Cypriot Journal of Educational Sciences

Volume 17, Issue 9, (2022) 3327-3343

www.cjes.eu

Transition from computer-based testing of national benchmark tests to adaptive testing: Robust application of fourth industrial revolution tools

Musa Adekunle Ayanwale [a] *, University of Johannesburg, Johannesburg, Auckland Park, 6000, South Africa
https://orcid.org/0000-0001-7640-9898

Mdutshekelwa Ndlovu [b], University of Johannesburg, Johannesburg, Auckland Park, 6000, South Africa
https://orcid.org/0000-0002-1187-0875

Suggested Citation:

Ayanwale, M. A. & Ndlovu, M. (2022). Transition from computer-based testing of national benchmark tests to adaptive testing: Robust application of fourth industrial revolution tools. *Cypriot Journal of Educational Science*. *17*(9), 3327-3343. https://doi.org/10.18844/cjes.v17i9.7124

Received from April 16, 2022; revised from August 11, 2022; accepted from September 20, 2022.

Abstract

In 2020, the COVID-19 pandemic strongly affected all sectors, including education. Across the globe, governments enacted policies restricting people's movement, affecting the educational assessment industry. During the lockdown, South Africa's National Benchmark Tests (NBTs) were cancelled. In contrast to paper-pencil and linear tests deployed for NBTs, computer-adaptive testing (CAT) is a modern alternative. CAT technology is widely used for licensing and certification exams in most developed nations. CAT is expected to be used by many educational assessment companies in sub-Saharan Africa for their high-stakes exams due to its reduced testing time and algorithms that produce stable and reliable results. This paper aims to provide an overview of the 4IR tools that will enable creation of a comprehensive CAT, such as feasibility studies, item bank development, test-and-calibrate item banks, and specifications for the final CAT. To successfully implement CAT, these steps are essential. The importance of thorough research and documentation in each stage of the CAT cannot be overemphasized. Based on Fourth Industrial Revolution (4IR) artificial intelligence tools, we conclude that CAT is an excellent modality to adopt for ensuring accurate evaluations of examinees' abilities within high-stakes exams.

Keywords: CAT framework, Computerised adaptive testing, Fourth industrial revolution, Linear testing, National benchmark tests

* ADDRESS FOR CORRESPONDENCE: Musa Adekunle, Ayanwale, Department of Science and Technology Education, Faculty of Education, University of Johannesburg, Johannesburg, Auckland Park, 6000, South Africa.
E-mail address: ayanwalea@uj.ac.za / Tel.: +27 (0) 64053-4022

1. Introduction

Despite scientific advances and technology, the novel coronavirus, known as the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), caused the highly contagious disease known as the coronavirus disease 2019 (COVID-19) in recent years, taking the world by surprise. A disease that originated in the small region of Wuhan, China, spread throughout the region and fast evolved into a pandemic that turned the world into a nightmare within a couple of years. In addition, the Africa continental disease control and prevention agency reported as of 25th April 2022 that the virus has spread to 58 countries throughout Africa, with 11,869,944 and 509,812,230 infected worldwide (Worldometer, 2022; WHO, 2022; Xinhu, 2022). In April 2022, the Africa CDC reported 253,736 deaths caused by the pandemic in Africa, compared with 6,244,471 deaths worldwide (Worldometer, 2022; Xinhu, 2022). Among the countries experiencing the most significant number of Covid-19 cases on the African continent, the most affected countries are South Africa (3,764,865), followed by Morocco (1,164,717), Tunisia (1,039,532), Egypt (515,645) and Libya (501, 862) (Worldometer, 2022). Within the past two years, South Africa has been hit by four waves of COVID-19, with each wave having a higher peak or a greater number of new cases than the previously reported one. Surges in SARS-CoV-2 have been largely attributed to new variants of the virus, which are highly transmissible but not necessarily more deadly than earlier waves. There has been a more effective response with each subsequent wave, with each surge having shortened by 23% on average compared to the one prior. In contrast to the first wave, which lasted nearly 29 weeks, the fourth wave ended after six weeks, which is about a fifth of the time (WHO, 2022).

The devastating COVID-19 pandemic severely impacted education and all related activities. Globally, the World Health Organisation (WHO) has an important role in finding solutions to the pandemic and recommending policy directions. Policy perspectives from around the globe proposed strategies for ensuring high levels of environmental hygiene and implementing measures such as lockdowns and curfews to curb the spread of diseases (Steward, 2020). Due to the Covid-19 guidelines, all physical gatherings, contacts, and educational activities ceased or were severely curtailed world-wide, including South Africa. During the lockdown that began on 27th March 2020, the Centre for Educational Testing, Access, and Placement (CETAP) in South Africa had to cancel the 2020 National Benchmark Tests (NBTs). The government's risk-adjusted strategy did not permit NBT candidates access to their examination venues across all the provinces. The cancellation of NBTs is presumed to significantly impact candidates' enrollment in the South African higher education institutions (HEIs). Consequently, this necessitates shifting the assessment paradigm to high-ended technology off-site. In a swift response to these challenges, on 25th July 2020, CETAP announced the migration of NBTs assessment to a securely proctored computer-based test (CETAP, 2020).

One of the entry requirements into select programmes at South African Universities is a score in the relevant NBTs. The NBTs are required especially for Science, Engineering, and Technology (SET) oriented programmes for the first-year applicants to HEIs. NBTs are intended to assess a student's ability to integrate Academic Literacy, Quantitative Literacy, and Mathematics into tertiary coursework. The NBT consists of three multiple-choice tests, one of which is a combined Academic Literacy and Quantitative Literacy test. Academic and quantitative literacy take three hours in the AQL (NBT, 2022; Prince et al., 2018). Each section of the test is scored separately. Second, the Mathematics (MAT) also takes three hours to complete and is multiple choice. The Academic Literacy (AL) test measures the ability of a writer to communicate effectively in a medium of instruction conducive to academic study. Quantitative Literacy (QL) tests measure a writer's ability to solve issues using

fundamental quantitative knowledge presented vocally, visually, tabularly, or symbolically in a natural setting relevant to higher education. An evaluation of a writer's ability to write within the context of secondary school math ideas relevant to higher education is conducted through the Cognitive Academic Mathematics Proficiency Test (CAMP).

