are three standard methods for ability estimation: Maximum Likelihood Estimation with Fences [MLEF], Maximum Likelihood Estimation [MLE], and Bayesian models (Han, 2016). Maximum Likelihood Estimation with Fences [MLEF] allows for calculating responses based on a log-likelihood function, which considers extreme response patterns that other estimation methods might miss. As a result, MLEF demonstrates minimal non-convergence and provides unbiased estimates regardless of test length (Han, 2018b). Previous studies (Cella et al., 2007; Seo, 2017; Veldkamp & Matteucci, 2013) have noted that applying Bayes' theorem enables the modeling of the conditional probability of item and person parameters given the data. This involves combining prior beliefs with a parametric model based on item and person parameter values. In contrast, Maximum Likelihood Estimation heavily relies on item quality and is evaluated based on its parameters. Van der Linden and Pashley (2009) explained that ML estimation may not yield finite estimates for response patterns with all items correct or all incorrect, which presents challenges, especially in the early stages of CAT administration with short test lengths (Han, 2018a; Oladele et al., 2022). Additionally, Maximum Likelihood methods that treat a person's abilities as fixed effects may lead to undesirable skewness, which can be addressed through bias correction methods (Robitzsch, 2021). To address these challenges, MLEF sets lower and upper bounds for theta estimation, truncating score estimation to fall within those bounds when the log-likelihood function fails to reach its peak with the dichotomous response pattern. Alternatively, Bayesian procedures enhance ability estimation by incorporating prior information on the distribution of the target population, reducing errors in item parameter estimation, especially for the discrimination parameter, particularly with small sample sizes (Cella et al., 2007; Olea et al., 2012).

The fifth and final step involves the termination criterion, determining when the testing process should be concluded. This criterion can be based on standard error (variable length) or a fixed length test (specifying a certain number of items to be attempted). If a fixed length is specified, the testing process stops once that length is reached. Alternatively, to attain the desired level of measurement precision, the test should continue until the 'standard error' is achieved. The simulation study in Computerized Adaptive Testing reveals that users can meet a given criterion with varying numbers of items (Tsaousis et al., 2021; Zhang et al., 2019). Consequently, CAT, lacking thorough feasibility studies through simulation research at each stage of the development process, faces the risk of inefficiency. This renders its potential advantages meaningless and legally indefensible (Thompson & Weiss, 2011).

The Monte Carlo simulation technique is efficient for evaluating a CAT system. The Monte Carlo configuration can simulate item and person parameters, including true item parameters (usually *a*, *b* and *c*), which provide helpful information in evaluating a newly developed

psychometric tool. The simulation studies are beneficial for testing and assessing CAT systems before implementing them on real data. CAT estimates are compared with those from the simulation study. An effective CAT implementation should have a high convergence rate (Aybek & Demirtasli, 2017). Therefore, the main goal of this study was to evaluate the practicality of incorporating Computerized Adaptive Testing for NBTs. The study specifically focused on examining the accuracy of fixed and variable-length tests as well as the impact of item exposure in CAT. The research questions can be summarized as follows: (a) How does the accuracy of fixed-length tests compare to variable-length tests in the context of CAT for NBTs? (b) What effect does item exposure have on the practicality and effectiveness of CAT for NBTs? And (c) How does the implementation of CAT for NBTs contribute to advancements in assessment practices, particularly within the African context? SimulCAT software (Han, 2018b) was utilized to calculate various statistical indices, including the conditional BIAS statistic [CBIAS], the conditional mean absolute error [CMAE] statistic, and the conditional root mean square error [CRMSE] statistic.

### 2.2. Previous Studies

CAT has been extensively researched in various regions, as evidenced by studies such as Aybek (2021), Aybek and Demirtasli (2017), Burhanettin and Selahattin (2022), Choe and Fu (2018), Han (2016), Kantrowitz et al. (2011), Magis et al. (2017), Oladele et al., 2020, 2022; Ogunjimi et al. (2021), Mills and Steffen (2016), Thompson (2009), Thompson and Weiss (2011), Tsaousis et al. (2021), van der Linden and Glas (2010), and Wang and Kingston (2019). These studies collectively reinforce the validity and effectiveness of using CAT in educational and psychological assessments. Also, Thompson (2017) suggest that implementing simulated national CATs in the UK yields significant advantages over fixed tests, particularly in a formative educational setting. Shorter tests tailored to individual learners, with content suitable for their level, enhance learner engagement and provide a better overall learning experience. Additionally, the results of these tests can be processed more quickly, allowing for prompt review and discussion with the learner while their assessment experience is still fresh in their mind. In the realm of high-stakes exams, simulation techniques have been emphasized by scholars like Erdem Kara (2019) and Han and Kosinski (2016) as crucial for the development and evaluation of CATs. These techniques aid in assessing the effectiveness and reliability of CATs in such contexts, ensuring that they meet the required standards.

## 3. Methodology

### 3.1. Study Design

The purpose of this study is to examine the precision of measurement in fixed and variable-length test designs when using the conventional computer adaptive testing item selection algorithm. The algorithm consists of three components: item selection criteria, and exposure controls, and it was developed by Han (2012). Several software packages are available for simulating CAT, including SimulCAT (Han, 2012), CATsim (Assessment Systems Corporation [ASC], n.d.), Firestar (Choi, 2009), and WinGen (Han, 2007). Additionally, simulation can be performed using the catR package in the R programming language (Magis & Barrada, 2017; Magis & Raĭche, 2012), while the mirtCAT package (Chalmers, 2016) is specifically designed for developing live CAT applications. In this study, SimulCAT was selected for the NBTs feasibility analysis due to its suitability, despite not being a commercial CAT software. The primary objective of CAT is to provide the most efficient and informative items for each group of test-takers (Embretson & Reise, 2013). Different items are administered based on the varying proficiency levels of the candidates. Each item is utilized to estimate and update the ability level of the test-takers, and the subsequent item selection is based on this updated level. This process is repeated until specific stopping criteria are met (Erdem Kara, 2019). SimulCAT, being a Monte-Carlo simulation program, is well-suited for this study. Please refer to Table 1 for a detailed description of the simulated study design.

