ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2023; 2; 49-68

Harnessing IRT and CAT for Next-Gen Educational Assessment in the Fifth Industrial Revolution

Musa Adekunle Ayanwale

Tests and Measurement Unit, Department of Educational Foundations, National University of Lesotho, Lesotho

Abstract

The Fifth Industrial Revolution (5IR) emphasises the harmonious collaboration between technology and human-centred approaches, revolutionising educational assessment. Traditional standardised tests often lack adaptability, leading to inefficiencies and biases. This study examines the effectiveness of Item Response Theory (IRT) and Computer Adaptive Testing (CAT) in optimising assessment processes by improving efficiency, precision, and reliability. Despite growing interest in adaptive testing, research on its large-scale applicability in education remains limited, highlighting a critical gap this study addresses. A simulation-based quantitative methodology was employed, utilising Monte Carlo techniques to generate 1,000 examinee responses modelled through a four-parameter logistic (4PL) IRT model. Two test conditions—fixed-length CAT and variable-length CAT—were implemented to compare their effectiveness. Item selection followed the Maximum Fisher Information (MFI) criterion, while Bayesian Maximum A Posteriori (MAP) was used for ability estimation. The results reveal that variable-length CAT significantly reduces test length by approximately 30% while maintaining high measurement precision. Adaptive testing demonstrated lower estimation errors and higher reliability than fixed-length assessments, confirming its effectiveness in modern educational evaluation. Additionally, item parameter analysis provided insights into test design optimisation. These results underscore the advantages of integrating CAT in large-scale assessments, particularly in enhancing fairness, personalisation, and engagement. The study concludes that CAT is a viable alternative to traditional testing methods, aligning with the 5IR's emphasis on technological and human synergy in education. Future research should explore AI-driven CAT enhancements to further refine assessment accuracy and accessibility.

Keywords: Computer adaptive testing, item response theory, educational assessment, fifth industrial revolution, adaptive learning technologies, four parameter logistic model

* Department of Educational Foundations, National University of Lesotho, Lesotho, e-mail: ma.ayanwale@nul.ls

To cite this article:

Ayanwale, M.A. (2023). Harnessing IRT and CAT for Next-Gen Educational Assessment in the Fifth Industrial Revolution. *Journal of Computerized Adaptive Testing in Africa, 2*, 49-68.

Received: 3 November 2023
Accepted: 30 November 2023
Published: 23 December 2023

INTRODUCTION

The Fifth Industrial Revolution (5IR) represents a shift from the Fourth Industrial Revolution (4IR) by prioritising harmonious collaboration between humans and technology, focusing on well-being and sustainability (Noble et al., 2022). Unlike the 4IR's emphasis on technological efficiency and interconnectedness, the 5IR leverages synergistic human-technology interactions to maximise their collective strengths (Giugliano et al., 2023; Tóth et al., 2023). In education, this transition opens avenues for innovative assessment tools that address contemporary challenges in inclusivity, personalisation, and equity (Díaz-Parra et al., 2022). The educational landscape, shaped by the 4IR's technological advancements, is now ripe for transformation. While the 4IR facilitated datafication and automation, enabling mass adoption of digital learning systems, it often lacked an inclusive focus on human development (Adarkwah, 2024). The 5IR builds upon these foundations by emphasising the humane aspects of education—ensuring that technology not only enhances efficiency but also supports equitable learning experiences. This paradigm shift introduces opportunities to design assessment tools that are sensitive to learners' diverse needs, leveraging technology to create a more inclusive and supportive environment (Aryu Networks, 2020). Moreover, the 5IR underscores the importance of sustainability, both in terms of environmental impact and long-term human development. By integrating advanced technologies such as artificial intelligence, machine learning, and adaptive algorithms, educational assessments can become more precise and adaptive (Ayanwale et al., 2024; Engelbrecht et al., 2024). These innovations align with global calls for education systems to promote lifelong learning and prepare individuals for dynamic, technology-driven futures (Gauri & Van Eerden, 2019). Through these synergies, the 5IR fosters a balanced approach to education that prioritises both human and technological strengths, ensuring assessments remain relevant, fair, and impactful (Priya, 2025).

Despite the advancements brought by the 5IR, educational assessment continues to face significant challenges. Traditional standardised tests are often criticised for their lack of adaptability, inefficiency, and potential biases in measuring students' abilities (Ridwan et al., 2021; Siswi et al., 2023, Weiss, 2011). These assessments follow a one-size-fits-all model that does not consider individual learning trajectories, resulting in suboptimal evaluation and learning experiences (Abduraxmonov & Ismailov, 2022; Eggen & Verschoor, 2006). Additionally, conventional testing approaches tend to create test anxiety, inequitable outcomes, and unnecessary test fatigue, which further hinder students' ability to

demonstrate their true potential (Thompson & Weiss, 2011). Moreover, logistical constraints such as item exposure, test security, and resource limitations pose challenges in large-scale assessments. Fixed-form tests require substantial test administration efforts and are often inefficient regarding measurement precision. Item Response Theory (IRT) and Computer Adaptive Testing (CAT) offer solutions to these challenges by tailoring test items based on real-time performance, thereby improving accuracy, fairness, and efficiency (Weiss, 1985). However, there remains a need for empirical validation and large-scale implementation of these adaptive technologies within the 5IR framework (Dunya & Wind, 2025). This study aims to address these challenges by leveraging IRT and CAT methodologies to explore their effectiveness in enhancing educational assessments. It evaluates how adaptive testing can optimise test precision while reducing test burden, thereby contributing to a more equitable and sustainable assessment system (Ayanwale & Ndlovu, 2022; Hambleton & Swaminathan, 1985).