However, the NBTs provide a different and complementary perspective to those found in the school-leaving examinations, although they cover the same type of content. When placing students in the right courses for their development, extending their education, or identifying other academic support options, most South African universities take NBT results into account, as well as school academic performance and examination results (Prince & Frith, 2017; Ayanwale et al., 2022).

More importantly, NBTs are still primarily deployed as a paper-pencil test, which had to be cancelled due to the need to comply with Covid-19 rules. Based on this unexpected development, there is a need for CETAP to shift its paradigm to more advanced technology-led assessment solutions. Digital technologies have turned into a method of individuals' work, life-molding, play, live, think, and this holds promise for educational assessments often implemented as computer-based testing (CBT). The CBT method involves administering examinations through a computer terminal and electronically recording and evaluating responses. Additionally, CBT can be administered in five different ways (Birdsall, 2011): (1) on a stand-alone computer; (2) in a dedicated centre; (3) at temporary test centres; (4) computer labs; or (5) when using a personal computer, laptop, netbook, tablet, or handheld device connected to the internet, preferably remotely proctored. Unless stand-alone computers are configured, for CBT to be successful, it usually requires network connectivity, with the most successful systems being able to link multiple computers together with the test delivery software and item banks, as well as transmit test materials, scores, and results quickly and efficiently (Birdsall, 2011).

CBT can be either linear or adaptive. A linear CBT, one of two types of CBT currently used in assessments leading to NBT scores (Redecker & Johannessen, 2013), belongs to the first of four generations of computerised educational assessments. In linear tests, individuals are given different questions without considering their performance level. This test consists of a full spectrum of questions ranging from the easiest to the most challenging. The same scoring process applies in the linear test as in a paper-pencil-based test (Kimura, 2017; Oladele et al., 2020). There are no differences between the linear and paper-pencil tests since the same set of test items is administered to every examinee.

In contrast, adaptive CBT is the second type of computer-based test. The test items are specifically tailored to each individual's ability level in this type of testing. As with adaptive tests, this tailoring occurs by tracking an examinee's performance on each item and then adjusting the next item to suit their abilities (Luecht & Sireci, 2011). Accordingly, the adaptive CBT process is explained by following these steps; (a) after receiving instruction regarding its use, an examinee reviews the first item selected from a bank of items that meets a predefined criteria. (b) If the correct answer to a question in the item bank is given, a more difficult question will be selected. When an item is answered incorrectly, another item from the item bank will be selected. (c) Each question is answered in a manner that reflects the examinee's provisional ability level, and the following items are selected with difficulty levels that are commensurate with the examinee's initial ability level. An estimate of the examinee's final score is generated after the test has met a prescribed end criterion. (d) Throughout the testing process, the process continues until the predefined end criteria has been reached. In this way, it is possible to understand the process in a more detailed manner. In practice, however, CAT is

more sophisticated, taking into account balanced content, the likelihood of cheating, and items specific to subgroups. As a primary criterion in computer adaptive testing (CAT), the test information function should be maximised while the measurement error should be minimised, thereby assuring accuracy and precision in estimating the examinee's ability (Kimura, 2017).

Succinctly, while the battle is against COVID-19 today, tomorrow remains uncertain. Therefore, educational experts should continue to invest in the best assessment practices that could survive unforeseen contingencies. Ensuring the quality of assessments undertaken, a solution that would comply with the COVID-19 safety rules becomes necessary. This calls for the urgent deployment of the Fourth Industrial Revolution (4IR) tools to boost the validity of assessment procedures undertaken by CETAP. Therefore, it is imperative at present to place the assessment mode of NBTs on the same pedestal as those of other high-stakes exams administered in developed countries, such as Medical school candidates in Australia take a Multiple Choice Question (MCQ) exam, while medical school candidates in Canada take the Medical Council of Canada Qualifying Examination Part 1 (MCCQE Part 1), the Test of English as a Foreign Language (TOEFL), the Physical Therapist licensing exam, the Graduate Record Examinations (GRE), the Armed Services Vocational Aptitude Battery (ASVAB), and the Graduate Management Admission Test (GMAT). Thus, this paper aims to demonstrate that 4IR tools can be effectively applied and outline the steps to develop the CAT framework suggested by Thompson & Weiss (2011). These steps are feasibility study, item bank development, or using an existing bank, pre-testing and calibrating the item bank, determining specifications for the final CAT, and publishing a live, online CAT. For CAT to be successful from a practical standpoint, these procedures must be followed because there can be no validity without them. In summary, a CAT that is not properly researched and documented in each stage is prone to inefficiency and, in some cases, an indefensible outcome (Thompson & Weiss, 2011).

2. Feasibility Study

CAT development is a feasible approach for CETAP's testing and assessment program, but it must be determined if it is appropriate or not. The CAT algorithm is theoretically fascinating and offers some well-known advantages (such as significantly more reliable results compared to paper and pencil tests, require less time to administer, and therefore save funds), provide students, teachers, and others with immediate feedback, shorter test length, can be administered on demand at several locations and increases statistical accuracy of an assessment, etc), but non-psychometricians may become enamored of the concept and turn to CAT implementation without having an in-depth understanding and technical understanding of its operation. Thus, migration to CAT can be quite risky from a psychometric and business standpoint. Converting an assessment program from fixed-form tests (paper-pencil test or linear CBT) to CAT is not a decision to be made frivolously (Birdsall, 2011; Thompson & Weiss, 2011). Consequently, educational assessment organisations need to ask themselves the following questions: do they have psychometric expertise, or can they afford a consultant if they use an external one? Is the organisation well prepared to create a large data bank of items? Do they have the resources to develop their own CAT engine, or do they have access to an affordable CAT delivery engine? Is there a likelihood of the test length being reduced when converted to CAT? Is there enough time saved by reducing test length to translate to actual monetary savings? Seat time is often costly, so will the reduced test length translate into actual savings? The cost of CAT is higher, and CAT does not substantially reduce seat time but is the higher precision and security of CAT enough to compensate for those limitations to make it worthwhile (Steffen, 2016; Thompson & Weiss, 2011). These questions can be answered through psychometric research, not simply

conjecture. In addition to estimating the length of the test and the precision of the results of CAT (van der Linden & Glas, 2010; van der Linden & Ren, 2020). In addition, Monte Carlo simulation studies offer insight into the difficulty of item exposure as well as the size of the item bank that will produce the desired precision for examination scores.