Table 1
*Computerised adaptive testing simulation design Simulees*

| *Step* | *Activity* | *Description of the activity* |
|---|---|---|
| One (Examinee and item data) | Simulees | 10,000 Simulees with Θ ~N (0, 1) |
| | Item pool | 500 items based on a 3-parameter logistic model of IRT<br>a–U (0.5, 1.2)<br>b–U (-3, 3)<br>c–U (0, 0.35) |
| Two (Item Selection) | Item selection criterion | Maximum Fisher Information (MFI) |
| | Item exposure control | Randomesque (randomly select an item from among the five best items) |
| | Test length | |
| | Fixed | Terminate when 50 items are responded to |
| | Variable | • Terminate when the standard error of estimation becomes smaller than 0.35<br>• Maximum of 50 items |
| | Content balancing | None |
| Three (Test Administration) | Score estimation | Maximum likelihood estimation with fences (lower and upper fences at −3.5 and 3.5, respectively<br>The initial score was randomly chosen between −0.5 and 0.5<br>Limit the estimated jump by 1 for the first five items |
| | Outputs | Save item usage (exposure) |

Table 1 simulated the random selection of 10,000 candidates using a normal distribution with different $\theta$ values for each time and 500 simulated item responses. Using a CAT simulation is an effective way to evaluate the performance of a CAT administration given the item pool distribution and examinee distribution.

### 3.2. Data Analysis

The measurement precision of the test program is an important aspect of CAT evaluation. In an evaluation of CAT system precision, several different indices were suggested (Han, 2018b). First, an index measuring CAT measurement precision is the BIAS statistic. Here we show the average difference between estimated and true θ across all candidates. The difference is calculated as follows:

$$\text{Bias} = \frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)}{I} \quad \text{..... Eqn. 1}$$

where $I$ is the number of examinees who have taken the test. A summary statistic that measures the accuracy of the test is provided by this statistic. Additionally, we estimated a conditional BIAS (CBIAS) statistic because the characteristics of the tests may differ considerably across the range of $\theta$ (Han, 2018a). This statistic indicates the bias between the different ability levels (e.g., $\theta$ <-2, -2 ≤ $\theta$ < -1, -1 ≤ $\theta$ < 0, 0 ≤ $\theta$ < 1, 1 ≤ $\theta$ < 2, and $\theta$ ≥ 2), that is, the condition for theta. A second statistic that depicts the overall measurement error is the mean absolute error (MAE). The estimated $\theta$ differs from the true θ for all examinees. Additionally, the conditional MAE (CMAE) was calculated based on the means MAE values at each point in the $\theta$ range. The MAE statistic is computed as:

$$\text{MAE} = \frac{\sum_{i=1}^{I}|\hat{\theta}_i - \theta_i|}{I} \quad \text{..... Eqn. 2}$$

Another useful statistic used in CAT for measurement precision is the root mean square error [RMSE]. Unlike previous ones, this statistic is based on the same ability scale $\theta$, while the conditional RMSE [CRMSE] within each $\theta$ point was also estimated. Thus, the square of the bias and the square root of the result is estimated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)^2}{I}} \quad \text{..... Eqn. 3}$$

The last statistic for measurement precision is the standard error estimate [SEE] associated with a conditional estimate for different $\theta$ values [CSEE]. As shown below, the SEE is estimated as follows:

$$\text{SEE} = \frac{1}{\sqrt{\text{TIF}}} \quad \text{..... Eqn. 4}$$

Where TIF represents the test information function for the specific test form, each examinee completes. More importantly, using SEE statistics, CATs with variable-lengths are frequently terminated (Han, 2018a).

## 4. Results

For a 500-item pool with 50 items of fixed and variable length, the following item parameter estimates are presented in abridged Table 2. Based on the three-parameter logistic IRT framework, Table 2 shows the psychometric quality of simulated items for the fixed and variable lengths. The b values ranged from −2.99 to 2.99, with a mean of 0.11 and a standard deviation of 1.71, indicating the item pool covers a wide range of difficulty levels across the θ continuum (Baker, 2001, 2004). Across items, the discrimination parameters range from 0.50 to 1.19, with a mean of 0.85 and a standard deviation of 0.20, suggesting the items in the item pool adequately distinguished between low and high-ability test-takers (Baker, 2001; Baker & Kim, 2017). For the entire item pool, estimates of the c-parameter ranged from 0.00 to 0.34, with a mean of 0.17 and a standard deviation of 0.09, indicative of random guessing (Baker, 2001).