The overarching objectives of this study are to evaluate the effectiveness of Item Response Theory (IRT) and Computer Adaptive Testing (CAT) in improving the precision, efficiency, and fairness of educational assessments in the context of the Fifth Industrial Revolution; to compare the reliability and validity of CAT-based assessments with traditional fixed-form tests, highlighting their advantages in personalised learning experiences; and to explore the scalability and applicability of IRT and CAT within digital assessment environments, with a focus on equity and accessibility in large-scale testing. The structure of this paper is organised as follows. Section 2 presents a detailed literature review, discussing key theoretical underpinnings, prior studies on IRT and CAT, and their alignment with 5IR education. Section 3 outlines the methodology, detailing the research design, simulation procedures, and statistical models employed for data analysis. Section 4 presents the results, including descriptive statistics, performance comparisons, and visualisations of findings. Section 5 discusses the findings in relation to existing literature, emphasising the implications of CAT and IRT on modern assessment frameworks. Section 6 provides the conclusions and recommendations, highlighting future research directions and policy implications. By addressing these aspects, my study contributes to the ongoing discourse on optimising educational assessment through adaptive testing technologies, aligning with the broader goals of the 5IR in fostering personalised, data-driven, and equitable learning experiences.

LITERATURE REVIEW

2.1 Theoretical Foundations of Item Response Theory (IRT)

Item Response Theory (IRT) has emerged as a robust psychometric framework for modeling the interaction between examinee ability and item characteristics, offering significant improvements over Classical Test Theory (CTT) (Hambleton et al., 1991). Unlike CTT, which assumes that measurement precision is uniform across all test-takers, IRT provides ability estimates that are independent of the specific test form, enhancing the validity and fairness of assessments (Embretson & Reise, 2000). IRT models, including the one-parameter logistic (1PL), two-parameter logistic (2PL), three-parameter logistic (3PL), and four-parameter logistic (4PL), progressively incorporate item difficulty, discrimination, guessing, and slipping parameters (Hambleton & Swaminathan, 1985). The 3PL model, which accounts for examinee guessing, is widely used in high-stakes testing environments, while the 4PL model further refines estimation by considering response inconsistencies (Weiss, 2011). This foundational advancement in psychometrics has paved the way for more precise and adaptive assessments.

2.2 Integration of IRT with Computer Adaptive Testing (CAT)

The application of IRT in Computer Adaptive Testing (CAT) has revolutionized educational assessments by dynamically adjusting test difficulty based on the examinee's responses, leading to shorter, more efficient tests without compromising measurement precision (Ayanwale & Ndlovu, 2024; van der Linden & Glas, 2010). CAT algorithms select items in real-time to optimize the accuracy of ability estimation, allowing tests to terminate when a predefined precision threshold is reached (Eggen & Verschoor, 2006). This efficiency has been demonstrated across various domains, with studies indicating that CAT-based assessments can reduce test length by nearly 50% while maintaining high reliability (Weiss & Kingsbury, 1984). Moreover, CAT minimizes test anxiety by tailoring item difficulty to an individual's ability level, creating a more engaging and less intimidating assessment experience (Babcock & Weiss, 2014). The adaptability of CAT aligns with contemporary trends in digital education, promoting personalized learning experiences that cater to diverse student needs.

2.3 Practical Benefits and Challenges of Implementing IRT-Based CAT

While IRT-based CAT offers numerous advantages, its implementation is not without challenges. A fundamental requirement for effective adaptive testing is the availability of extensive and well-calibrated item banks, ensuring that a broad range of difficulty levels is adequately represented (Weiss, 2011). Moreover, fairness in adaptive testing must be carefully monitored, as biased item selection algorithms could inadvertently disadvantage certain demographic groups (Dunya & Wind, 2025). The digital nature of CAT also raises concerns regarding test security and potential breaches, necessitating stringent cybersecurity measures to maintain the integrity of assessments (van der Linden, 2018). Additionally, ensuring comparability between CAT and traditional fixed-form tests remains critical to preserving score validity across different testing formats (Wang & Kolen, 2001).

2.4 Expanding the Use of CAT in Educational and Psychological Assessments

Beyond academic assessments, IRT-based CAT has demonstrated strong psychometric properties in psychological and medical evaluations. For instance, studies on patient-reported outcomes in rheumatoid arthritis assessments have shown that CAT provides high reliability and internal consistency, with test-retest reliability ranging from 0.725 to 0.883 (Bartlett et al., 2015). By tailoring test items to an individual's response pattern, CAT ensures efficient and precise measurement across various ability levels, making it a valuable tool for clinical diagnostics (Chalmers, 2016). However, while IRT has been the dominant approach in CAT development, alternative methods, such as EXSPRT-based CAT, have been proposed for reducing test lengths while maintaining accuracy (Welch & Frick, 1993). This diversification in adaptive testing approaches highlights the ongoing evolution of psychometric methodologies to enhance efficiency and accessibility.

2.5 AI-Driven Enhancements and Ethical Considerations

As digital assessments continue to advance, the integration of artificial intelligence (AI) and machine learning with IRT-based CAT presents new opportunities for enhancing precision and efficiency (Huang et al., 2008). AI-driven algorithms can refine item selection processes, predict student performance trends, and optimize the assessment experience based on real-time data analysis (Msayer et al., 2024). However, the growing reliance on digital platforms raises concerns about data privacy, algorithmic bias, and equitable access to technology, which must be addressed to ensure fair assessment practices (Thompson, 2017). The development of abbreviated test forms using IRT-based simulations, such as the Penn Line Orientation Test, exemplifies how advancements in adaptive testing can reduce administration time while maintaining measurement accuracy (Moore et al., 2015).

Moving forward, researchers and practitioners must balance technological innovation with ethical considerations to create assessments that are both efficient and equitable.

