International Journal of Learning, Teaching and Educational Research
Vol. 20, No. 5, pp. 1-17, May 2021
https://doi.org/10.26803/ijlter.20.5.1

A Review of Standardised Assessment Development Procedure and Algorithms for Computer Adaptive Testing: Applications and Relevance for Fourth Industrial Revolution

Jumoke I. Oladele and Mdutshekelwa Ndlovu
University of Johannesburg, Johannesburg, South Africa
https://orcid.org/0000-0003-0225-7435
https://orcid.org/0000-0002-1187-0875

Abstract. Teaching and learning have gone online in response to the pandemic, which reveals the need for accurately tailored educational assessments to ascertain the extent to which learning outcomes or objectives are achieved. Computer Adaptive Testing (CAT) is a technology-driven form of assessment that tailors items to a candidate's ability level with empirically proven benefits over the fixed-form computer based test. A systematic review was employed which shows that item bank is a key requirement for CAT and the items must through a rigorous item development process to ensure and maintain quality in terms of content, criterion constructs and internal consistency, determining the psychometric validation of behavioural measures while leveraging on variances of Item Response Theory (IRT). Following the item development stage is the need to compile validated items into administrable forms using advanced computer software for automatic test assembly and administration, such as FastTest which allows specifying empirically tried algorithms for CAT from start to termination of the test. This helps to ensure that assessment properly leverages the advantages that CAT holds. Furthermore, the review revealed that CAT has been widely applied with large-scale testing in various fields by educational, health and psychological professionals utilising different IRT models; however only in developed countries. This brings to bear the need for adoption in other parts of the world, for improvements in educational assessments. The interjections of 4IR with AI considering emerging technology aids the CAT algorithm for achieving expert and knowledge-based systems, being a requirement for survival in today's world.

Keywords: item bank development; test-forms; administration algorithms; CAT; 4IR

©Authors
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

The world is battling with Covid-19 pandemic which has impacted the continents in unimaginable ways. First reported in Wuhan City, Hubei Province, China, on December 31, 2019, the virus has spread like wildfire worldwide with 106,673,989 recorded cases and a death toll of 2,326,773 as of GMT 01.31 on February 8, 2021 (Worldometers, 2021). Statistics show that the virus has spread into 58 African countries, having over four million recorded cases and a death toll of one hundred and twenty-two thousand, one hundred and three, (122,103); 53,757 of which was recorded in South Africa as reported on April 19, 2021 (COVID-19 South African Online Portal, 2021; APO Group, 2021). With over 1.602 m, South Africa remains the worst-hit African country with about 34% of the continent's recorded cases and 44% of its death toll (Worldometers, 2021). Efforts to flatten the curve in South Africa necessitated the adoption of a five-level lockdown approach starting from Level 5 in March 2020 with severe measures to curb the spread of the virus. The country moved gradually down to (adjusted) Level 1 by September 2020 whereby day-to-day activity could recommence, while adhering strictly to Covid-19 safety guidelines (The Presidency, Republic of South Africa, 2020; South African Government Disaster Management Act, 2020 Staff Writer, 2020). The ease of lockdown restrictions that started in May 2020, despite the rapid rise of Covid-19 cases, by South Africa's President was necessitated to salvage the country's deteriorating economic situation, as experienced in other parts of the world (BBC News, 2020; Vecchiatto *et al.*, 2020).

The pandemic has resulted in national and international lockdowns to curb the spread of the virus. The lockdown has impacted the educational sector just like all other sectors of the economy with learning having gone virtual in most technologically advanced countries of the world applicable to higher learning institutions (Li & Lalani, 2020; The World Bank Group, 2020a, 2020 b). With virtual learning comes the need for virtual assessments, also known as off-site assessments, which requires the use of high-end technology, such as Computer-Based Testing (CBT). CBT is a method of administering tests where examinations are deployed through a computer terminal, and the responses are recorded and assessed electronically, which can be fixed-form or adaptive (Alabi *et al.*, 2012). A fixed-form CBT is an examination in which the computer presents all items to candidates regardless of their performance, usually presented from the easiest to most difficult items with a limited number of parallel forms (Alabi *et al.*, 2012; Becker & Bergstrom, 2013, Oladele *et al.*, 2020). The adaptive CBT, also known as Computer Adaptive Testing (CAT), is a testing procedure that employs on-the-fly techniques aligned to candidates' ability levels to enhance the accuracy of testing while reducing test length by up to 50% (Han, 2018; Kimura, 2017, Reckase, 2010). As such, examinees are served items according to their ability levels (difficult/easier), thus guaranteeing a personalised assessment format (Aybek & Demirtasli, 2017). With CAT, a large bank of administrable test items categorised by content, difficulty and parallel forms is required. This review centres on item development, test-forms and CAT algorithms while considering its broad applications and relevance for the Fourth Industrial Revolution. The limitation of the review was identified while giving directions for future research.