Table 2
*Item parameters for item pool of fixed and variable length*

| Item pool | *a* | *b* | *c* |
|---|---|---|---|
| 1 | 0.88 | −2.43 | 0.13 |
| 2 | 0.66 | −2.75 | 0.14 |
| 3 | 0.73 | 0.12 | 0.20 |
| 4 | 1.12 | 0.96 | 0.26 |
| 5 | 0.74 | 0.35 | 0.25 |
| 6 | 0.69 | −0.89 | 0.11 |
| 7 | 0.76 | −1.94 | 0.16 |
| 8 | 0.56 | −2.49 | 0.19 |
| 9 | 1.12 | −0.63 | 0.16 |
| 10 | 0.73 | 0.35 | 0.33 |
| + | + | + | + |
| 491 | 1.04 | −0.21 | 0.12 |
| 492 | 0.55 | −1.04 | 0.30 |
| 493 | 0.89 | −1.77 | 0.20 |
| 494 | 0.85 | 1.66 | 0.31 |
| 495 | 0.52 | −2.51 | 0.07 |
| 496 | 0.68 | 2.23 | 0.27 |
| 497 | 0.77 | −1.58 | 0.14 |
| 498 | 1.01 | −0.76 | 0.22 |
| 499 | 0.79 | 1.25 | 0.30 |
| 500 | 1.05 | 1.76 | 0.01 |
| Mean | 0.85 | 0.11 | 0.17 |
| SD | 0.20 | 1.71 | 0.09 |
| Max. | 0.50 | −2.99 | 0.00 |
| Min. | 1.19 | 2.99 | 0.34 |

*Note*: *a*-Discrimination; *b*- Difficulty; *c*- Guessing; sd- Standard deviation; max- Maximum and min- Minimum.

Next is the assessment of measurement precision for simulated adaptive tests that are fixed in length (i.e., 50) across various $\theta$ areas using statistics such as CBIAS, CMAE, and CRMSE. These CAT system's statistics return mean CBIAS, a measurement of how well it recovers the true $\theta$ parameters, was -0.03440, very close to zero. Similar results were obtained for the mean of CMAE with 0.19046 and the CRMSE with 0.22680. Consequently, for different $\theta$ areas of the fixed test length, Table 3 presents the CBIAS, CMAE, and CRMSE scores, respectively.

Simulation statistics for 50 items of fixed length are shown in Table 3. CBIAS across theta continuum showed systematic error was less than zero for the fixed test length, as shown in Figure 1. A fixed test length is deemed adequate measurement precision because it can achieve zero systematic error. Next, two similar measurements precision indices are displayed; Figure 2 displays the CMAE along the theta continuum, summarising the overall measurement error for the fixed test length. While Figure 3 displays the CRMSE along the theta continuum for fixed-length tests, the smaller the CMAE values, the more accurate the CAT system. In addition, a closer look at Figures 2 and 3 shows a similar pattern, corresponding to higher levels of ability with greater rates of error. However, there appear to be a few very difficult items across the range of ability levels, which might prevent the CAT system from providing accurate ability estimates for the test-takers.

Table 3
*Fixed test length statistics for measurement precision*

| Theta Area ($\Theta$) | Number of Cases | Test Length | CBIAS | CMAE | CRMSE |
|---|---|---|---|---|---|
| −3.5 | 9 | 50 | 0.221 | 0.221 | 0.247 |
| −3 | 41 | 50 | 0.019 | 0.156 | 0.191 |
| −2.5 | 165 | 50 | −0.016 | 0.151 | 0.187 |
| −2 | 435 | 50 | 0.002 | 0.159 | 0.201 |
| −1.5 | 872 | 50 | −0.001 | 0.155 | 0.196 |
| −1 | 1524 | 50 | −0.001 | 0.152 | 0.193 |
| −0.5 | 1953 | 50 | −0.002 | 0.157 | 0.199 |
| 0 | 1927 | 50 | 0.003 | 0.155 | 0.196 |
| 0.5 | 1493 | 50 | 0.008 | 0.157 | 0.198 |
| 1 | 890 | 50 | 0.007 | 0.152 | 0.192 |
| 1.5 | 456 | 50 | 0.013 | 0.150 | 0.186 |
| 2 | 178 | 50 | 0.017 | 0.166 | 0.207 |
| 2.5 | 35 | 50 | −0.017 | 0.157 | 0.206 |
| 3 | 15 | 50 | −0.210 | 0.210 | 0.243 |
| 3.5 | 4 | 50 | −0.559 | 0.559 | 0.560 |