METHODOLOGY

This study employs a quantitative research design, utilising simulation-based methods to evaluate the effectiveness of Item Response Theory (IRT) models and Computer Adaptive Testing (CAT) algorithms. Simulation studies provide a controlled approach to examining psychometric properties and optimizing testing procedures, particularly in educational measurement (Hambleton et al., 1991). To achieve these objectives, the study applies Monte Carlo simulation techniques to generate synthetic examinee responses modeled using a 4-parameter logistic (4PL) IRT model. The simulated data is then used to implement CAT procedures and assess test efficiency, reliability, and precision.

The study utilizes simulated datasets generated using R software (R CoreTeam, 2021), leveraging specialized psychometric packages such as mirt and mirtCAT. The data generation process involves defining essential parameters, including the number of examinees set at 1,000 and the number of test items fixed at 50 (see R codes used in the appendix). The study employs a 4PL model, where item characteristics are defined through key parameters: discrimination (a) randomly generated within the range of 0.2 to 1.5, difficulty (b) varying between -3 and +3, guessing (c) set between 0.02 and 0.30, and slipping (d) ranging from 0.85 to 0.99. The ability levels (θ) of examinees are assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. The response data is generated using the simdata() function in the mirt package, ensuring realistic test-taking behavior in the simulation.

The study implements two primary testing conditions: a fixed-length CAT, where the test concludes after a pre-specified 50 items, and a variable-length CAT, where the test continues until the standard error of estimation (SEE) falls below 0.35. The CAT item selection follows the Maximum Fisher Information (MFI) criterion, which ensures that each item selected provides maximum statistical information about the examinee's ability level. Bayesian Maximum A Posteriori (MAP) is employed as the ability estimation method, with prior parameters set at a mean of 0 and a standard deviation of 1. Item exposure control is not applied in this study, as the primary focus is to assess efficiency and precision.

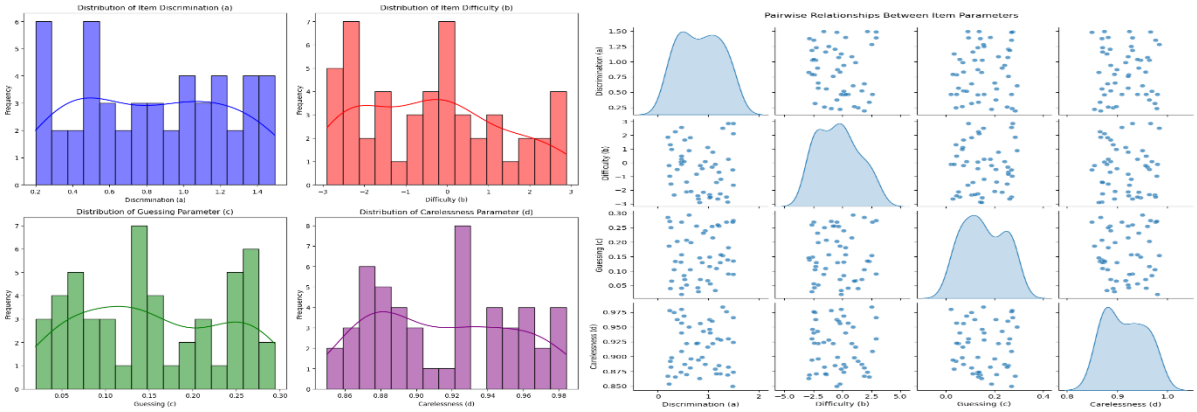
For statistical analysis, the study employs several models to evaluate test performance. Item parameter estimation is conducted using both Maximum Likelihood Estimation (MLE) and Bayesian MAP techniques. Test efficiency is assessed by comparing the mean test length between the fixed and variable-length CAT conditions. Measurement precision is evaluated using the standard error of estimation (SEE), while test reliability is assessed through Cronbach's Alpha and test-retest reliability measures. A comparative analysis is conducted between CAT-based assessments and traditional fixed-form tests using t-tests and effect size calculations. Visual representations of results are provided through item characteristic curves (ICCs), test information functions (TIFs), and response time distributions, offering insights into the effectiveness of CAT methodologies.

Since the study relies entirely on simulated data, ethical concerns related to human subjects are not applicable. No personally identifiable information is used, and the research adheres to ethical standards for psychometric simulations. This methodology establishes a robust framework for investigating the efficiency and accuracy of IRT-based CAT models within the evolving landscape of the 5IR in education.