2. Item Bank Development for CAT

Item bank for CAT is a collection of calibrated test items based on the parameters of difficulty, discrimination and pseudo guessing having gone through rigorous item development procedures while indicating the history of the items developed. Also, an initial CAT item bank could start with existing paper-and-pencil items while adding new ones, which guarantees that items maintain their psychometric properties while impacting the cost implication of an additional number of items (Linacre, 2000; Thompson & Weiss, 2011). Germain (2006) stressed that quality in item development must be valid in terms of content, criterion constructs and internally consistent, determining the psychometric validation of behavioural measures such as a test. A typical test consists of items carefully developed to ensure that the test is valid, testing what it purposes to test and reliable; that is, tests are consistent over multiple administrations. Preliminary activities for item banking are discussed below:

Planning leads to decision-making: planning is necessary before drawing an item bank for CAT, which is premised on a range of decisions arrived at based on the test's purpose. Planning entails determining the test's objectives for curriculum evaluation, students' motivation, placement and selection, remedial work diagnosis, and formative and summative evaluation. Also, it is necessary to consider the likely decisions based on the test results. Another major decision for planning item bank development rests on the available resources considered regarding resources needed in the test development, such as expertise and personnel as well as the technology needed for test administration (Cella *et al.*, 2007).

Content analysis and test blueprint: provides a summary of curricular objectives designed by a subject-specific expert in selecting testing domains. The content is supposed to provide the learning experiences that will enable the test to achieve its stated objectives. This analysis of the content helps the test planner determine the relative importance of the content's various aspects and the emphasis on the specifics. Based on this, a test blueprint, also known as Table Of Specification (TOS), can be constructed.

As provided by Bloom's Taxonomy, a TOS aligns with test content rather than the curriculum content and, as such, the latter may be narrower than the former in scope (Anderson & Krathwohl, 2001). It is a practical word given to the plan for scripting items for a test. TOS is a two-dimensional table relating instructional objectives to course content and specifying what proportions of these are to be sampled by the test items. The table of specification enables test experts to gauge examinees over knowledge (cognitive), skill (psychomotor) and attitude (affective) depending on the domain of testing interest. It provides the operational guides to ensure that a test addresses what it sets out to address.
The preparation of a table of specification requires:

1. The total number of items that will constitute the test. It is important to note that a large item bank is required with adaptive testing and this should have been adequately catered for at the planning stage in terms of expertise as well as personnel engagements; and

2. The proportion of items developed per content areas, depending on the emphasis placed on it during instruction and the amount of time spent, as illustrated below drawn from topics in a statistics course:
   - ❖ Frequency Distribution 10%
   - ❖ Measure of Central Tendency 15%
   - ❖ Measure of Variability 15%
   - ❖ Measure of Relationship 40%
   - ❖ Relative Standing 20%
3. Deciding on the proportion of items in each process objective depending on the level of the cognitive behavioural objectives. The illustration is as follows:
   - ❖ Remembering (Recall of facts from short-term memory) 20%
   - ❖ Understanding (Recovering appropriate knowledge from long-term memory) 30%
   - ❖ Applying (Using a procedure in a given situation) 15%
   - ❖ Analysing (Breaking instruction into its constituent parts, how it relates to one to another and to an overall structure) 15%
   - ❖ Evaluating (Making judgments based on criteria and standards) 10%
   - ❖ Creating (Placing components of instruction together to form a new, lucid whole) 10%
4. Deciding on the quantity of test items to be constructed in each of the content areas by finding out the respective percentage of the total number of items (see Table 1).
5. Deciding on the quantity of test items to be written in each content area of the cognitive behavioural objectives (see Table 1).

It is good to ensure that the sum of the approximated numbers of items should be equal to the total number of the items desired in each of the content areas. This is shown in Table 1.

Table 1: Test Blueprint for a Test in Statistical Methods

	Cognitive Learning Domains						
Content Areas	*Remembering* **(20%)**	*Understanding* **(30%)**	*Application* **(15%)**	*Analysis* **(15%)** *Evaluating* **(10%)**		*Creating* **(10%)**	**Total**
A (10%)	1	1	1	1	1	1	**6**
B (15%)	2	3	1	1	1	1	**9**
C (15%)	2	3	1	1	1	1	**9**
D (40%)	5	7	4	4	2	2	**24**
E (20%)	2	4	2	2	1	1	**12**
Total	**12**	**18**	**9**	**9**	**6**	**6**	**60**

A: Frequency Distribution; B: Measure of Central Tendency; C: Measure of Variability; D: Measure of Relationship; E: Relative Standing (Revision of Bloom's Taxonomy of Educational Objectives)

Table 1 provides a framework for organising information about the students' instructional activities (Anderson & Krathwohl, 2001). The foundation of the practice of educational assessment is the extent to which students' learning outcomes are achieved guided be a table of specification when writing test items, especially with standardised tests.

Item writing for CAT: this is activity-centred which entails preparing assessment tasks for gauging students' knowledge and skill gained from exposure to teaching and learning. It is required that assessment tasks be precise and aligned to learning objectives important for CAT leveraged on using the item information function in terms of difficulty, discrimination and guessing (Veldkamp & Verschoor, 2019). As such, professionalism is required for item writing, which is germane to the effectiveness of CAT.