3. Development of item banks for CAT

Organising and storing test items in a central database is known as item banking, which enhances efficiency and quality by storing test items (Smarter Assessment, 2022). Even though items are referred to as questions, their format does not need to be limited, and they can also include situations to evaluate and problem-solving activities. Item banking uses best practices as a foundation for the test development cycle to develop valid, reliable content and secure test forms. For those responsible for educational assessment, such as CETAP, having a quality item bank makes the process easier and more efficient, which reduces the cost of item development. CAT item banks can include both existing paper and pencil items as well as new items, which will maintain psychometric properties while minimizing the cost implications associated with adding additional items (Birdsall, 2011; Thompson & Weiss, 2011). Oladele and Ndlovu (2021) citing Germain (2006), suggested that items should be developed considering content, test domains, and internal consistency, thus ensuring that behavioural measures are valid psychometrically. Most tests are composed of carefully designed items to ensure that they are valid, able to test the subjects intended, and reliable, i.e., they are consistent between administrations. CAT item banks must be prepared according to a range of decisions made according to the test's purpose.

3.1 Planning

The most pertinent stage in the CAT test development is planning. In this stage, three things are crucial: the goals of the test must be determined, the test blueprint preparation, and the selection of suitable item types. Thus, the purpose of an examination must be patently stated in conformity with the taxonomy of educational objectives (Huit, 2011). Educational learning is classified into three main domains as observed by taxonomy. They include cognitive, affective, and psycho-productive aspects. This section focuses on cognitive learning because it has to do with knowledge and acquiring intellectual skills. In accordance with revised Bloom's taxonomy, cognitive learning outcomes are categorised according to six levels of complexity (Radmehr & Drake, 2019). Among them are: remembering (to bring to mind or recall previously learned information), understanding (understanding how instructions and problems are translated into other languages, interpolation, and interpretation, and stating a problem in one's own terms), applying (to use a concept to solve a problem unprompted), analysing (dividing materials into parts in order to determine their organisational structure and distinguishing between facts and inferences), evaluating (judging the value of ideas, materials, or objects) and creating (through the combining of elements into something new or a new pattern), i.e., creating new meaning or structure from them).

Specifying test blueprints is another important part of test development. The test blueprint must specify how it will measure a representative sample of instructional objectives and content areas. The blueprint serves as a guide for the development of test items. There are three dimensions to a test blueprint. They include the behavioural objectives, content areas, and types of items. There are several steps involved in this area; the weight to each of the distinctive behavioural objectives is determined, the weight to each of the content areas is determined, the item types to be included are determined, and then a chart or table is prepared. In Table 1, we present a cognitive classification of content areas assessed for quantitative literacy, academic literacy, and mathematics.

Table 1. NBTs Cognitive classification of Content Areas

Content Areas		
Mathematics	Quantitative Literacy	Academic Literacy
The NBT mathematics achievement test is designed to assess candidates' ability in several mathematical topics, such as algebraic processes, trigonometry, Spatial perception (angles, symmetries, measurements, etc.), interpreting three-dimensional objects, analytic geometry data handling, and probability (le Roux & Sebolai, 2017) cited in (Ayanwale et al., In press).	According to Prince & Frith (2017), QL assessments measure students' ability to interpret and reason analytically about quantitative situations presented in different contexts, such as Quantity, number, and operations, Shape, dimension, and space, relationships, patterns, and permutations, Change, and rates, as well as data presentation and analysis.	This assessment measures the ability of first-year students to name, distinguish, and use a variety of different communication purposes within the academic language, such as, for example, distinguishing between essential information and less-essential information, extrapolating, inferring, and applying, using parallel and metaphorical language, using academic and general vocabulary, using text genre, using grammar and syntax, and using textual coherence features (Cliff, 2015).

It is also important to determine what item types are appropriate for a test blueprint. Items are developed according to a test blueprint. A blueprint describes the nature of skills to be measured and the balance of test content. Test format and item types should be selected according to the skills to be assessed, not how one feels about particular item formats. First, the selected-response format refers to items that give the examinee several choices, from which they must select the correct answer. Among these are multiple-choice, true/false, and matching items. The second type of format is constructed response, in which an examinee must produce or generate their responses. The second type can also be divided into three categories: essay, completion, and short answer. Stems, alternatives, keys, and decoys define multiple-choice items. The stem is the part of the item that specifies the problem for the examinee. A man bought a plot of land for N 480,000.00 and sold it for a profit of 20%. How much did he sell the land for? The options for the question include the correct answer, also known as the key, and many incorrect choices, also known as decoys.

3.2 Items writing

After planning, preparing items that will form the test is another crucial stage. The test items are developed according to the test blueprint in this stage. We need many more items than we would have in a traditional paper-and-pencil measure in a CAT. This is because we will convert one item into multiple forms. When items are developed based on test blueprints, the items becomes relevant. Thus, in developing relevant test items, all the item-writing rules must be considered. For CAT item writing, it involves developing what is considered an activity-centered assessment task to gauge the amount of knowledge and skills students have gained through exposure to teaching and learning. CAT leverages the item information function to determine the assessment tasks' difficulty, discrimination, and guessing to be precise and aligned to learning objectives (Veldkamp & Verschoor, 2019). This is why CAT must have a professional approach to item writing, contributing to its effectiveness. CAT items were written following these steps: Conducting literature reviews, developing new items or modifying existing test items, conducting field testing through computers, and performing psychometric analysis to select the final items (Dirven et al., 2021; Petersen et al., 2016). Prior to field testing, an expert evaluation is required to verify face and content validity.

In addition, there are four-step processes for writing items; however, the method requires tailor-fit software (Smarter Assessment, 2022). During the first stage, feasibility and planning studies are conducted using SimulCAT, after that, a comprehensive assessment ecosystem called FastTest is used

to create an item bank (Smarter Assessment, 2022). Pilot testing of items is undertaken on FastTest in the third stage, while in-depth analysis is conducted on Jmetrik in the fourth stage. In Jmetrik, the item response theory model can be used to calibrate assessments based on dichotomous models and polytomous models alike (Aksu et al., 2019). Although we emphasise that CAT is not a simple task, the ultimate objective is to simplify the process while following best practices and international standards by using free, clean software without any coding. Thus, ascertaining quality item writing skills requires constant practices and painstaking reviews by the subject experts.