Figure 1
*Conditional BIAS across θ range for fixed length*

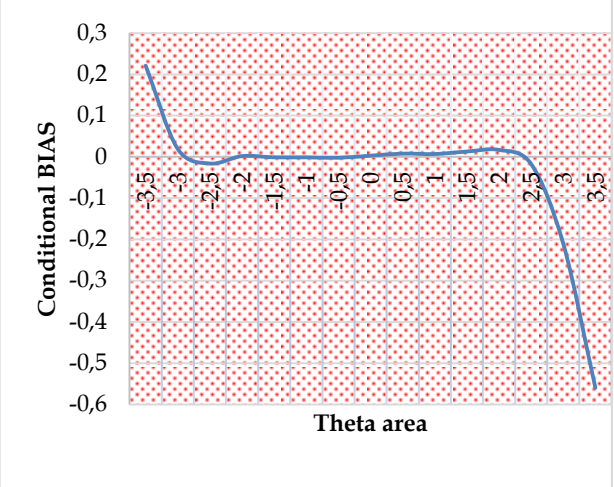



Figure 2
*Conditional MAE across θ range for fixed length*

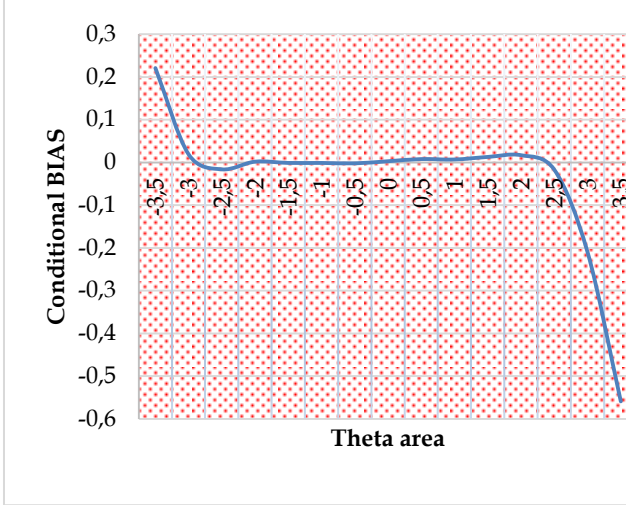
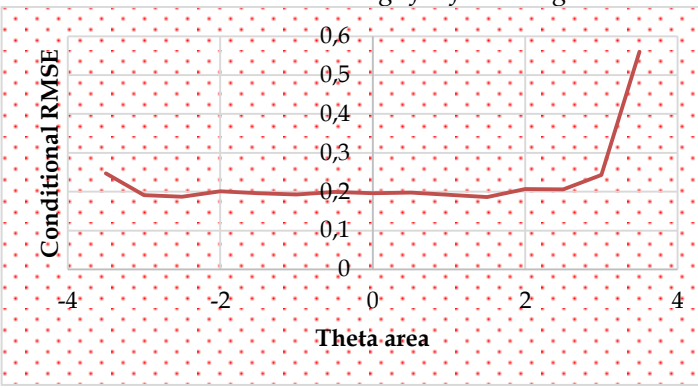



Figure 3
*Conditional RMSE across θ range for fixed length*




*Note.* Graphical Figure [1] alt text: The CBIAS values on Y-axis were plotted against theta (θ) areas on the X-axis for the fixed test length. In the vicinity of zero, there was a more significant line for CBIAS; For the fixed test length, alt text Graphical Figure [2] plots CMAE values on Y-axis against theta areas on X-axis. There was a greater significance of CMAE values at positive sides of the theta area; Alt text Graphical Figure [3]: The CRMSE values were plotted along theta (θ) areas on the X-axis for the fixed test length. As shown in Figure 2, this curve/shape is very similar.

Furthermore, the CAT system's measurements precision statistics were examined for variable length. Across all $\theta$ areas of the variable-length test, the simulation results tightly controlled the Conditional Standard Error of Estimation [CSEE]. The CAT system's statistics return a mean CBIAS close to zero, which was -0.02767. For the CMAE, 0.259067 was obtained, and for the CRMSE, 0.3193. As a result, Table 4 provides CBIAS, CMAE, and CRMSE scores for different $\theta$ areas of the variable length.

Table 4
*Variable-length statistics for measurement precision*

| Theta Area ($\Theta$) | Number of Cases | Test Length | CBIAS | CMAE | CRMSE |
|---|---|---|---|---|---|
| −3.5 | 9 | 23.67 | 0.270 | 0.270 | 0.323 |
| −3 | 41 | 20.05 | 0.121 | 0.215 | 0.274 |
| −2.5 | 165 | 18.76 | 0.058 | 0.233 | 0.288 |
| −2 | 435 | 19.31 | −0.003 | 0.224 | 0.283 |
| −1.5 | 872 | 19.61 | 0.002 | 0.239 | 0.299 |
| −1 | 1524 | 19.97 | 0.004 | 0.240 | 0.301 |
| −0.5 | 1953 | 20.41 | 0.004 | 0.249 | 0.314 |
| 0 | 1927 | 19.95 | 0.006 | 0.242 | 0.305 |
| 0.5 | 1493 | 19.11 | 0.010 | 0.235 | 0.296 |
| 1 | 890 | 19.02 | 0.001 | 0.255 | 0.317 |
| 1.5 | 456 | 18.90 | 0.013 | 0.242 | 0.301 |
| 2 | 178 | 18.74 | 0.013 | 0.225 | 0.281 |
| 2.5 | 35 | 19.51 | −0.068 | 0.171 | 0.229 |
| 3 | 15 | 20.60 | −0.261 | 0.261 | 0.307 |
| 3.5 | 4 | 20.00 | −0.585 | 0.585 | 0.591 |

In Figure 4, the observed systematic error for the variable test length was also close to zero. However, it was greater than the fixed test length. It produces close results with a fixed length when measurement precision is as low as possible. Figure 5 shows CMAE across the theta continuum, highlighting an overall measurement error higher than that found for the fixed-length test. RMSE for the variable-length tests was consistent at 0.3, as shown in Figure 6. It appears that the fixed-length tests are more precise than the variable. Similarly, variable-length tests are longer than the typical ones and provide little precision improvement.