Steps identified for CAT item writing procedure were literature search, formulation of new items or acquiring items from existing test forms where available, field-testing conducted through a computer terminal, and psychometric analyses for the final items selection (Cella *et al.*, 2007; Petersen *et al.*, 2016). Expert evaluations should be carried out to ascertain face and content validation leading to field testing. Thompson (2018) also outlined a four-step procedure for item writing; however, it uses tailor-fit software. The first stage consists of feasibility and planning studies using CATSim, and this precedes the item bank development using FastTest, a comprehensive assessment ecosystem (Thompson, n.d.). In the third stage, items are pilot tested using FastTest while, at the fourth stage, item analysis is performed and other due diligence using Iteman or Xcalibre. Xcalibre provides item response theory calibration for a wide range of assessment types, using all the major dichotomous and polytomous models. Its unique features allow for automatic report generation, with full result tables and figures (item response functions and standard error functions) already embedded. While reiterating that CAT is not easy, the goal is to ease the task using clean software with no need for code writing while aligning with best practices and international standards.

Zhang *et al.* (2019) developed CAT to assess internet addiction while investigating related validity issues. The standardised scales used had a total of 59 carefully calibrated polytomous scored items and satisfying the IRT assumptions of unidimensionality, as well as a good item-model fit. Also, items did not function differentially. According to Downing (2006), specialised training on item writing is as important as content knowledge. Quality item writing skills can be ascertained by constant practice and critical reviews from experts (Jozefowicz *et al.*, 2002).

Item Review: High expertise is required in writing test items after which item review is mandatory. Item review ensures clarity to all and gives evidence about the quality of the items carried out by test and subject experts for content

quality; items void of ambiguity; identifying unintended clues to the correct answer; items with no correct or multiple answers plausible distracters; language difficulties; redundant words; grammatical faults; and sensitivity to issues that could bring bias to the test, such as cultural and gender among others. Item review is essential in the item development process leading to the empirical trial, also known as face validity, which is geared towards ascertaining that a test measures only the intended.

The gold standard is to have independent subject experts' evaluatively correlate test items with instructional objectives and blueprint (National Research Council-NRC, 2004). Izrad (2005) reiterated that item review is central to the test development process, assessment approaches and curriculum intentions. The author also stressed the benefits of having a team of item reviewers as it provides the benefit of interaction with colleagues to avoid the possibility of idiosyncrasy and limited view of the topic to be assessed. As such, selecting an appropriate sample of evidence will foster accuracy on decisions made from educational assessments and enhance efficiency. Other relevant aspects of the item review process are items scoring, availability of practice items, need for separate answer sheet, the appropriate time for the actual test, score key to be used as errors in score keys will create interpretation problems and test administration information to be provided at the trial test stage (Izrad, 2005).

Trial/Pilot Testing: this is a means of subjecting proposed test items to testing with a comparable group of students to the target group as a selection criterion. Data generated from this exercise are used to assess test item quality based on the item parameters of difficulty (denoted by b), discrimination (denoted by a) and guessing (denoted by c) in alignment with the test model fit (Oladele *et al.*, 2020). Pilot testing in the item development process is essential before use with the target group and requires sound planning concerning gender, age and schooling level group required for the trials and administration modes. Generally, Izard (2005) explained that trial testing would help establish item parameters, the appropriate number of final test items, ascertain the administration instructions' adequacy and if practice items would be required, adequacy of testing time and students' responses pattern analysis.

Item Analysis: students' responses are analysed using a variety of methods. It is a systematic evaluation of the effectiveness of each test item. Zhang *et al.* (2019) explained that developing an item bank for CAT requires evaluation for ascertaining unidimensional assumption of the item pool, a measure of only the main latent trait; selecting the test IRT model-fit, assessing local independence of the item pool for ensuring that within and across examinee response on an item will not be influenced by other test items; assessing item pool monotonicity, connoting that examinees with higher latent trait levels have a probability of higher scores and that items functions at par for examinees who are of the same ability level, also known as Differential Item functioning (DIF) (Aybek & Demirtasli, 2017). According to Izard, 2005, item analysis is aimed at determining:

Item difficulty: this index is a function of the skill level required by items administered to a particular group and reported for a particular test. Therefore, item difficulty is a measure of the proportion of examinees that answers an item correctly, and so it is a direct function of examinees' ability level. An achievement test aims to have at least 90% of students completing all the items unless the purpose is to test speed.
Item discrimination power: is the correlation between the item responses and correct responses. It is a measure of how a single item separates high from low ability level examinees. At worst, items analysis aids the identification and deletion of items that do not fulfil this role and at best calls for necessary amendment.

Pseudo Guessing: this connotes that examinees with very low ability levels have some probability of answering one item correctly. For example, an examinee with no requisite knowledge on a multiple-choice item with four options still has a 25% chance to answer it correctly, based on guessing.