3.3 Review of items by Experts

The writing of test items requires high expertise, and item review is mandatory. The item should be well stated (free from ambiguity), inappropriate sentence structure should be shunned, and items should be worded so that all examinees understand the task. The writer should also avoid any clues in the item's description, which may help students to answer correctly or solve another question - such as grammatical inconsistencies, verbal associations, extreme words, or mechanical features. It is important for the examinees to have a shared understanding of all the test items. In addition, a written statement of the testing objectives maintains the integrity of the test. As part of the item development process, it is crucial to review items in preparation for empirical testing, which is supposed to confirm the test measures what it claims to measure. Steffen (2016) suggests that the best practice, when evaluating test items relating to instructional objectives and blueprints, is to have independent subject experts evaluate them. In Birdsall (2011), he discusses how item review drives test development, assessment strategies, and curriculum design. Furthermore, the review process should include consideration of item scoring, the availability of practice items, and the appropriate time for actual testing; truncating the score key would create interpretation problems and provide test administration information during the pre-test assessment.

4. Pre-testing and item calibration

It is important to test items in exam-like conditions after being developed and reviewed. Wheadon et al. (2009) state Pretesting is the most reliable way for those who set and evaluate tests to ensure quality. Typically, pre-test items are seeded into existing exams as part of a pre-test. The practice of seeding in a live test is to place items that will not count towards a candidate's score. Pommerich et al.(2009) noted that even when the testing method is different, seeding appears to produce comparable results. Pre-tested items are seeded to ensure that candidates respond with the same enthusiasm as they would if they were answering real items, although those items may not affect a candidate’s score.

Furthermore, as selection criteria, the developed items can be pre-tested by representative groups of students from the target group. A calibration (item parameter) is used to assess item quality based on alignment with the test model specification, namely discrimination (corresponds to a), difficulty (corresponds to b), and pseudo-guessing (corresponds to c) (Oladele et al., 2020). The testing and administration modes should be carefully planned based on the target group's demographic profile. Ayanwale and Adeleke (2020) opined that pre-testing would assist in establishing item parameters, determining the number of final test items, determining whether practice items should be included in the final test, determining if testing time is appropriate, and analyse student responses.

4.1 Analysis of pre-test data

This calibration stage involves additional statistical analysis. Analysing students' responses using various methods is an essential component of item analysis. Each item is evaluated systematically to determine how effective it is. A CAT item bank must be evaluated to confirm the unidimensional assumption, which states that responses to each item are influenced by a single latent characteristic of the participants (Embretson & Reise, 2013); when the ability is conditional, determine the appropriate item response theory model based on test-level model-fit indices, and ensure that internal and external examinee distributions on one item are not relevant to other test items (Cohen, 2013); item response function refers to how examinees' likelihood of answering correctly on a particular item corresponds to their abilities, and items with different item response functions between subgroups and other items are said to be biased. Different methods exist for detecting items that behave differently, such as the Mantel-Haenszel procedure, the Logistic Regression procedure, the Multiple Indicators Multiple Causes (MIMIC) model, the likelihood ratio test of item response theory (IRT-LR), the Lord's IRT-based Wald test, and the simultaneous item bias test (SIBTEST) (Aybek & Demirtasli, 2017; Zhang et al., 2019). CAT is an item response theory-based technique used for solving a wide variety of measurement problems, which is useful for building tests, identifying items that may be biased, equating scores from different tests or forms of the same test, and reporting test scores. The IRT comprises a family of models (such as the one-parameter logistic 1-PL, the two-parameter logistic 2-PL, the three-parameter logistic 3-PL, and the four-parameter logistic 4-PL), which have been widely used to design, develop, and evaluate educational tests, as well as to establish their psychometric properties.

4.2. Models in item response theory based CAT

In unidimensional item response models, the number of items described by each parameter is one of the primary differences. It is up to the user to choose one of these models, but their choice involves making assumptions about the data that can be checked by examining how well they explain the observed results. Unidimensional IRT models can be classified into three groups according to the number of item parameters they incorporate: one-, two- and three-parameter logistic models; the four-parameter logistic model has not yet been fully explored. With these models, CAT dichotomous responses can be fitted.

4.2.1. *One-Parameter* Logistic *Model*

The Rasch model is the simplest of the four models (Rasch, 1960). Ayanwale et al. (2018) argue that this model assumes that the chance of a student answering a question correctly is based upon a logistic function of the examinees' ability (θ) and the difficulty of the question (b). It is expressed as:

$$Pi(\theta) = \frac{e^{1.7(\theta - b_i)}}{1+e^{1.7(\theta - b_i)}} \qquad \text{Eqn. 1}$$

Where $Pi(\theta)$ assesses the likelihood that an examinee has the necessary ability (θ) to correctly answer item i, b_i is the difficulty parameter of item *i* known as item location parameter, e is the transcendental number (like π) whose value is 2.718, ϑ (Theta) is the ability level of a particular examinee and D (1.7) is the scaling factor for logistic function, respectively. More so, for an item, the b_i parameter indicates the point on the ability scale at which the probability of a correct response is 50%. This parameter indicates the ability scale's item characteristics curve (ICC) position (Baker & Kim, 2017). As b_i increases, examinees must have more ability to get 50% of the items correct. Item

difficulty is determined by where it is located on the ability scale: to the right or higher-end; to the left or lower end. Next is a logistic model with two parameters.

4.2.2. Two-Parameter Logistic *Model*

Birnbaum (1968) proposed the two-parameter logistic function. The 2-PLM differs from 1-PLM by adding two additional parameters. The scaling factor provides a function close to a normal Ogive if D is introduced. Two-PLM also includes a parameter called 'a,' which is also known as item discrimination. The slope of the ICC at *bi* determines the "a" parameter on the ability scale. Clearly, steeper slopes are better at distinguishing between different abilities than items with a lower slope. The model is as follows:

$$Pi(\theta) = \frac{e^{Da(\theta - b_i)}}{1+e^{Da(\theta - b_i)}} \qquad \text{Eqn.2}$$

Theoretically, item discrimination parameters are defined as (-∞, +∞). The items that negatively discriminate from ability tests are discarded because they are flawed. In general, the likelihood of the examinee answering a question correctly decreases with increasing ability (de Ayala, 2009). A value of 'a' greater than 2 is uncommon, so the usual range is between 0.20 and 0.30. High values of 'a' result in steep increasing ICCs, while low values result in ICCs that increase gradually. This is followed by the 3-PLM.