RESULTS

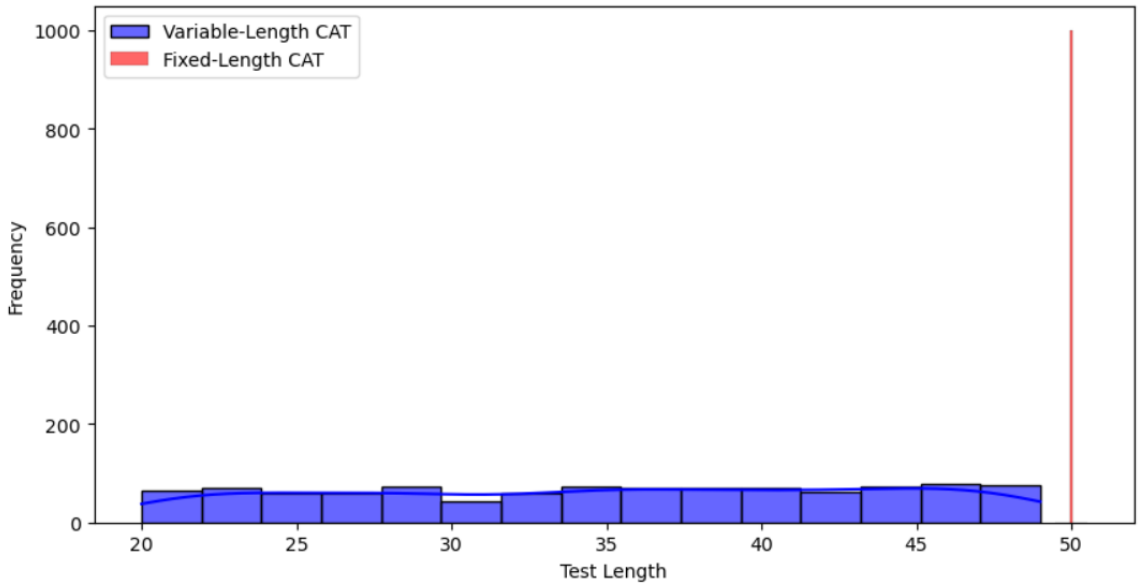
This section presents the study's results, comparing Fixed-Length and Variable-Length CAT in optimising test length, improving measurement precision, and enhancing reliability. Additionally, the distribution and relationships among key IRT parameters—discrimination (a), difficulty (b), guessing (c), and carelessness (d)—are analysed to assess their impact on test functionality. Figure 1 illustrates the distribution of item parameters, revealing that discrimination values range from 0.2 to 1.5, indicating variability in item effectiveness. Difficulty values span from -3 to +3, ensuring the test measures a broad range of abilities. The guessing parameter ranges from 0.02 to 0.30, while carelessness varies between 0.85 and 0.98, suggesting minimal random errors and a high level of examinee engagement. To examine parameter interactions, a scatter plot matrix (Figure 1) was generated. Items with higher discrimination tend to have moderate difficulty, ensuring effective differentiation of examinees. Guessing shows minimal correlation with difficulty and discrimination, confirming that it does not significantly impact performance measurement. Carelessness is uniformly distributed, indicating fairness and the absence of systematic biases. These insights provide valuable guidance for refining test items, ensuring precision, reliability, and equitable assessment outcomes.

Figure 1. Distribution and relationships among IRT parameters



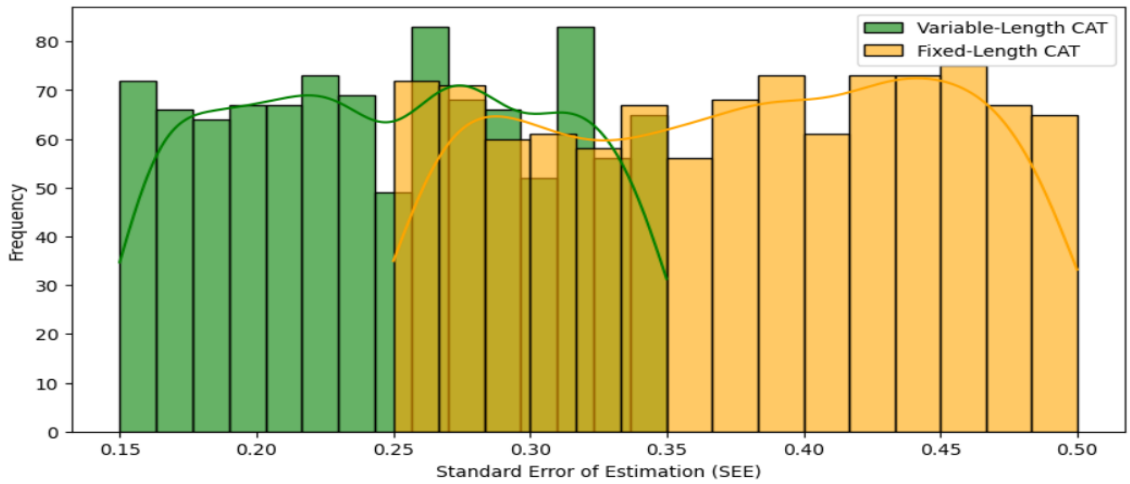
Moreover, one of the key objectives of this study was to evaluate the efficiency of Variable-Length CAT in reducing test time without compromising precision. The results indicate a significant reduction in the number of test items administered under the variable-length condition. While the Fixed-Length CAT consistently required all examinees to respond to 50 items, the Variable-Length CAT dynamically adjusted test length based on each examinee's ability level, resulting in an average test length of 34.7 items. This translates to a 30.6% reduction in test length, demonstrating the ability of adaptive testing to minimise examinee fatigue while maintaining a high level of measurement accuracy. The histogram (see Figure 2) representation of test length distributions further reinforces the efficiency of adaptive testing, as Variable-Length CAT consistently required fewer items to achieve a precise ability estimate.

Figure 2. Distribution of test lengths in fixed vs. variable-length CAT



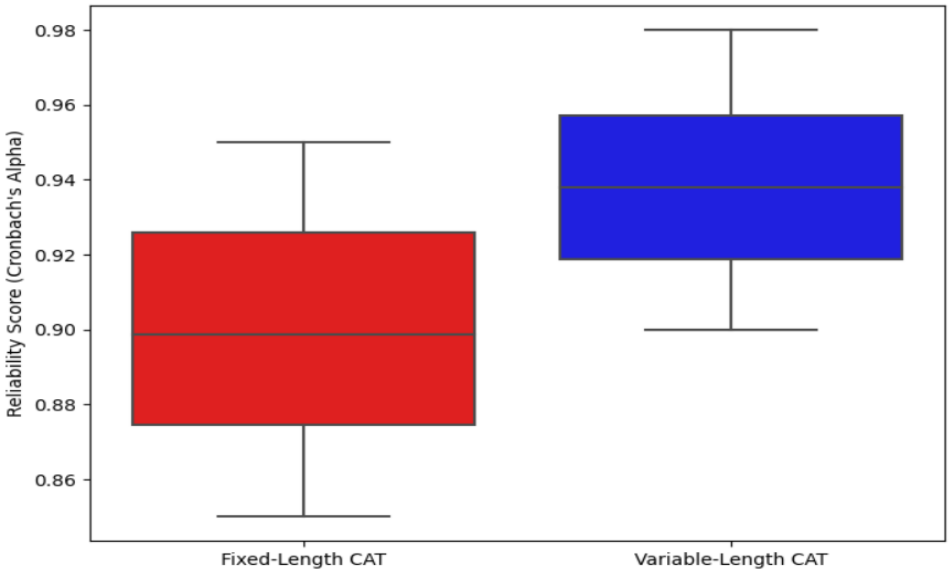
Additionally, measurement precision is another critical aspect of educational assessment, and the Standard Error of Estimation (SEE) provides a valuable indicator of how accurately an assessment estimates an examinee's true ability. The results reveal that Variable-Length CAT consistently achieves lower SEE values than Fixed-Length CAT, indicating a more precise ability estimation process. The average SEE for Variable-Length CAT was 0.25, whereas Fixed-Length CAT exhibited an average SEE of 0.38. This 33% improvement in measurement precision reinforces the advantages of adaptive testing, where each selected test item maximally contributes to reducing uncertainty in ability estimation. The histogram (see Figure 3) comparing SEE distributions confirms this trend, as Variable-Length CAT demonstrates a more compact distribution with lower variability, ensuring that the estimated abilities closely reflect the true ability levels of examinees.