Figure 4
*Conditional BIAS across θ range for variable length*

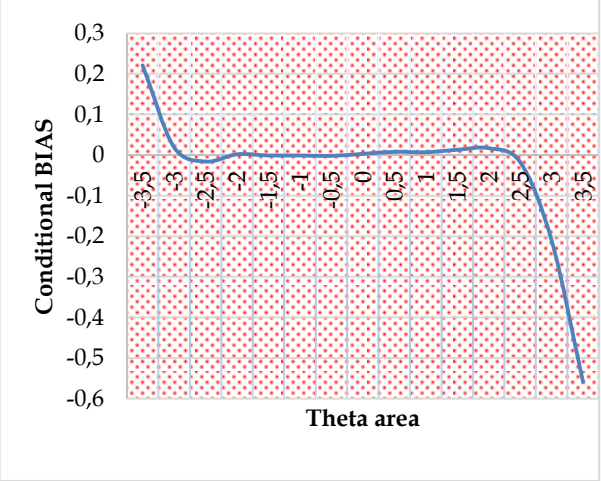



Figure 5
*Conditional MAE across θ range for variable length*

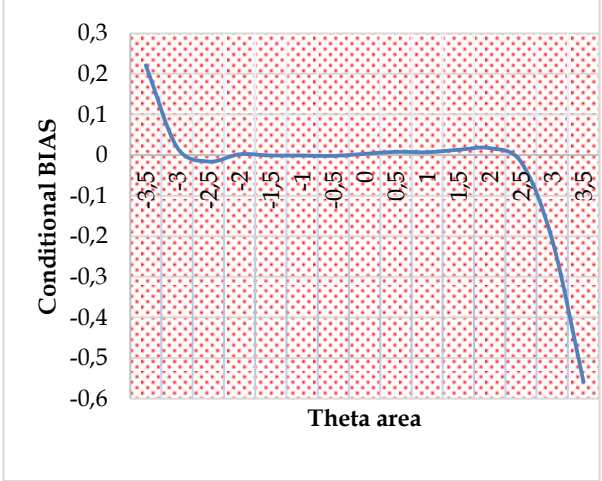
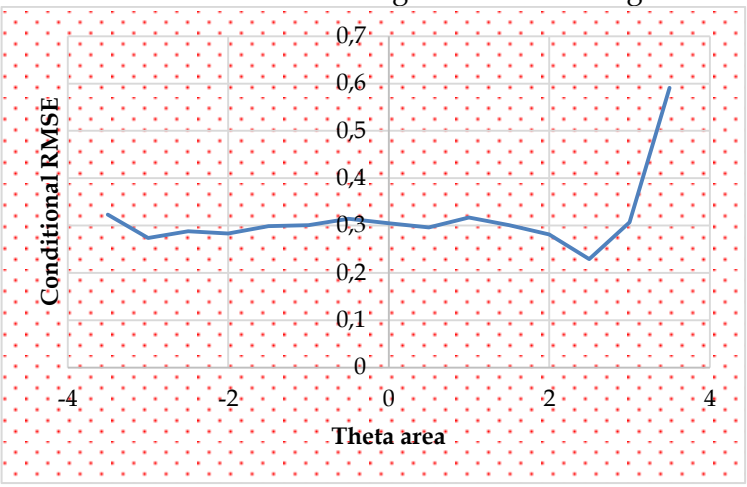


Figure 6
Conditional MAE across $\theta$ range for variable length




*Note.* Graphical Figure [4] alt text: CBIAS values of the variable test length are plotted against theta areas. Instances of zero are significantly higher for CBIAS; In alt text Graphical Figure [5], CMAE values are plotted against theta areas on Y-axis for the variable test length. When theta values were positive, CMAE values were more significant; Alt text Graphical Figure [6]: The CRMSE values were plotted along theta (θ) areas on the X-axis for the variable test length. It is very similar to the curve/shape shown in Figure 5.

Next is the evaluation of fixed-length item exposure control. SimulCAT's item usage output file (*.scu) was used to determine the exposure profile. Seven thousand, three hundred and ninety-seven (out of 10,000 test administrations/simulees) were considered the maximum observed item exposure rate for fixed length. This suggests that the randomesque method and its setting (one of the best five items) were not compelling enough, as more than two-thirds of the simulated candidates were exposed to the item, implying that it was overexposed. The fixed-length item pool also had 227 items not used (45.4%). Therefore, item exposure and item redundancy seem to be inversely related. A variable-length test would be more conclusive.

Evaluation of variable test-length item exposure follows. Based on the item usage output file from SimulCAT, we investigated the item exposure profile and found the maximum observed item exposure rate was 6140 (out of 10,000 simulations). As with the fixed-length test, an item was overexposed when more than half of the simulated candidates saw it. Furthermore, 342 items (68.5%) of the 500 items that made up the variable test length were not used. With fixed-length tests, items were maximized better than with variable-length tests. More so, Figure 7 showed that the most discriminating value (*a*-values) were the items most preferred in CAT design studies (Han, 2018b). An alternative method for handling this issue is to switch from using the maximum fisher information method to using a *b*-matching approach, which does not consider *a*-values (Han, 2018a). A glance at Figure 8 shows item difficulty (*b*-values) in the pool; there was no shortage of items with various difficulty levels.

## 5. Discussion and Implications

The study delves into the feasibility and effectiveness of integrating Computerized Adaptive Testing into the National Benchmark Tests, employing the three-parameter logistic Item Response Theory framework. The examination of item psychometric characteristics reveals a diverse range of difficulty levels, affirming the comprehensive coverage of the item pool. Discrimination parameters further underscore the item pool's effectiveness in distinguishing between low and high-ability test-takers. This finding resonates with the observations made by Ogunjimi et al. (2021), indicating that when item parameter statistics fall within the specified range, it signifies a thorough representation of the theta level across all items in the item bank. In fixed-length tests, the CAT system demonstrates a commendable ability to recover true θ parameters with minimal systematic error (CBIAS close to zero). Low Cumulative Mean Absolute Error and Cumulative