4.2.3. Three-Parameter Logistic *Model*

A lower asymptote parameter "c" is included in 3-PLM for multiple-choice and true-false testing. An item's parameter is novel and independent of the examinee's ability. ICC is expressed as:

$$P_i(\theta_s) = Pr(X_{is} = 1|\theta_s,\, a_i, b_i\,, c_i) = \; c_i + (1 - c_i)\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \qquad \text{Eqn. 3}$$

On an asymptotic scale, an item with a non-zero ICC is more likely to be answered correctly by examinees with low abilities. The parameter c was therefore added to the equation in order to accommodate selected-response tests (multiple choice) at the low end of the ability spectrum. The letter *"c"* indicates the probability of receiving the correct answer based on a single guess. As c is a constant, its value does not change according to the level of ability. Due to this, the chances of a low ability examinee winning the item by guessing are identical to that of a high ability examinee winning the item by guessing. Generally, the parameter c has a theoretical range of $0 \leq c \geq 1.0$; however, for practical purposes, values higher than 0.35 are not acceptable. Therefore, the range $\theta \leq c \leq 0.35$ is typically adopted when 3-PLM is used. In the IRT model for CAT, only appropriate item pools are used to select items according to examinee ability levels over a series of iterative steps from item selection to stopping criteria until an accurate estimation of theta (θ) level is achieved. An ideal theta level should contain hundreds of items that are uniformly distributed in difficulty, highly discriminating, and have a low guessing parameter for the purpose of effectively assessing the test assembly (Birdsall, 2011; Oladele & Ndlovu, 2021).

4.2.4. Four-Parameter Logistic *Model*

To overcome some estimation problems 3-PL model posed with, 4-PL model was developed where an upper asymptote known as carelessness parameter *'d'* was added to the model (Barton & Lord, 1981). This was made possible such that a high-ability examinee who as a result of carelessness (that could result from mistake, stress, tiredness, inattention, anxiety, afraid of computers, frustrated

with poor testing conditions, and unable to understand questions) responded to an easy item incorrectly. Reise and Waller (2009) submits that including both a lower (c) and upper (d) bound to the ICC will improve the fit of the model. 4-PLM mathematical formulation is given as:

$$Pi(\theta_s) = Pr\left(X_{is} = 1 \middle| \theta_s,\ a_i, b_i, c_i, d_{i,}\right) = c_i + \left(d_{i,} - c_i\right)\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \qquad \text{Eqn.4}$$

On the contrary, lack of consensus on the effectiveness of the 4-PL model and its overwhelming dominance in the literature strongly argue against its use in developing CAT item banks (Loken & Rulison, 2010). The specifications for the final CAT must then be determined based on the five components shown in figure 1.

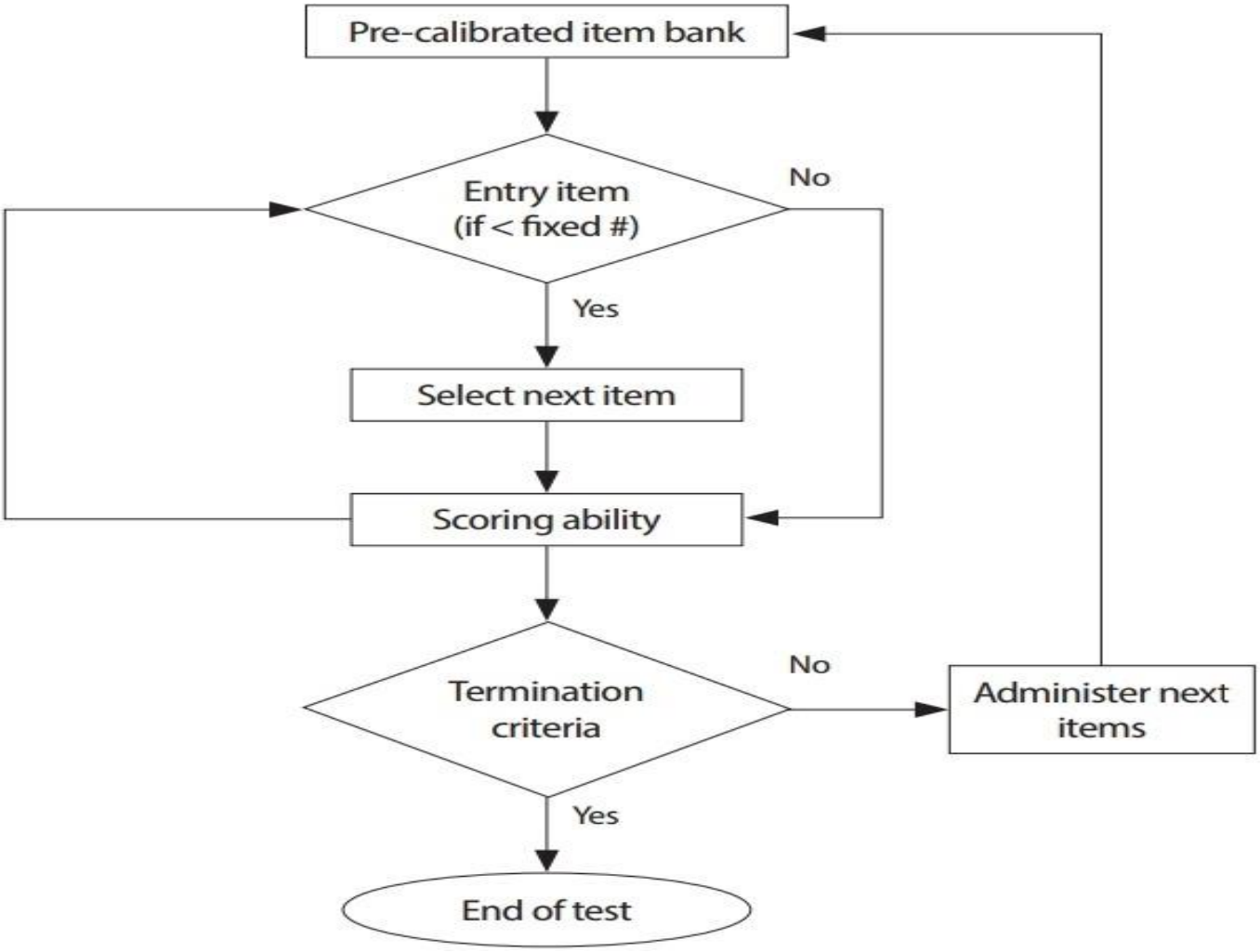

Figure 1. CAT Algorithm Components.