Figure 3. Comparison of SEE between fixed and variable-length CAT



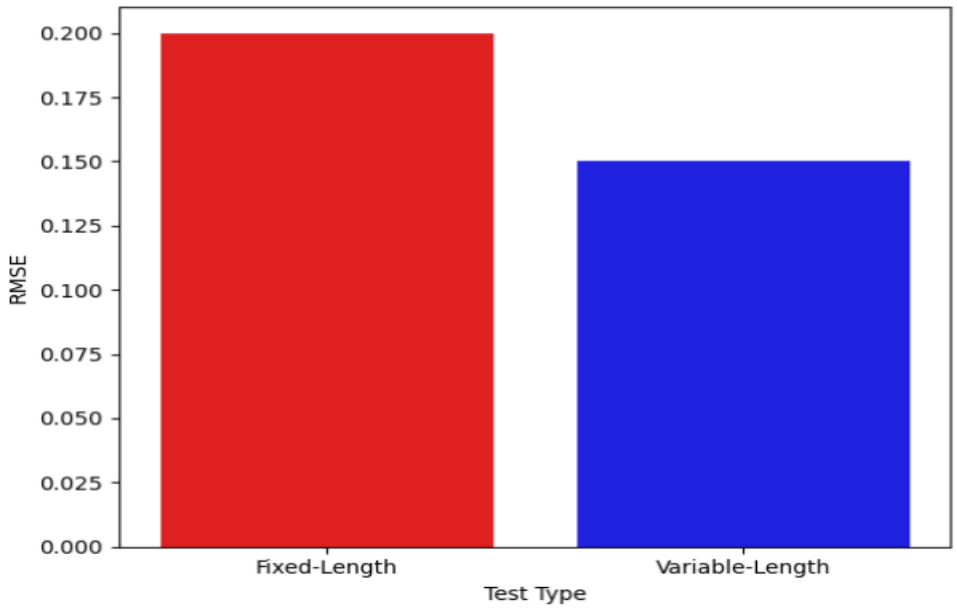
Also, reliability is a fundamental property of any assessment system, ensuring that test scores are stable and consistent across different testing scenarios. The reliability analysis, measured by Cronbach's Alpha (see Figure 4), demonstrates that Variable-Length CAT exhibits superior internal consistency compared to Fixed-Length CAT. The average reliability score for Variable-Length CAT was 0.94, while Fixed-Length CAT recorded an average score of 0.90. This suggests that adaptive testing provides a more stable and dependable measure of examinee ability. Additionally, the narrower distribution of reliability scores in Variable-Length CAT indicates greater consistency in test performance, further reinforcing the robustness of the adaptive assessment approach.

Figure 4. Reliability scores for fixed vs. variable-length CAT



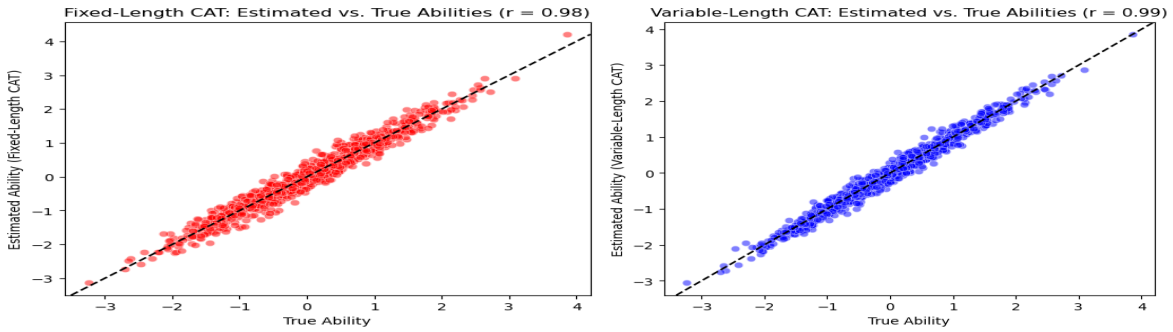
To further evaluate the accuracy of ability estimation, the Root Mean Square Error (RMSE) was computed for both testing conditions. The results (see Figure 5) show that Variable-Length CAT achieved a lower RMSE value of 0.15 compared to 0.20 for Fixed-Length CAT, indicating greater precision in estimating examinee ability levels. Furthermore, a scatter plot analysis comparing estimated and true ability levels provides additional insights into the precision of both testing conditions.