Figure 7
*Item exposure by a-values*

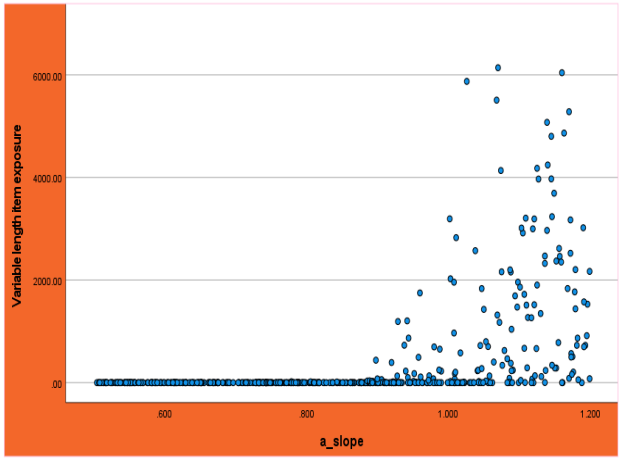
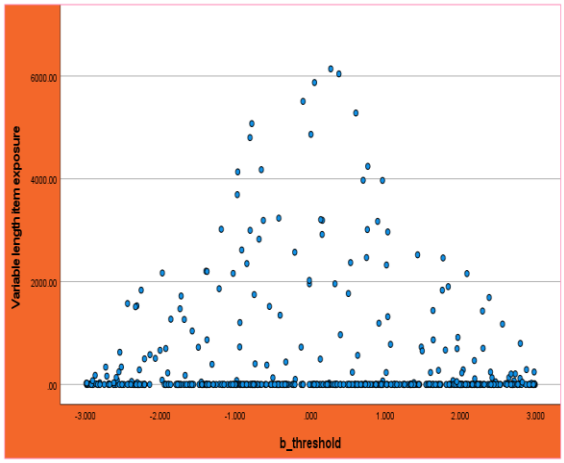



Figure 8
*Item exposure by b-values*




*Note.* A plot of variable test length item exposure against discriminating index values (a-slope) is shown in alt text Graphic Figure [7]. A skewed pattern of items was observed; Alt text Graphical Figure [8] plots variable test length item exposure values against the difficulty index (b-threshold). The distribution of items across the b-thresholds was normal.

Root Mean Square Error values reinforce the system's overall low measurement error, emphasizing the reliability of the CAT system in providing accurate ability estimates. This aligns with the fundamental objective of CAT, emphasizing precise and efficient measurement. The results of this study agree with those of Ogunjimi et al. (2021), Han (2018b), and Tsaousis et al. (2021) study indicating that fixed-length tests guarantee higher testing precision with a less than zero systematic error.

In variable-length tests, the CAT system exhibits robust control over the Conditional Standard Error of Estimation. The CBIAS, CMAE, and CRMSE values affirm the system's accuracy and reliability in accommodating variable-length tests, reinforcing its adaptability and potential to optimize measurement precision across different ability levels. However, challenges emerge in fixed-length tests regarding item exposure, with overexposure observed in a significant proportion of simulated candidates. This highlights the need for meticulous item selection and exposure control in fixed-length CAT to ensure fair and reliable assessments. In contrast, variable-length tests showcase better control over item exposure, aligning with CAT's flexible nature and its ability to dynamically adapt the test length based on individual performance. This confirmed the conclusion of Ogunjim et al. (2021) and Burhanettin and Selahattin (2022) that fixed-length tests have a higher item exposure rate, which can be overcome by relying less on the accuracy parameter.

Practical implications for NBT examination emerge from these findings. Designing adaptive tests that effectively control item exposure could enhance the precision of the NBT examination, dynamically adapting to individual abilities for optimized measurement accuracy and fairness. The insights into item psychometric properties and exposure patterns offer practical guidance for managing the item bank, emphasizing the importance of careful curation based on difficulty levels and discrimination parameters. Consideration of the trade-offs between fixed and variable-length tests becomes pivotal for NBT administrators. Balancing measurement precision, fairness, and practical considerations in test administration will be crucial in optimizing the effectiveness of the NBT examination. The study underscores the importance of continuous feasibility studies, particularly through simulation research, to evaluate the performance of the CAT system at each developmental stage. This ongoing assessment and refinement process contributes to the efficiency and effectiveness of CAT within the NBT context, providing valuable insights into the evolution of assessment practices, particularly within the African educational landscape. The integration of adaptive testing has the potential to enhance the precision and fairness of examinations like the

NBTs, contributing to the ongoing discourse on the advancement of educational assessments in the region.

## 6. Conclusion, Limitations, and Future Work

In conclusion, the findings of this study suggest that it would be beneficial to conduct a feasibility study for the purpose of providing practicable or informed benchmarks for the CAT specification in a real-life setting. Ultimately, the fixed-length method produced better measurement precision, and that is why it is recommended for high-stakes exams, such as the NBTs. More so, CATs have the potential to replace full-length measures in many situations. In addition to providing accurate results with a minimum amount of measurement error, this method reduces the number of items required to be administered. It is important, however, to acknowledge that this study has limitations. Using simulation techniques introduces assumptions and models that may not be fully representative of real-world testing conditions, and the findings may not be fully generalizable outside the specific context of the research. Further, the study only compared fixed-length and variable-length CAT designs, potentially omitting other variations and modifications. Researchers can explore additional variations in test design and investigate other components of CAT in future research. CAT-based assessments could also benefit from validation studies that examine the validity and reliability of different test designs. It would also be more comprehensive to evaluate measurement precision when empirical data is gathered from actual test administrations in real-world situations and studies that look at the real-world implementation of CAT.

**Acknowledgements:** Dr Kyung (Chris) T. Han of the Graduate Management Admission Council developed SimulCAT, a free simulation program that generated the data for this study.

**Author contributions:** Musa Adekunle Ayanwale: Writing - Original Draft, Conceptualization, Methodology, Writing -Review & Editing, Data Simulation, Analyses, References alignment. Mdutshekelwa Ndlovu: Writing -Review & Editing, Supervision.