Source: (Seo, 2017)

5. Specifications for final CAT

5.1 Item bank

CAT requires the development of an item bank containing the items that may be administered during the test, as well as the parameter values related to those items. Consequently, to develop a large item bank, subsets of items administered to different groups must be linked to a reference group. It is possible for IRT to provide pre-calibrated parameter sets and a reasonable method of linking exam items based on the invariance properties of item and candidate parameters. In this manner, thousands of items can be pre-calibrated before CAT starts, resulting in a better CAT (Wang & Kingston, 2019). McClarty (2007); Oladele et al. (2020) rely on the choice of IRT model to estimate the relationship between examinee trait level and likelihood of responding. Thomson & Weiss (2011) suggested 500 items to be sufficient for dichotomous IRT models, but a higher number would be more appropriate for high stakes testing.

5.2. Starting item

It is imperative that a starting point is identified prior to implementing CAT. As a result of the difficulty in gathering accurate prior information about a candidate's ability level, the items in the CAT are usually chosen randomly. The selection of an initial item with a difficulty level that is close to the candidate's ability level is thought to enhance the effectiveness of CAT (Weiss, 1985) as cited in (Seo, 2017). In practice, it is not effective to start all tests at the same point, for example zero. Based on an estimate, the CAT algorithm determines the most appropriate item for each examinee, so if all candidates have the same estimate, all candidates will receive the same item. Such a situation could result in overexposure of the item. To reduce the exposure rate, several methods have been developed. Thompson and Weiss (2011) suggest selecting a subset of the item bank at random to select the first few items.

5.3. Item selection procedure

Item selection is the key component of CAT, and it determines what items an examinee is to be tested on after a starting item has been assigned. A selected item will be flagged so that the item cannot be selected again for the same examinee if it has already been selected (Han, 2018a). In CAT, item selection rules are determined by the item information function in IRT. It is closely related to the examinee's current trait level, thus enabling the selection of the most informative item among the remaining options (Seo, 2017). Additionally, several methods are available for selecting items. Among them are the maximum information procedure and the Bayesian selection procedure. The maximum information procedure allows the computer to choose an item that contains the most information for a given ability level. After each item, the procedure is repeated. Responses to previous items are used to estimate ability or trait level. The item selected by Bayesian selection minimizes the expected posterior variance of the ability estimates (Owen, 1975) as quoted in (Thompson, 2009). A likelihood ratio approach would be more effective to use as the item selection rule if the purpose of the exam is to categorise candidates by their cut-off score (Thompson, 2009).

5.4 Scoring procedure

When a test such as the NBTs exam is administered, the first item is expected to correspond to the level of ability required to pass. Based on the scores needed for the tests, an initial estimate of ability would be established (Ayanwale et al., 2022). After each item in a CAT is administered, a candidate's ability level can be updated based on their responses. The next item will be administered based on their ability level and responses. There are four ways to estimate a candidate's ability level: Maximum Likelihood, Maximum Likelihood Estimation with Fences, Bayesian Maximum a Posteriori and Bayes Expected a Posteriori (EAP) (Han, 2018b; Seo, 2017). The most common method for estimating ability

is maximum likelihood estimation (Lord, 2012). The goal of this procedure is to determine the level of the trait (theta value), which is most likely to be reflected in the examinee's response pattern. This procedure estimates how likely it would be for a person of a certain theta level to show the pattern of responses observed (x1, x2, ..., xn) for items with known item parameters. In this situation, the ability level of a candidate can be determined by combining the ICCs of the items. It should be noted, however, that Bayesian analysis can be applied to any response pattern since it is based on Bayes' law, which is proportional to the maximum likelihood plus the prior probability, assuming that the sample has a standard normal distribution (Birdsall, 2011). Bayesian types of estimators are used to determine the maximum value in a posterior distribution of abilities. It is estimated by applying the expected a posteriori method to the posterior distribution of abilities.

5.5. *Termination criterion*

In the final component of the CAT, a termination criterion is used for determining when a candidate has successfully completed the exam, based on a predetermined level of accuracy. The stopping rule may differ depending on the purpose of CAT. A number of approaches can be employed to accomplish this. These options can be used: after a predetermined number of items have been administered (fixed length method), when a minimum level of precision has been reached (variable length method), or while combining both approaches (Mcclarty, 2007). Examinations are administered to all examinees with a fixed-length CAT consisting of the same number of items. Users tend to understand the implementation easily, and the implementation is typically simple. Yet all examinees are expected to complete the same number of items, despite the degree of precision with which their abilities are estimated. Individuals at either end of the continuum may find the first few items to be less informative than those at the middle (Thompson & Weiss, 2011). The item pool also tends to contain fewer items at the extreme ends, which makes it more difficult to precisely measure the abilities of the extreme ends.

A CAT may cease to operate when there is a standard error or if there are no items in the item pool that meet a minimum level of information (Han, 2018; Oladele et al., 2020; Zhang et al., 2019). To determine whether another item is to be administered based upon the precision of measurement, the standard error procedure calculates the precision of measurement after the examinee responds to each item. This method ensures that all participants estimate results with the same degree of precision, though they may reach their stopping point by taking different numbers of test items. When an examinee responds to each item in the minimum level of information procedure, an item from the item pool is selected to provide additional information about the examinee. Examinees are terminated when they do not have any further items to complete in the CAT. There is compelling evidence to suggest that the standard error stopping rule is more effective for polytomous CATs than the minimum information-stopping rule (Birdsall, 2011; Lawless et al., 2002; Mcclarty, 2007). Conversely, students find that variable-length stopping rules are more difficult to comprehend than fixed-length ones (Dirven et al., 2021; Petersen et al., 2016).

6. Publish the live CAT

After all the specifications for all the components and any additional algorithms have been established, the final CAT can be published. There is little difficulty involved in this step if the test development and delivery software already exist. This step can be the most difficult if the organisation develops its platform.