There are various models in testing, and the IRT model is commonly used with CAT. IRT models the relationship between examinees' performance on the test of their ability levels. It is a theory that focuses on the item level of performance. As such, IRT models examinees' performance at each ability level to each item on the test. Standard unidimensional models are the one-parameter logistic (1PL) model (difficulty parameter- b), the two-parameter logistic (2PL) (difficult and the discrimination parameters- b, a) and the three-parameter logistic (3PL) model (pseudo guess parameter- c to b and a)(Aybek & Demirtasli, 2017; Oladele *et al.*, 2020). Some benefits IRT brings to the educational testing table include putting the examinees and items on the same scale: sample independent score equating enables score correspondence between two tests expressed as the item's characteristic curves; examinee specific Standard Error of Measurement (SEM) is based on individual ability levels computed as a reciprocal of the test information across different ability levels. So the more information a test provides at an ability level, the lesser the SEM and the features of examinees and item on the scale enables the selection of items that provide full information for examinees at theta ability level on which CAT rests as an advanced passing scheme (Wang & Thompson, 2020).

Theoretical Models for CAT Item Analysis

Items analysis for CAT can be approached with dichotomous Item Response Theory (IRT) deployed as the one, two or three-parameter logistic model (Oladele *et al.*, 2020). With the one-parameter logistic model, (Rasch model), the probability of getting an item (i) correct at an ability level (θ) is expressed as:

$$P_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}} \qquad \text{Eq. 1}$$

e in the equation is an exponential constant with the value 2.718 (approximately) while D is a scaling factor with the value 1.7 regarded as the "normal metric". However, the common practice is to set D to 1.0 as the "logistic metric" since the normal ogive model is seldom used in real testing situations. Considering D's

value is paramount when studying or generating item parameters to ensure that items are tailored to provide maximal information in examining examinee proficiency using the IRT modelled unidimensionally and determinant of the response model adopted (Wise & Kingsbury, 2000).

With the 2PLM, each item has its discrimination parameter denoted as a in the equation as against fixing as '1' across all items as practised with 1PLM. Thus, the model is mathematically expressed as:

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \qquad \text{Eq. 2}$$

Lastly, the 3PLM allows an Item Characteristic Curve to have non-zero lower asymptotes; this is suitable for response data with high likelihood for guessing, such as multiple-choice items, and is expressed as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1+e^{-Da_i(\theta-b_i)}} \qquad \text{Eq. 3}$$

The ci is the pseudo-guessing parameter which signifies the likelihood of low ability examinees responding to item i correctly. Segall (2005) stressed that the 3PL is commonly used to model multiple-choice items scored dichotomously. Using any of the IRT models permits the comparison of the examinees' ability level based on item parameters used to estimate the probability of the individual's response to that item (Aybek & Demirtasli, 2017). For CAT, employing IRT models, only suitable examinee ability level is selected from the item pool in an iterative cycle from item selection to stopping criteria until the individual's θ level is estimated accurately. A large items pool is required for each theta level with uniformly distributed difficult, highly discriminating items and low guessing parameters, which can provide greater measurement efficiency (Segall, 2005; Wainer *et al.*, 2000). Carrying out item analysis concurrently with test administration is highly recommended (Linacre, 2000).

3. Test Forms and Administration

The next activity is to assemble the test into administrable forms, having carefully undergone item writing. Through reviews to trial testing and analysis, the test assembly process also impacts the final test's validity and score interpretation for effective quality control (Izard, 2005). Test administration is then carried out publicly in the test development process, as a way of striking a balance between theory and practice (McCallin, 2006). Therefore, there is a positive correlation between the standardisation of testing conditions and test administration quality. Proctoring is also germane with off-site testing to curb examination malpractices (Downing, 2006).

According to Wise and Kingsbury (2000), a CAT administration is a two-stage process. At the first stage, a test item with an average level of difficulty is administered after which the response is scored, and this forms the basis for the next item selected. At the second stage, candidates' responses are scored leading to updating the examinees' proficiency level. These stages go through a cyclic pattern until some stopping criterion is met for a fixed or variable-length test

(predetermined number of items and a desired level of measurement precision, respectively). The CAT algorithm follows an iterative process until the test converges on a final proficiency estimate for a candidate. The author stressed that, while adaptive testing administration is relatively simple in theory, the practice is more complicated; as discussed under the CAT algorithm section of this literature review.

Administered as a CAT, advanced computer software for automatic test assembly such as FastTest is a tried-and-tested enterprise platform for high stakes assessment which leverages on Artificial Intelligence for creating equivalent test forms (Luecht, 2006; Thompson, n.d., 2018). Worthy of note is that such technologies can provide greater ease for the test construction and administration process while also enhancing the items' quality.

4. Algorithm for CAT

CAT has mainly been tagged with advantages that give it an edge over the fixed-form of the test. Some of these advantages are shorter tests leading to reduce testing time by 50% or more, equi-precision, examinee experience as CAT provides appropriate challenges for each examinee leading to increased motivation, and much greater security with item-set specific test administration, all these being possible by leveraging on computer technologies (Thompson, 2011). While the advantages of CAT are appealing, there are strict procedures that must be adhered to in ensuring that assessment properly leverages the advantages that CAT brings to the table.