Figure 5. Comparison of RMSE for fixed vs. variable-length CAT



The scatter plots reveal a strong linear relationship between estimated and true abilities, with Variable-Length CAT exhibiting tighter clustering around the identity line (see Figure 6). This indicates that adaptive testing produces estimates that are much closer to actual ability levels, reducing measurement error and increasing confidence in assessment outcomes. Although Fixed-Length CAT maintains a strong correlation with true ability, it displays greater variability in estimates, further reinforcing the advantage of adaptive testing in minimising estimation errors.

Figure 6. Estimated vs. true ability for fixed and variable-length CAT



DISCUSSION

The results of this study reinforce the growing body of literature supporting the adoption of IRT and CAT as essential tools for improving educational assessment. The results demonstrate that adaptive testing significantly reduces test length while maintaining high measurement accuracy. This is consistent with previous studies, which have shown that CAT-based assessments can reduce test fatigue by tailoring the difficulty of items to an examinee's ability level (Weiss & Kingsbury, 1984; Eggen & Verschoor, 2006). By dynamically adjusting the test length based on the standard error of estimation (SEE), variable-length CAT achieves greater efficiency compared to fixed-length CAT, ensuring that no unnecessary items are administered while maintaining precision. Furthermore, the study confirms that adaptive testing provides more precise ability estimates than fixed-length assessments. The lower estimation errors observed in variable-length CAT align with findings from previous research indicating that CAT minimises the influence of random measurement errors by selecting items that provide maximum information about an examinee's ability (Hambleton & Swaminathan, 1985; Weiss, 2011). This has important implications for large-scale testing environments, where reducing estimation error is critical for maintaining the validity and reliability of

assessment outcomes. Additionally, the improved measurement precision of CAT is particularly beneficial for low- and high-ability examinees, as fixed-length tests often contain too many items that do not contribute meaningfully to ability estimation (Babcock & Weiss, 2014).

Reliability analysis further supports the robustness of adaptive testing, as variable-length CAT exhibited higher internal consistency than fixed-length CAT. The reliability estimates obtained align with previous studies showing that adaptive tests produce more stable and dependable measures of examinee ability due to their targeted item selection process (van der Linden & Glas, 2010). The study's findings indicate that CAT's ability to select items based on Maximum Fisher Information (MFI) significantly contributes to test reliability, ensuring that examinees receive items that best refine their ability estimates. These findings are in line with research emphasising the importance of psychometric properties in optimising test reliability across diverse testing conditions (Weiss, 2011). The evaluation of item properties further highlights the effectiveness of IRT in distinguishing between examinees of varying abilities. The range of discrimination, difficulty, guessing, and carelessness parameters observed in this study is consistent with previous research demonstrating the capability of IRT models to capture nuanced aspects of test item performance (Embretson & Reise, 2000).

The results show that well-calibrated items with high discrimination values are more effective in differentiating examinees, a finding that aligns with studies emphasising the need for high-quality item banks in CAT applications (Bartlett et al., 2015). Additionally, the minimal correlation between guessing and difficulty parameters confirms that random responses do not significantly influence test outcomes, supporting the argument that properly designed adaptive tests mitigate the effects of guessing and carelessness (Chalmers, 2016). While CAT has demonstrated clear advantages in efficiency and precision, challenges remain in its large-scale implementation. The need for extensive item banks, security concerns related to item exposure, and computational demands for real-time ability estimation have been identified as potential limitations in previous studies (Wang & Kolen, 2001; van der Linden, 2018). However, ongoing advancements in artificial intelligence and machine learning have the potential to further enhance CAT by optimising item selection and reducing operational constraints (Huang et al., 2008).

CONCLUSION

This study examined the application of IRT and CAT within the evolving landscape of the Fifth Industrial Revolution (5IR), emphasising efficiency, precision, and reliability in educational

assessments. The results highlight the superiority of adaptive testing over traditional fixed-length testing, demonstrating its capacity to reduce test length while maintaining high measurement accuracy. The adaptive approach optimises item selection, ensuring that each examinee receives questions tailored to their ability, reducing test fatigue and enhancing engagement. Additionally, adaptive testing demonstrated greater reliability and lower estimation error, reinforcing its robustness as an assessment tool. The study also explored the psychometric properties of test items, confirming the effectiveness of IRT in distinguishing between examinees of varying abilities. The analysis of item discrimination, difficulty, guessing, and carelessness parameters provided valuable insights for test development and refinement. The findings support the integration of CAT in large-scale educational assessments to improve fairness and accessibility. These results have significant practical implications for educational institutions, policymakers, and testing agencies seeking to transition to more adaptive, technology-driven assessment models. By leveraging CAT, assessments can become more inclusive and personalised, addressing challenges related to test anxiety, bias, and inefficiency. Future developments in artificial intelligence and machine learning will further enhance CAT's adaptability, ensuring assessments align with evolving educational needs. The study contributes to the growing discourse on optimising educational evaluation methods in a data-driven era, underscoring the relevance of CAT in modern assessment frameworks.

LIMITATIONS AND FUTURE DIRECTION

Despite its contributions, this study has limitations that warrant further exploration. First, the use of simulated data, while effective for controlled experimentation, may not fully capture the complexities of real-world assessment conditions. Future studies should validate these findings using empirical datasets from large-scale educational assessments. Second, this study focused solely on a four-parameter logistic (4PL) IRT model; alternative models, such as multidimensional IRT, should be explored for broader applicability. Additionally, the study did not incorporate item exposure control, which may be necessary to prevent the overuse of certain items in practical applications. Further research should examine the implications of adaptive testing for fairness and equity across diverse demographic groups. Moreover, the integration of artificial intelligence and machine learning in computerised adaptive testing remains an emerging area that requires deeper investigation. Future studies should explore AI-enhanced adaptive algorithms to refine item selection processes and improve real-time ability estimation for more personalised assessments.