7. The implication of 4IR for the CAT Algorithm

4IR is the present and future environment in which technologies and trends are transforming the way we live and work (Kayembe & Nel, 2019; Schwab, 2016), such as Internet of Things (IoT), Cyber Physical Systems, Smart Factory, 3D printing, robotics, blockchain technology, cryptocurrency, quantum computing, nanotechnology, bioengineering and artificial intelligence (AI). In the 4IR, information and communication technology (ICT) also plays an important role. As outlined by Lee et al. (2018), the 4IR represents the revolutionary changes that take place when ICT thrives across all industries, including education. A new generation of materials, products, and services are being developed and consumed thanks to these technologies, as noted by Skilton and Hovsepian (2018). Several implications arise from the 4IR in terms of skills development and education assessment, such as a high-end computer terminal provides multiple connectivity options and high storage capacities for educational assessments using CAT technology. Artificial intelligence breakthroughs based on emerging technology can enhance educational testing possibilities (Oladele & Ndlovu, 2021; Schwab, 2016).

Schwab (2016) argues that 4IR leverages AI to create benefits that are becoming realities, with strong evidence that the 4IR technologies will significantly influence businesses, and the educational sector is not immune. Compared to human intelligence, AI can carry out complex tasks efficiently. In contrast, AI can operate through computer algorithms and commands to respond to a certain behaviour (Perez et al., 2018; Xu et al., 2018). AI refers to the study of machines' ability to learn like humans through computer algorithms and commands and to act upon those actions. In the age of new technologies, computers can process large quantities of data and recognise patterns within that data to perform specific functions. Thus, using expert and knowledge-based systems, CAT relies on robust AI applications to place examinees according to their abilities.

8. Conclusion and Recommendation

For the quality of assessments to be assured, a solution that complies with the safety rules of COVID-19 is needed. To accomplish this, high-tech tools must be deployed immediately. As a result, this paper outlines the development of CAT for educational assessment based on the robust application of 4IR tools. A CAT is based on IRT, which explains examinees' responses to test items with a mathematical model using multiple parameters to explain examinees' interaction with the test items based on their likelihood of answering correctly. With this framework, the measurement accuracy is increased, ceiling and floor effects are reduced, the test management and scheduling is flexible, and it is easier to administer since each examinee gets a unique test, gets immediate feedback which motivates them, reduces test anxiety by getting items that are appropriate for the ability level. It is cost saving compared to conventional paper-pencil and linear tests, among other benefits.

There are some shortcomings to CAT, even though it has many advantages. These security issues may be compromised if the item banks are small or if an item exhibits item bias and is administered. In the case of such items, the scores of the affected examinees are likely to be affected more significantly. In terms of the first two issues, however, the development and implementation of large item banks and accurate item bias analysis practices can alleviate them. There has been much research in CAT on DIFs, DTFs, and item selection algorithms in recent years. Some candidates may be disadvantaged by the way the CAT is administered. When the hardware or software is not powerful enough, or when the CAT is not robust enough, items may be delayed. Consequently, CETAP should transit from linear CBT-based assessments leading to NBT scores, and placement into higher education institutions to adaptive CBT that incorporates the components of 4IR into their algorithm process.

References

Aksu, G., Güzeller, C. O., & Eser, M. T. (2019). Jmetrik: Classical test theory and item response theory data analysis software. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 165–178. https://doi.org/10.21031/EPOD.483396

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., Ndlovu, M., & Ramdhany, V. (2022). The Modus Operandi of National Benchmark Test Project in South Africa: A Systematic Review. *Journal of Higher Education Theory and Practice*, 22(4), 105-125. https://doi.org/10.33423/jhetp.v22i4.5133

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 & Education Policy*, *14*(2), 260–285. http://bjsep.org/getfile.php?id=312

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., & Kim, S. (2017). *The Basics of Item Response Theory Using R* (S. E. Fienberg (ed.)). Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8_1

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *19*(1), 388–402. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Birdsall, M. (2011). *Implementing Computer Adaptive Testing to Improve Achievement Opportunities* (Issue April). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606023/0411_MichaelBirdsall_implementing-computer-testing-_Final_April_2011_With_Copyright.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Reading MA: Addison-Wesley.itle. In S. T. of M. T. S. In: Lord, F.M. and Novick, M.R., Eds. (Ed.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

CETAP. (2020). *Test dates National Benchmark Test (NBT)*. https://www.cut.ac.za/nbt

Cliff, A. (2015). The national benchmark test in academic literacy: How might it be used to support teaching in higher education? In *Language Matters* (Vol. 46, Issue 1, pp. 3–21). Routledge. https://doi.org/10.1080/10228195.2015.1027505

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences*. https://doi.org/10.4324/9780203771587

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory* (1st ed.). The Guilford Press.

Dirven, L., Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., Velikova, G., Verdonck-de Leeuw, I. M., Young, T., & Groenvold, M. (2021). Development and Psychometric Evaluation of an Item Bank for Computerized Adaptive Testing of the EORTC Insomnia Dimension in Cancer Patients (EORTC CAT-SL). *Applied Research in Quality of Life*, *16*(2), 827–844. https://doi.org/10.1007/S11482-019-09799-W

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. *Item Response Theory for Psychologists*, 1–371. https://doi.org/10.4324/9781410605269/ITEM-RESPONSE-THEORY-SUSAN-EMBRETSON-STEVEN-REISE

Germain, M. (2006). Stages of psychometric measure development: the example of the generalized expertise measure (GEM). *AHRD ConProceedings*, *25*(4), 893–898.

Han, K. C. T. (2018a). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*, 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*, 20. https://doi.org/10.3352/jeehp.2018.15.20

Huit, W. (2011). Bloom et al.'s taxonomy of the cognitive domain. *Educational Psychology Interactive*, *10*(2), 1–4. http://www.edpsycinteractive.org/topics/cogsys/bloom.html%0Ahttp://www.edpsycinteractive.org/topics/cognition/bloom.html

Kayembe, C., & Nel, D. (2019). *Challenges and Opportunities for Education in the Fourth Industrial Revolution*. *11*(3).