Starting point/Item selection: the starting point for CAT is taken as given, it could be based on fixed values, randomly chosen values within a range or mean items parameters. Using a pre-defined IRT models, candidates' previous response determines item selection from a calibrated bank, which is usually large. With CAT, for candidates a small number of ability-appropriate items is required for accurate ability compared to the fixed-form test, which presents the full length of items to all candidates (Cella *et al.*, 2007). Cella et al. (2007) further stressed that initial item selecting should cover as much of the concept's continuum being measured as possible. Item selection has three significant components, which are item selection, item exposure control and content balancing (Han, 2018). Item selection is driven by item information, with a preference for the most appropriate items. CAT administration combines item selection and ability estimation concurrently with little or no human intervention as the test adapts to the examinee's ability level (van der Linden & Pashley, 2009). For example, administering easy items to a high ability level examinee makes no sense with passing guaranteed, and vice-versa (Eggen, 1999; Eggen & Straetmans, 2000; Thompson, 2009).

Score estimation: with CAT, psychometricians must select on the initial, interim and final score estimates methods (van der Linden & Pashley, 2009). Some modern score estimation methods are Maximum Likelihood Estimation (MLE), Maximum Likelihood Estimation with Fences (MLEF), Bayesian Maximum a Posteriori (MAP) and Bayes Expected a Posteriori (EAP) (Han, 2018). At the initial stage, Bayesian methods are advised over maximum likelihood estimates

for a dichotomously scored test as it is not capable of producing accurate estimates with related response patterns. At the interim score estimation stage, it is expected that ability estimates converge quickly, which is satisfied by an appropriate combination of ability estimator and item-selection criterion; a popular choice being the EAP estimator combined with maximum-information item selection. At the final stage of CAT, the goal is to provide the candidate with an accurate estimate of their performance. Performance estimation can be resolved using Bayesian methods such as EAP estimator, which aids in fixing the ability estimate until accurate estimates are obtained. This shows that scores estimation methods should be made carefully considering that these methods all have their drawbacks. Other aspects that impact score estimates are the quality of item pool, use of candidates' collateral information, issues concerning on item exposure, and item selection constraints (van der Linden & Pashley, 2009).

Van der Linden (2005) explained that, with CAT, the candidate's ability estimate is updated after each new response, leading to the next item selected based on the full information of the updated estimate. Although there are a variety of item selection methods, Han (as cited in Oladele *et al.* 2020) stressed that modern methods requiring less computer time are the Maximised Fisher information, the b-matching, a-stratification with or without b-blocking, Kullback-Leibler information, weighted likelihood information, and efficiency balanced information. The choice of an item selection method with the highest measurement precision is crucial to the assessment process. Adopted the Monte-Carlo simulation approach for CAT feasibility studies is necessary to determine the viability of method selection. Oladele *et al.* (2020) reported a-Stratification/b-Blocking an empirically proven method for CAT leading to accurate ability placement.

Termination criterion: algorithms for CAT should be specified as fixed-length where all candidates receive an equal number of items adaptively selected from the bank or variable-length tests items and needed number are adapted to the candidate. Termination criteria could be the candidate ability level (θ estimate), standard error of measurement (SEM) or item bank to be evaluated. While the first two methods are premised on the candidates' parameter, the third method is premised on item parameters. CAT is terminated when the ability estimate stops varying significantly by administering additional items and which hinges on the minimum information criterion. Therefore, a test terminates when there are no items left in the bank, which provides a minimal level of information, as specified by the item selection algorithm. Most utilised of these termination criteria is SEM (Thompson & Weiss, 2011).

With a carefully calibrated item bank in place with the appropriate technological integrations, a simple CAT begins by presenting an item with average difficulty to a candidate as practised using the maximum likelihood approach to item-choice early in the Adaptive Sequence (Segall, 2005). Although the starting point of the test may not be critical to measurement, it could impact the psychological state of the candidate wherein administering an item with high difficulty may immediately lead candidates into despair while administering an item with low difficulty may result in the candidate not taking the test seriously and so making

careless mistakes (Linacre, 2000). IRT is clearly at the heart of CAT in which modern algorithms concepts are taken from and maximum likelihood and Bayesian statistical estimation theories.

It is pertinent for test development to be based on a sound scientific basis and that evidence of the scientific approach should be documented (NRC, 2004). Downing (2006) reiterated that, although creating useful test items is greatly improved by constant engagement, there are well-established item writing ethics. This goes a long way to ensure the writing of cognitive appropriate items in the process of test development. Psychometric testing is broad in its potential application cognitive and non-cognitive measures. The outlined stages provide an appropriate organisational structure for validating a testing program and maintaining relevant educational testing standards (Downing, 2006).