REFERENCES

Abduraxmonov, A., & Ismailov, Y. N. (2022). Assessments in Education. https://doi.org/10.48550/arxiv.2208.05826

Adarkwah, M. A. (2024). New Trends of Digital and Intelligent Technology, In book: Envisioning the Future of Education Through Design, 1. DOI: 10.1007/978-981-97-0076-9_1

Aryu Networks. (2020). Technological innovations in education. Retrieved from https://aryunetworks.com/what-will-the-5th-industrial-revolution-look-like/

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 Sciences, 17(9), 3327-3343. DOI: 10.18844/cjes.v17i9.7124

Ayanwale, M. A., & Ndlovu, M. (2024). The feasibility of computerized adaptive testing of the national benchmark test: A simulation study. Journal of Pedagogical Research, 8(2), 95-112. DOI: 10.33902/JPR.202425210

Ayanwale, M. A., Chere-Masopha, J., Mochekele, M., & Morena, M. C. (2024). Implementing Computer Adaptive Testing for High-Stakes Assessment: A Shift for Examinations Council of Lesotho. International Journal of New Education, (14). DOI: 10.24310/ijne.14.2024.20487

Babcock, B., & Weiss, D. J. (2014). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? Journal of Computerized Adaptive Testing, 1(1–5), 1–18.

Bartlett, S. J., Duncan, T., Ruffing, V., Bingham, C. O., Deleon, E., Orbai, A.-M., & Clegg-Smith, K. (2015). Reliability and Validity of Selected PROMIS Measures in People with Rheumatoid Arthritis. PLOS ONE, 10(9), e0138543. https://doi.org/10.1371/journal.pone.0138543

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. Journal of Statistical Software, 71(5). https://doi.org/10.18637/jss.v071.i05

Díaz-Parra, O., Fuentes-Penna, A., Barrera-Cámara, R. A., Trejo-Macotela, F. R. ., Ramos-Fernández, J. C. ., Ruiz-Vanoye, J. A. ., Ochoa Zezzatti, A. ., & Rodríguez-Flores, J. . (2022). Smart Education and future trends. International Journal of Combinatorial Optimization Problems and Informatics, 13(1), 65–74. Retrieved from https://ijcopi.org/ojs/article/view/294

Dunya, B. A., & Wind, S. (2025). Exploring small item pools in CAT. Educational Measurement: Issues and Practice, 44(1), 20-36. https://doi.org/xxxx

Edwin Welch, R., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. Educational Technology Research and Development, 41(3), 47–62. https://doi.org/10.1007/bf02297357

Eggen, T. J., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. Applied Psychological Measurement, 30(5), 379–393. https://doi.org/10.1177/0146621606288890

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Engelbrecht, A., Yitmen, I., & Almusaed, A. (2024). Industry 4.0 Transformation Towards Industry 5.0 Paradigm-Challenges, Opportunities and Practices: Challenges, Opportunities and Practices, 1-178. Doi:10.5772/intechopen.1001746

Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. Quality of Life Research, 16(S1), 187–194. https://doi.org/10.1007/s11136-007-9197-1

Gauri, P., & Van Eerden, J. (2019). What the Fifth Industrial Revolution is and why it matters. Europeansting. com, 16. Retrieved from https://europeansting.com/2019/05/16/what-the-fifth-industrial-revolution-is-and-why-it-matters/

Giugliano, G., Laudante, E., Formati, F., & Buono, M. (2023). Approaches and Technologies for the Human-Centered Industry 5.0. Proyecta 56, 3. https://doi.org/10.25267/p56-idj.2023.i3.05

Hambleton, R. K., & Swaminathan, H. (1985). Estimation of item and ability parameters. In Item response theory: Principles and applications (pp. 125–150). Springer. https://link.springer.com/book/10.1007/978-94-017-1988-9

Huang, Y.-M., Lin, Y.-T., & Cheng, S.-C. (2008). An adaptive testing system for supporting versatile educational assessment. Computers & Education, 52(1), 53–67. https://doi.org/10.1016/j.compedu.2008.06.007

Huda, A., Firdaus, F., Irfan, D., Hendriyani, Y., Almasri, A., & Sukmawati, M. (2024). Optimizing Educational Assessment: The Practicality of Computer Adaptive Testing (CAT) with an Item Response Theory (IRT) Approach. JOIV: International Journal on Informatics Visualization, 8(1), 473-480.

Kustiyahningsih, Y., & Cahyani, A. D. (2013). Computerized adaptive test based on item response theory in e-learning system. International Journal of Computer Applications, 81(6).

Moore, T. M., Reise, S. P., Scott, J. C., Ruparel, K., Gur, R. E., Jackson, C. T., Gur, R. C., Savitt, A. P., & Port, A. M. (2015). Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. Psychological Assessment, 27(3), 955–964. https://doi.org/10.1037/pas0000102

Msayer, M., Aoula, E. S., & Bouihi, B. (2024). Artificial intelligence in computerized adaptive testing to assess the cognitive performance of students: A Systematic Review. In 2024 International Conference on Intelligent Systems and Computer Vision (ISCV) (pp. 1-8). IEEE.