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

Lawless, R., Bejar, I. I., Morley, M. E., Wagner, M. E., & Bennett, R. E. (2002). *A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing*. *December*. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

le Roux, N., & Sebolai, K. (2017). The national benchmark test of quantitative literacy: Does it complement the grade 12 mathematical literacy examination? *South African Journal of Education*, *37*(1), 1–11. https://doi.org/10.15700/saje.v37n1a1350

Lee, M. H., Yun, J. H. J., Pyka, A., Won, D. K., Kodama, F., Schiuma, G., Park, H. S., Jeon, J., Park, K. B., Jung, K. H., Yan, M. R., Lee, S. Y., & Zhao, X. (2018). How to Respond to the Fourth Industrial Revolution, or the Second Information Technology Revolution? Dynamic New Combinations between Technology, Market, and Society through Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity 2018, Vol. 4, Page 21*, *4*(3), 21. https://doi.org/10.3390/JOITMC4030021

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. https://doi.org/10.1348/000711009X474502

Lone, S. A., & Ahmad, A. (2020). COVID-19 pandemic–an African perspective. In *Emerging Microbes and Infections* (Vol. 9, Issue 1, pp. 1300–1308). Taylor & Francis. https://doi.org/10.1080/22221751.2020.1775132

Lord, F. M. (2012). Applications of item response theory to practical testing problems. In *Applications of Item Response Theory To Practical Testing Problems*. Taylor and Francis. https://doi.org/10.4324/9780203056615

Luecht, R., & Sireci, S. (2011). A Review of Models for Computer-Based Testing. *College Board Research Reports*, *1*, 1–56.

Mcclarty, K. L. (2007). A feasibility study of a computerized adaptive test of the international personality item pool NEO. In *Dissertation Abstracts International: Section B: The Sciences and Engineering* (Vol. 67). http://ezproxy.lib.ucf.edu/login?URL=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2007-99012-060&site=ehost-live

NBT. (2022). *More about the NBTs | National Benchmark Test Project*. https://www.nbt.ac.za/content/about

Oladele, J. I, & Ndlovu, M. (2021). A review of standardised assessment development procedure and algorithms for computer adaptive testing: Applications and relevance for fourth industrial revolution. *International Journal of Learning, Teaching and Educational Research*, *20*(5), 1–17. https://doi.org/10.26803/ijlter.20.5.1

Oladele, Jumoke Iyabode, Ayanwale, M. A., & Owolabi, H. O. (2020). Paradigm Shifts in Computer Adaptive

Testing in Nigeria in Terms of Simulated Evidences. *Journal of Social Science*, *63*, 9–20. https://doi.org/10.31901/24566608.2020/63.1-3.2264

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*(350), 351–356. https://doi.org/10.1080/01621459.1975.10479871

Perez, J.A; Deligianni, F.; Ravi, D.; Guang-Zhong, Y. (2018). Artificial intelligence and robotic assembly. *Engineering with Computers*, *2*(3), 147–155. https://doi.org/10.1007/BF01201262

Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., Hjermstad, M. J., Kaasa, S., Loge, J. H., Velikova, G., Young, T., & Groenvold, M. (2016). Development of an item bank for computerized adaptive test (CAT) measurement of pain. *Quality of Life Research*, *25*(1), 1–11. https://doi.org/10.1007/S11136-015-1069-5

Pommerich, M., Segall, D.O.; Moreno, K. E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. *In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved on April 20, 2022 from www.psych.umn.edu/psylabs/CATCentral/

Prince, R., & Frith, V. (2017). The quantitative literacy of South African school-leavers who qualify for higher education. *Pythagoras*, *38*(1). https://doi.org/10.4102/pythagoras.v38i1.355

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* (Issue May). www.nbt.ac.za

Radmehr, F., & Drake, M. (2019). Revised Bloom's taxonomy and major theories and frameworks that influence the teaching, learning, and assessment of mathematics: a comparison. *International Journal of Mathematical Education in Science and Technology*, *50*(6), 895–920. https://doi.org/10.1080/0020739X.2018.1549336

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Redecker, C., & Johannessen, Ø. (2013). Changing Assessment - Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, *48*(1), 79–96. https://doi.org/10.1111/ejed.12018

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/ANNUREV.CLINPSY.032408.153553

Schwab, K. (2016). *The Fourth Industrial Revolution.* World Economic Forum.

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

Skilton, M., & Hovsepian, F. (2018). *Responding to the Impact of Artificial Intelligence on Business*. *April 2019*, 342. https://eclass.hmu.gr/modules/document/file.php/ECE113/Χρήσιμο Υλικό %26 Παρουσιάσεις/eBook - The 4th Industrial Revolution/2018_Book_The 4th Industrial Revolution.pdf

Smarter Assessment. (2022). *What is Item Banking? What are item banks?* -. https://assess.com/item-banking-can-improve-assessments/

Steffen, G. N. M. ; M. (2016). *Computerized Adaptive Testing: Theory and Practice* (W. J. van der; G. A. W. G. Linden (ed.)). Kluwer Academic.

Steward, K. (2020). SARS-CoV-2 Is Re-emerging Following the Relaxation of Lockdown. *Immunology & Microbiology*, *5*(2), 1–5.

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

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

van der Linden, W. J., & Ren, H. (2020). A Fast and Simple Algorithm for Bayesian Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58–85. https://doi.org/10.3102/1076998619858970

Veldkamp, B. P., & Verschoor, A. J. (2019). *Robust Computerized Adaptive Testing*. 291–305. 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

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. and Charman, M. (2009). Principles and practice of on-demand testing. In *Principles and practice of on-demand testing* (Vol. 14, Issue 1, pp. 67–85). www.ofqual.gov.uk/files/2009-01-principles-practice-on-demand-testing.pdf

WHO. (2022). *Africa on track to control COVID-19 pandemic in 2022 | WHO | Regional Office for Africa*. https://www.afro.who.int/news/africa-track-control-covid-19-pandemic-2022

Worldometer. (2022). *COVID Live - Coronavirus Statistics - Worldometer*. Worldometer. https://www.worldometers.info/coronavirus/#countries

Xinhu. (2022). *Africa's COVID-19 cases near 11.34 mln: Africa CDC-Xinhua*. http://www.news.cn/english/20220402/9de2ab8f88bc44f6b28b8b18e125752e/c.html

Xu, M., David, J. M., & Kim, S. H. (2018). The Fourth Industrial Revolution: Opportunities and Challenges. *International Journal of Financial Research*, *9*(2), 90. https://doi.org/10.5430/IJFR.V9N2P90

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). https://doi.org/10.3389/FPSYG.2019.01010/Full