5. Applications of CAT

CAT has been extensively applied in various fields by educational, health and psychological professionals utilising different IRT models with over four decades of practice. Weiss and Kingsbury (1984) examined the application of CAT to educational problems, which were Adaptive Mastery Testing (AMT) using the 1, 2 and 3PL models in a simulated study to compare the average items used to reach a mastery/non-mastery decision for the conventional and adaptive AMT procedures. Findings revealed that the adaptive test results in higher ability estimation precision than fixed-form tests with fewer items.

Eggen and Straetmans (2000) employed CAT for classifying candidates through simulation studies. Computation procedures used were based on statistical estimation and statistical testing with five item selection methods (Maximum Information (MI) at the candidate's current ability estimate, MI with content control, MI with exposure control and MI with both content and exposure controls). The effects of adding content and item exposure control based on the 1PL model were also investigated, and real data from a mathematics placement test for adult learners were used. Findings revealed that the item bank's quality is satisfactory for adaptive testing with a maximum of 25 items for each test administration, reducing the number of required items to between 22-44% of the required number with paper-and-pencil versions.

Ware Jr. *et al.* (2003) applied CAT to assess the impact of pain as a simulated study using real data to select the most informative items for each candidate and estimate impact scores according to pre-set precision standards. Findings revealed that adaptive-based administrations impacted achievement without compromising testing validity over time. Also, Kane *et al.* (2020) and Theunissen *et al.* (2020) applied CAT in developing more concise Patient-Reported Outcome Measures (PROM) using the Veterans RAND 12 Item Health Survey (VR-12) deployed adaptively to decrease patients' question burden, a 33% decrease. Therefore, the CAT model was termed efficient in improving PROM as well as patient experience.

CAT has also been applied to large-scale language testing programmes for placement purposes such as The Quick Placement Test (QPT), Test of English as

a Foreign Language (TOEFL), Computerised Oral Proficiency Instrument (COPI) and Basic English Skills Test (BEST) Plus. Others are Scholarship Aptitude Test (SAT), the Test of Standard Written English, Student Description Questionnaire, and ATP Achievement Tests, all under the College Board Admissions Testing Program (ATP), Graduate Records Examination (GRE) and Graduate Management Admission Test (GMAT) (Cella *et al.*, 2007; Giouroglou & Economides, 2004). These testing programmes are full-scale paper-and-pencil testing before being implemented adaptively (Eignor *et al.*, 1993). Way *et al.* (2006) examined practical questions needed to be adequately answered before the transition of testing programmes to online delivery forms using CAT concerning K-12Assessments. CAT has also been implemented extensively in licensing health professionals in the United States, such as the National Council Licensure Examination-Registered Nurses and the National Registry of Emergency Medical Technicians (Han, 2018; Seo, 2017).

6. Fourth Industrial Revolution and CAT

The Fourth Industrial Revolution (4IR) occupies a digital sphere driven by the merging of technologies that makes it almost impossible to distinguish between the physical, digital and biological divides. Some of the possibilities brought to bear by the 4IR include the ease of connecting people by mobile devices, with high processing power, large storage volume and a knowledge economy, rapidly influencing intelligent behaviour in living and systemic engineering multiplied exponentially by evolving expertise in many fields, one of which is Artificial Intelligence (AI) (Bartneck *et al.*, 2021; Schwab 2016; Singh et al., 2013). AI is a wide-ranging branch of computer science premised on smart technologies capable of performing human intelligence-based tasks. It adopts an interdisciplinary approach, creating a paradigm shift in virtually every tech industry sector (Builtin, 2019).

4IR riding on AI drives possibilities which are fast turning into realities, with strong indication that the technologies underpinning the 4IR have a significant impact on businesses (Schwab, 2016); and the educational sector cannot be left out. As such, the sectoral response to 4IR must be unified and inclusive of all global stakeholders, such as the public and private sectors, academia and civil societies. The CAT algorithm leverages AI to achieve expert and knowledge-based systems for accurate ability placement. The possibilities that 4IR brings to the table, such as multiple connectivity through a high-ended computer device and high storage capacity, strengthens CAT technology for educational assessments. These possibilities are coupled with emerging technology breakthroughs premised on AI that could be leveraged for educational testing (Schwab, 2016). Butler-Adam (2018) challenged educational researchers to identify the link of AI to curricula, teaching and learning while stressing the need for people to have the skills required to thrive with evolving technology, and be more of problem solvers, being adaptable and adequate in expressing themselves in both the written and spoken word. These are achievable by accurate educational assessments through sophisticated algorithms for adaptive testing with CAT, an emerging technology-driven by 4IR. Applications of AI for educational assessment hold the potential of shaping higher education with exponential technologies such as CAT (Penprase, 2018).