**Declaration of interest:** The authors declare that no competing interests exist.

**Funding:** No funding source is reported for this study.

## 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 and Development*, *2*(3), 68-80. https://doi.org/10.5296/ijld.v2i3.1775

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), AB157. https://doi.org/10.1016/j.jaci.2015.12.642

Asiyai, R. I. (2014). Improving quality higher education in Nigeria: The roles of stakeholders. *International Journal of Higher Education*, *4*(1), 61-70. https://doi.org/10.5430/ijhe.v4n1p61

Assessment Systems Corporation [ASC] (n.d.). *Adaptive testing simulations with CATSim*. https://assess.com/catsim/

Ayanwale M.A., Chere-Masopha, J. & Morena, M. (2022). The classical test or item response measurement theory: the status of the framework at the examination council of Lesotho. *International Journal of Learning, Teaching and Educational Research, 21*(8), 384-406. https://doi.org/10.26803/ijlter.21.8.22

Ayanwale, M. A. (2019). *Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria* [Unpublished doctoral dissertation]. University of Ibadan, Nigeria. https://doi.org/10.13140/RG.2.2.17461.22247

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2019). Invariance person estimate of Basic Education Certificate Examination: Classical test theory and item response theory scoring perspective. *Journal of the International Society for Teacher Education, 23*(1), 18–26.

Ayanwale, M.A. & Adeleke, J.O. (2020). Efficacy of item response theory in the validation and score ranking of dichotomous response mathematics achievement test. *Bulgarian Journal of Science and Education Policy,*

*14*(2), 260-285.
Ayanwale, M.A., Adeleke, J.O. & Mamadelo, T.I. (2018). An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach. *International Journal of Educational Research Review, 3*(4), 55-67. https://doi.org/10.24331/ijere.452555
Ayanwale, M.A., Isaac-Oloniyo, F.O. & Abayomi, F.R. (2020). Dimensionality assessment of binary response test items: a non-parametric approach of bayesian item response theory measurement. *International Journal of Evaluation and Research in Education, 9*(2), 412-420. https://doi.org/10.11591/ijere. v9i2.20454
Aybek, E. C. (2021). catIRT tools: A "Shiny" application for Item Response Theory calibration and computerized adaptive testing simulation. *Journal of Applied Testing Technology*, *22*(1), 12–24.
Aybek, E. C., & Demirtasli, R. N. (2017). Computerized adaptive test (Cat) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science*, *3*(2), 475–487. https://doi.org/10.21890/IJRES.327907
Baker, F. B. (2004). Item Response Theory: Parameter Estimation Techniques. *Biometrics*, *50*(3), 896. https://doi.org/10.2307/2532822
Baker, F. B., & Kim, S. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8_1
Baker, F.B. (2001). *The basics of item response theory* (ED458219). ERIC.
Burhanettin, O. & Selahattin, G. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies, 27,* 6273–6294. https://doi.org/10.1007/s10639-021-10853-0
Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16,* 133-141. https://doi.org/10.1007/s11136-007-9204-6
CETAP. (2019). *The national benchmark tests national report*. Centre for Higher Education, University of Cape Town.
CETAP. (2020). *Test dates National Benchmark Test (NBT)*. Central University of Technology.
Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1-38. https://doi.org/10.18637/jss.v071.i05
Chen, S. Y., & Lei, P. W. (2015). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, *29*(3), 204–217. https://doi.org/10.1177/0146621604271495
Choe, E. M., & Fu, Y. (2018). Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR. *Measurement: Interdisciplinary Research and Perspectives*, *16*(4), 264–267. https://doi.org/10.1080/15366367.2018.1520560
Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, *33*(8), 644–645. https://doi.org/10.1177/0146621608329892
Cliff, A. & Yeld, N. (2006). Test domains and constructs: Academic literacy. In H. Griesel (Ed.), *Acccess and entry level benchmarks: The national benchmark tests project* (pp. 19–25). Higher Education South Africa.
Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, *47*(2), 202–226. https://doi.org/10.1111/j.1745-3984.2010.00109.x
Educational Testing Services. (2014). *GRE A Snapshot of the Individuals Who Took the GRE ® revised General Test*. Author.
Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.
Erdem Kara, B. (2019). Computer adaptive testing simulations in R. *International Journal of Assessment Tools in Education*, *6*(5), 44–56. https://doi.org/10.21449/ijate.621157
Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychological Assessment*, *28*(8), 953–962. https://doi.org/10.1037/pas0000240
Frith, V., & Prince, R. (2006). Quantitative literacy. In H. Griesel (Ed.), *Access and entry level benchmarks, the National Benchmark Tests Project* (pp. 47–54). Higher Education South Africa.
Frith, V., & Prince, R. N. (2018). The national benchmark quantitative literacy test for applicants to South African Higher Education. *Numeracy, 11*(2), Article 3.
Han, K. C. T. (2018a). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*, Article 7. https://doi.org/10.3352/JEEHP.2018.15.7

Han, K. C. T. (2018b). Conducting simulation studies for computerized adaptive testing using SimulCAT: an instructional piece. *Journal of Educational Evaluation for Health Professions*, *15*, Article 20. https://doi.org/10.3352/jeehp.2018.15.20

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Han, K. T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement*, *36*(1), 64–66. https://doi.org/10.1177/0146621611414407

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, *40*(4), 289–301. https://doi.org/10.1177/0146621616631317

Han, K. T., & Kosinski, M. (2016). Software tools for multistage testing simulations. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 411–420). CRC Press. https://doi.org/10.1201/b16858-39

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer adaptive testing (CAT): A faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology*, *26*(2), 227–232. https://doi.org/10.1007/s10869-011-9228-3

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, *14*, Article 12. https://doi.org/10.3352/JEEHP.2017.14.12

Linacre, J. M. (2000). Computer-adaptive testing: a methodology whose time has come. In S. Chae, U. Kang, E. Jeon & J. M. Linacre (Eds.), *Development of computerised middle school achievement tests* (p. 60). Komesa Press.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*(4), 615–630. https://doi.org/10.1177/0013164499594004