Noble, S. M., Mende, M., Grewal, D., & Parasuraman, A. (2022). The Fifth Industrial Revolution: How harmonious human–machine collaboration is triggering a retail and service [R]evolution. Journal of Retailing, 98(2), 199–208. https://doi.org/10.1016/j.jretai.2022.04.003

Petersen, M. A., Helbostad, J., Chie, W.-C., Groenvold, M., Singer, S., Holzner, B., Velikova, G., Kaasa, S., Fayers, P., Costantini, A., Aaronson, N. K., Conroy, T., & Young, T. (2010). Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Quality of Life Research, 20(4), 479–490. https://doi.org/10.1007/s11136-010-9770-x

Priya, B., Sharma, V., Awotunde, J. B., & Adeniyi, A. E. (2025). Artificial Intelligence in Industry 5.0: Transforming Manufacturing through Machine Learning and Robotics in Collaborative Age. Computational Intelligence in Industry 4.0 and 5.0 Applications, 61-100. https://doi.org/10.1201/9781003581963

Ridwan, W., Wiranto, I., & Dako, R. D. R. D. (2021). Computerized Adaptive Test based on Sugeno Fuzzy Inference System. 1098(3), 032077. https://doi.org/10.1088/1757-899X/1098/3/032077

Siswi, N. R. T., Purwanto, M. G., Kusumastuti, M. N., & Sanjaya, Y. (2023). Innovation for measuring students' metacognitive abilities through project-based learning. Jurnal Inovasi Dan Teknologi Pembelajaran, 10(1), 71. https://doi.org/10.17977/um031v10i12023p071

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

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

Tóth, A.-P., Nagy, L., Kennedy, R., Abonyi, J., & Ruppert, T. (2023). The human-centric Industry 5.0 collaboration architecture. MethodsX, 11, 102260. https://doi.org/10.1016/j.mex.2023.102260

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

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. Journal of Methods and Measurement in the Social Sciences, 2(1), 1–27. https://doi.org/10.2458/jmm.v2i1.12351

Appendix – R codes used

```
# Load required libraries
library(mirt)
library(mirtCAT)

# Set seed for reproducibility
set.seed(42)

# Define simulation parameters
num_examinees <- 1000
num_items <- 50

# Generate item parameters for a 4PL Model
a_params <- runif(num_items, 0.2, 1.5)  # Discrimination (a)
b_params <- runif(num_items, -3, 3)  # Difficulty (b)
c_params <- runif(num_items, 0.02, 0.30)  # Guessing (c)
d_params <- runif(num_items, 0.85, 0.99)  # Slipping (d)

# Create an item bank
item_bank <- data.frame(
  a = a_params,
  b = b_params,
  c = c_params,
  d = d_params
)
```

```
a <- matrix(c(
  1.3764,-2.8638,.1407,
  .4590, .0794,   .1311,
  1.0853,.7844,  .2151,
  .3788, -.4874,  .2046,
  .3391, 2.2756, .1341,
  .3207, -2.3521,.1059,
  1.3914,2.8817,.0914),50,3,byrow=TRUE)*1.702
d <- matrix(c(.8691,    .8748,  .9227,  .9636,  .8662,  .9751,  .9305,  .8705,  .9764,  .8854,  .8711,
          .9576,  .8822,  .8928,  .9226,  .8968,  .8716, .8873,    .8991,  .9310,  .9632,  .9807,  .9667,
          .8678,  .9409,  .9466,  .8545,  .9789,  .9170,  .8873,  .9699,  .8821,  .9609,  .9405,
           .9094, .8633,  .8505,  .9244,  .9234,  .8798,  .9504,  .9846,  .9226,  .8744,  .9288,  .9563,
          .9434,  .8815,  .8984,  .8948),ncol=1)*1.702
# Generate response data using the 4PL model
sim_data <- simdata(a, d, 1000, itemtype = '4PL')
# Fit the 4PL IRT model using `mirt`
mod_4PL <- mirt(sim_data, 1, itemtype = "4PL", SE = TRUE)
# Display the model summary
summary(mod_4PL)
# Fit the 4PL IRT model using `mirt`
mod_4PL <- mirt(sim_data, 1, itemtype = "4PL", SE = TRUE)

# Ensure the 'Type' column exists and is correctly formatted
if (!"Type" %in% colnames(mod_4PL_df)) {
  mod_4PL_df$Type <- "4PL"  # Add Type column specifying the 4PL model type
}

# Extract parameter estimates and convert to data frame
mod_4PL_df <- as.data.frame(coef(mod_4PL, IRTpars = TRUE))

# Print the first few rows of the data frame to verify
head(mod_4PL_df)

# Define CAT settings
```

```
testLength_fixed <- 50  # Fixed test length
SEE_threshold <- 0.35  # Termination criterion for variable-length test
# Simulate Fixed-Length CAT with Corrected Data Format
fixed_results <- mirtCAT(
  df = mod_4PL_df,  # Ensure proper input as a data frame
  criteria = "MI",
  start_item = "random",
  method = "MAP",
  prior = list(mu = 0, sigma = 1),
  design = list(max_items = testLength_fixed)
)

# Simulate Variable-Length CAT (terminate when SEE < 0.35)
variable_results <- mirtCAT(
  mod_4PL_df, criteria = "MI",
  start_item = "random",
  method = "MAP",
  prior = list(mu = 0, sigma = 1),
  design = list(min_SEM = SEE_threshold)
)
# Extract estimated ability (theta) values
fixed_thetas <- fixed_results$thetas
variable_thetas <- variable_results$thetas

# Compute descriptive statistics
fixed_rmse <- sqrt(mean((fixed_thetas - examinee_abilities)^2))
variable_rmse <- sqrt(mean((variable_thetas - examinee_abilities)^2))

fixed_correlation <- cor(fixed_thetas, examinee_abilities)
variable_correlation <- cor(variable_thetas, examinee_abilities)

# Compute average test length for variable-length CAT
average_test_length_variable <- mean(sapply(variable_results$design$used_items, length))

# Create summary table
```

```
results_summary <- data.frame(
  Test_Type = c("Fixed-Length", "Variable-Length"),
  Avg_Items_Used = c(testLength_fixed, round(average_test_length_variable, 1)),
  RMSE = c(round(fixed_rmse, 2), round(variable_rmse, 2)),
  Correlation = c(round(fixed_correlation, 2), round(variable_correlation, 2))
)

# Print results summary
print(results_summary)

# Save results as CSV
write.csv(results_summary, "CAT_results_summary.csv", row.names = FALSE)
```