7. Conclusion and Recommendation

The review has shown an exciting direction for ensuring accuracy in ability estimation premised on the 3Parameter Logistic Item Response Theory model made possible by the fourth industrial revolution and characterised by high-ended adaptive technologies such as CAT in the realm of AI and apparent with intelligent candidate ability estimation in an iterative process. Empirical studies have provided evidence of reduced test lengths with CAT without watering down score integrity. Lessons from the Covid-19 pandemic show the need for higher institutions of learning to have online arrangements for teaching and learning with suitable assessment platforms for accurate ability estimation while leveraging on the wide technological possibilities greatly enhanced by the 4IR. The onus lies on higher educational stakeholders in Africa to position technology for pedagogical gain. While the traditional linear tests have mainly been employed, it is imperative to ensure CAT practice in educational assessment in Africa to actualise its gains at the higher levels of education with most institutions moving to online teaching and learning. CAT for supporting teaching and learning is an undeniable reality in the Covid-19 era and a way of ensuring that the African continent falls in line with the rapid emerging technologies characteristic of the 4IR era.

8. Limitations and directions for future research

The review carried out is majorly premised on foreign literature as CAT as an assessment format is an emerging area of research in Africa. The workability of CAT considering the realities of the African continent should be considered. This calls for simulation studies on CAT as a direction for future research.

9. References

Alabi, A. T., Issa, A. O., & Oyekunle, R. A. (2012). The use of a computer-based testing method for the conduct of examinations at the University of *Ilorin. International Journal of Learning and Development, 2*(3), 68-80.

Anderson, L., & Krathwohl, D.A. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Longman.

APO Group. (2021). April 19. Coronavirus - South Africa: COVID-19 update 2021). africanews. https://www.africanews.com/2021/04/20/coronavirus-south-africa covid-19-update-19-april-2021//

Aybek, E. C., & Demirtasli, R. N. (2017). Computerised adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science (IJRES), 3*(2), 475-487. https://doi.org/10.21890/ijres.327907

BBC News. 2020. Coronavirus in South Africa: Restrictions ease as Covid-19 cases rise rapidly. https://www.bbc.com/news/world-africa-53093832

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). What Is AI? In C.Bartneck, C. Lütge, A. Wagner & S. Welsh (Eds.), *An Introduction to Ethics in Robotics and AI.* Cham: Springer. https://doi.org/10.1007/978-3-030-51110-4_2

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research, and Evaluation, 18*(1), 14.

Builtin. (2019). Artificial Intelligence. What is Artificial Intelligence? How Does AI Work? https://builtin.com/artificial-intelligence

Butler-Adam, J. (2018). The fourth industrial revolution and education. *South African Journal of Science, 114*(5-6), 1-1. http://dx.doi.org/10.17159/sajs.2018/a0271

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerised adaptive assessment. *Quality of Life Research, 16*(1), 133-141.

COVID-19 Online Resource & News Portal. (2020). COVID-19 Corona Virus South African Resource Portal. https://sacoronavirus.co.za/

Downing, S. M. (2006). Twelve Steps for Effective Test Development. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development.* Routledge. http://www.danangtimes.vn/Portals/0/Docs/314154729-0805852654.pdf#page=18

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerised adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). Case Studies in Computer Adaptive Test Design Through Simulation. *ETS Research Report Series, (2),* 1-41.

Germain, M. L. (2006a). *Development and preliminary validation of a psychometric measure of expertise: The generalized expertise measure (GEM)* [Dissertation, Barry University, Florida].

Giouroglou, H., & Economides, A. (2004). *State-of-the-Art and adaptive open-closed items in adaptive foreign language assessment*. Proceedings of the 4th Hellenic Conference of the Association of Informational and Communication Technologies in Education (pp. 747-756). Athens.

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

Izard, J. (2005). *Trial testing and item analysis in test construction. International Institute for Educational Planning/UNESCO.* http://www.sacmeq.org/sites/default/files/sacmeq/training-modules/sacmeq-training-module-7.pdf

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, H. (2002). *The quality of in-house medical school examinations. Academic Medicine, 77,* 156–161.

Kane, L. T., Namdari, S., Plummer, O. R., Beredjiklian, P., Vaccaro, A., & Abboud, J. A. (2020). Use of Computerized Adaptive Testing to Develop More Concise Patient-Reported Outcome Measures. *JBJS Open Access, 5*(1). http://dx.doi.org/10.2106/JBJS.OA.19.00052

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

Li, C., & Lalani, F. (2020, April 29). The COVID-19 Pandemic has changed education forever. This is how. World Economic Forum. https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (No. 69, p. 58). Mathematics, Engineering, Science, Achievement (MESA) memorandum. https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf

Luecht, R. M. (2006). Designing Tests for Pass-Fail Decisions Using Item Response Theory. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.) *Handbook of test*

development. Routledge. http://www.danangtimes.vn/Portals/0/Docs/314154729-0805852654.pdf#page=18

McCallin, R. C. (2006). Test Administration. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.) *Handbook of test development*. Routledge. http://www.danangtimes.vn/Portals/0/Docs/314154729-0805852654.pdf#page=18

National Research Council. (2004). "3 *The Test Development Process*." Redesigning the U.S. Naturalization Tests: Interim Report. The National Academies Press. https://doi.org/10.17226/11168

Oladele, J. I., Ayanwale, M. A., & Owolabi, H. O. (2020). Paradigm Shifts in Computer Adaptive Testing in Nigeria in Terms of Simulated Evidence. *Journal of Social Sciences, 63*(1-3, 9-20. Publication of Kamla-Raj Enterprises (KRE) Publishers. https://doi.org/10.31901/24566756.2020/63.1-3.2264