Luecht, R. M. (2005). *Computer-adaptive testing*. Wiley. https://doi.org/10.1002/0470013192.BSA125

Luecht, R. M. (2016). Computer-Adaptive Testing. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (pp. 1–10). Wiley. https://doi.org/10.1002/9781118445112.stat06405.pub2

Luecht, R., & Sireci, S. (2011). A review of models for computer-based testing. *College Board Research Reports*, *1*, 1–56.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *76*, Article 1. https://doi.org/10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48,* Article 8. https://doi.org/10.18637/jss.v048.i08

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R*. Springer. https://doi.org/10.1007/978-3-319-69218-0

Mills, C. N., & Steffen, M. (2016). The GRE Computer Adaptive Test: Operational Issues. In W. J. Linden & G. A. W Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 75-99). Kluwer Academic.

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.

Nandakumar, G. S., & 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.

NBT. (2022). *More about the NBTs: National Benchmark Test Project*. Author.

Ogunjimi, M. O., Ayanwale, M. A., Oladele, J. I., Daramola, D. S., Jimoh, I. M., & Owolabi, H. O. (2021). Simulated evidence of computer adaptive test length: Implications for high stakes assessment in Nigeria. *Journal of Higher Education Theory and Practice*, *21*(2), 202–212. https://doi.org/10.33423/JHETP.V21I2.4129

Oladele, J. I., Ndlovu, M., & Spangenberg, E. D. (2022). Simulated computer adaptive testing method choices for ability estimation with empirical evidence. *International Journal of Evaluation and Research in Education, 3,* 1392-1399. https://doi.org/10.11591/ijere.v11i3.21986

Oladele, J.I., Ayanwale, M.A & Owolabi, H. (2020). Paradigm shifts in computer adaptive testing in Nigeria in terms of simulated evidences. *Journal of Social Sciences, 63*(1-3), 9-20. https://doi.org/10.31901/24566756.2020/63.1 -3.2264

Oladele, J.I., Ayanwale, M.A. & Ndlovu, M. (2023). Simulated computer adaptive testing administration: trajectories for off-site assessment. *PONTE: International Journal of Sciences and Research, 79*(6), 41-50.

https://doi.org/10.21506/j.ponte.2023.6.4

Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish journal of psychology, 15*(1), 424-441. https://doi.org/10.5209/rev_sjop.2012.v15.n1.37348

Prince, R. N., Frith, V., Steyn, S., & Cliff, A. (2021). Academic and quantitative literacy in higher education: Relationship with cognate school-leaving subjects. *South African Journal of Higher Education, 35*(3), 163-181.

Prince, R., Balarin, E., Nel, B., Padayashni, R. P., Mutakwa D., & Niekerk, A. D. J. (2018). *The National Benchmark Tests national report: 2018 intake Cycle*. NBT.

Robitzsch, A. (2021). A comprehensive simulation study of estimation methods for the Rasch model. *Stats, 4*(4), 814-836. https://doi.org/10.3390/stats4040048

Sango, T., Prince, R., Steyn, S., & Mudavanhu, P. (2022). High-stakes online assessments: A case study of National Benchmark Tests during COVID-19. *Perspectives in Education*, *40*(1), 212–233. https://doi.org/10.18820/2519593X/PIE.V40.I1.13

Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment: new approaches to skills assessment and implications for large-scale testing.* European Commission. https://doi.org/10.2788/60083

Sebolai, K. (2014). Do the academic and quantitative literacy tests of the national benchmark tests have discriminant validity?. *Journal for Language Teaching, 48*(1), 131-147.

Seo, D. G. (2017). Overview and current management of computerized adaptive testing in licensing/certification examinations. *Journal of Educational Evaluation for Health Professions*, *14*, 17. https://doi.org/10.3352/JEEHP.2017.14.17

Thompson, G. (2017). Computer adaptive testing, big data and algorithmic approaches to education. *British journal of sociology of education, 38*(6), 827-840.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1), Article 1. https://doi.org/10.7275/fq3r-zz60

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*(5), 778–793. https://doi.org/10.1177/0013164408324460

Thompson, N. A., & Weiss, D. J. (2009). Computerized and adaptive testing in educational assessment. In F. Scheuermann & J. Björnsson (Eds.), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 127–133). European Commission.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research and Evaluation*, *16*(1), 1–9.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, and Evaluation, 16*(1), 4.https://doi.org/10.7275/wq8m-zk25

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2021). Evaluating a computerized adaptive testing version of a cognitive ability test using a simulation study. *Journal of Psychoeducational Assessment*, *39*(8), 954–968. https://doi.org/10.1177/07342829211027753

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing.* Springer. https://doi.org/10.1007/978-0-387-85461-8

Van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). Springer. https://doi.org/10.1007/0-306-47531-6_1

Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação, 21,* 57-82. https://doi.org/10.1590/S0104-40362013005000001

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 291–305). Springer. https://doi.org/10.1007/978-3-030-18480-3_15

Wang, W., & Kingston, N. (2019). Adaptive testing with a hierarchical item response theory model. *Applied Psychological Measurement*, *43*(1), 51–67. https://doi.org/10.1177/0146621618765714

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*(6), 774–789. https://doi.org/10.1037//0022-006X.53.6.774

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica*, *29*(1), Article 18. https://doi.org/10.1186/s41155-016-0040-x

Zhang, Y., Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a computerized adaptive testing for internet addiction. *Frontiers in Psychology*, *10*(5), 1010. https://doi.org/10.3389/FPSYG.2019.01010