Penprase, B. E. (2018). The Fourth Industrial Revolution and Higher Education. In N. Gleason (Ed.), *Higher Education in the Era of the Fourth Industrial Revolution.* Singapore: Palgrave Macmillan. https://doi.org/10.1007/978-981-13-0194-0_9

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 computerised adaptive test (CAT) measurement of pain. *Quality of Life Research, 25*(1), 1-11. https://doi.org/10.1007/s11136-015-1069-5

Reckase, M. D. (2010). Designing item pools to optimise the functioning of a computerised adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Schwab, K. (2016). *The Fourth Industrial Revolution: what it means, how to respond.* World Economic Forum. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Segall, D. O. (2005). Computerised adaptive testing. *Encyclopedia of social measurement, 1*, 429-438. http://iacat.org/sites/default/files/biblio/se04-01.pdf

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

Singh, G., & Sagar, A. M. D. (2013). An Overview of Artificial Intelligence. SBIT *Journal of Sciences and Technology, 2*(1). https://doi.org/10.13140/RG.2.2.20660.19840

Staff Writer. (2020). September 16. South Africa moves to lockdown level 1-here are the changes. BUSINESSTECH. https://businesstech.co.za/news/trending/433943/south-africa-moves-to-lockdown-level-1-here-are-the-changes/

South African Government Disaster Management Act. (2020a). Alert level 1 lockdown regulations.

South African Disaster Management Act. (2020b). Alert level 5 lockdown regulations. https://bit.ly/33r7iUM

The Presidency, Republic of South Africa (2020). Statement by President Cyril Ramaphosa on South Africa's response to the coronavirus pandemic. https://bit.ly/3mezqlz

Theunissen, M. H., de Wolff, M. S., Deurloo, J. A., Vogels, A. G., & Reijneveld, S. A. (2020). Computerised adaptive testing to screen children for emotional and behavioural problems by preventive child healthcare. *BMC paediatrics, 20*(1), 1-7. https://bmcpediatr.biomedcentral.com/articles/10.1186/s12887-020-2018-1

Thompson, N. A. (2011). *Advantages of computerised adaptive testing (CAT).* https://assess.com/docs/Advantages-of-CAT-Testing.pdf

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

Thompson, N. (2018, March 23). *Overview: Steps to develop an adaptive test*. Assessment Systems. https://assess.com/2018/03/23/develop-computerized-adaptive-test/

Thompson, N. (n.d.). *Secure Online Testing with FastTest.* https://assess.com/fasttest-secure-online-testing/

The World Bank Group. (2020a). *How countries are using edtech (including online learning, radio, television, texting) to support access to remote learning during the COVID-19 Pandemic*. https://bit.ly/3vULWvd

The World Bank Group. (2020b). *Digital Technologies in Education. The use of information and communication technologies in education can play a crucial role in providing new and innovative forms of support to teachers, students, and the learning process more broadly*. https://www.worldbank.org/en/topic/edutech

Van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*(3), 283-302. http://www.jstor.com/stable/20461794

Van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. Linden, W. J. van der Linden, & C. A. Glas (Eds.), *Computerised adaptive testing: Theory and practice.* https://doi.org/10.1007/0-306-47531-6_1

Vecchiatto, P., Kew, J., & Prinsloo, L. (2020) *August 12. South Africa's Ramaphosa Prepares to Ease Lockdown Rules*. Bloomberg. https://www.bloomberg.com/news/articles/2020-08-12/south-africa-s-ramaphosa-prepares-to-ease-lockdown-restrictions

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust Computerised Adaptive Testing. In B. P. Veldkamp & C. Sluijter (Eds.), Theoretical and Practical Advances in Computer-based Educational Measurement (pp. 291-305). Cham: Springer. https://library.oapen.org/bitstream/handle/20.500.12657/22945/1007216.pdf?sequence=1#page=291

Wang, Z., & Thompson, N. (2020). *Digital Module 19: Foundations of IRT Estimation*. Instructional Topics in Educational Measurement Series (ITEMS). https://ncme.elevate.commpartners.com

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerised adaptive testing: A primer.* Routledge.

Ware, Jr., J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlöf, C. G. H., Tepper, S., & Dowson, A. (2003). Applications of computerised adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*(8), 935-952. https://doi.org/10.1023/a:1026115230284

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Practical questions in introducing computerised adaptive testing for K-12 assessments*. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/testnav/research-report-cat-for-k-12-assessments.pdf

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

Wise, S. L., & Kingsbury, G. G. (2000). Practical Issues in Developing and Maintaining a Computerised Adaptive Testing Program. *Psicológica, 21*,135-155.

Worldometers. (2020, November 28). *COVID-19 Coronavirus Pandemic*. https://www.worldometers.info/coronavirus/

Zhang, Y., Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a computerised adaptive testing for internet addiction. *Frontiers in Psychology, 10*, 1010.