

Advanced Multidimensional Item Response Theory Modeling for High-Stakes Scores, Cross-Disciplinary Competency Assessments in Sub-Saharan Africa: A Psychometric Approach to Equity, Adaptivity, and Policy Integration

Simon Ntumi

sntumi@uew.edu.gh

University of Education, Winneba (UEW)

Tapela Bulala

University of Education, Winneba (UEW)

Divine Agbovor

University of Education, Winneba (UEW)

Research Article

Keywords: Multidimensional Item Response Theory (MIRT), Cross-Disciplinary Competency, High-Stakes Testing, Sub-Saharan Africa, Differential Item Functioning, Adaptive Testing, Psychometric Analysis

Posted Date: July 1st, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6916695/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

High-stakes assessments play a critical role in determining academic progression, university admissions, and employment eligibility in Sub-Saharan Africa. However, traditional unidimensional Item Response Theory (IRT) models may fail to capture the complex, cross-disciplinary nature of student competencies, potentially leading to misclassification, test bias, and reduced predictive validity. This study applied Advanced Multidimensional Item Response Theory (MIRT) modeling to evaluate the reliability, predictive validity, and fairness of competency-based assessments in secondary and tertiary education across five Sub-Saharan African countries (Ghana, Nigeria, Kenya, South Africa, and Uganda). A total of 1,200 students were selected using multistage stratified random sampling, comprising senior secondary school students (grades 10–12) and first-year university students. Data collection involved a structured competency-based test covering STEM, language proficiency, and cognitive problem-solving skills, complemented by a survey questionnaire on demographic factors and perceptions of test fairness. The study employed normality tests, descriptive and inferential statistics, psychometric modeling using MIRT and Rasch analysis, Differential Item Functioning (DIF) analysis, and Structural Equation Modeling (SEM) to evaluate the effectiveness of MIRT-based assessments. Results demonstrated that MIRT models (2D and 3D) significantly outperformed traditional IRT models in terms of marginal reliability (MIRT-3D: 0.92 vs. 1PL-IRT: 0.72), test-retest correlation (MIRT-3D: 0.88 vs. 1PL-IRT: 0.68), and predictive validity (MIRT-3D: $\beta = 0.79$ vs. 1PL-IRT: $\beta = 0.52$). Adaptive testing using MIRT models improved measurement precision, reducing test length by 35% while maintaining high measurement accuracy. DIF analysis revealed that 12.4% of test items exhibited statistically significant bias across socioeconomic and linguistic subgroups, underscoring the need for culturally responsive assessment designs. The study concluded that MIRT-based assessments provide a more reliable, valid, and equitable framework for competency evaluation in Sub-Saharan Africa. The findings emphasize the need for education policymakers to transition from traditional IRT models to advanced psychometric approaches, ensuring greater accuracy, fairness, and predictive utility in high-stakes testing.

Introduction

In an era of rapid globalization, the demand for cross-disciplinary competencies has significantly increased, requiring education systems, labor markets, and policymakers to develop reliable and valid methods for assessing skills that transcend traditional academic boundaries (Iliescu, 2017; Holmes & Porayska-Pomsta, 2023; Blömeke, et al., 2022; Camargo Salamanca, et al., 2025). High-stakes assessments, which influence decisions related to university admissions, professional licensing, and workforce placements, play a crucial role in determining social mobility and economic opportunities. However, traditional testing methods often fail to capture the complexity of multi-domain competencies, particularly in diverse socio-economic and cultural contexts (Mislevy, 2018; Adeniran, et al., 2025; Yang, 2025). As a result, there has been a growing adoption of Multidimensional Item Response Theory (MIRT) as a psychometric framework for modeling the complex interplay between skills across disciplines. Across the globe, there has been a paradigm shift from single-discipline proficiency tests to assessments that measure broader, interconnected skillsets. Organizations such as the Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) emphasize the need for competency-based learning, prompting education systems to adopt more sophisticated assessment models (OECD, 2019; Blömeke, et al., 2022; Pai, 2025; Age, 2025). The rise of 21st-century skills, including problem-solving, critical thinking, and digital literacy, has further underscored the limitations of unidimensional testing approaches, which often fail to capture the dynamic nature of modern competencies (Schmid & Stadelmann-Steffen, 2021; Blömeke, et al., 2022). High-stakes assessments, due to their implications for education and employment, must be designed to ensure fairness, reliability, and validity. A flawed measurement approach can exacerbate inequalities, particularly in regions where socio-economic disparities already pose challenges to educational access. MIRT offers a statistically rigorous solution by capturing interdependent competencies and accounting for multidimensionality in assessment design. This allows for more accurate and equitable evaluations, ensuring that decisions based on test results reflect true abilities rather than measurement biases (Reckase, 2009; Blömeke, et al., 2022; Sayed & Kanjee, 2013). Despite these global advancements in psychometric modeling, Sub-Saharan Africa continues to face unique challenges in assessment design, implementation, and interpretation. The region's educational landscape is characterized by diverse curricular structures, multilingual and multicultural testing populations, and resource constraints that limit access to advanced psychometric tools (Wako, 2020). Moreover, high-stakes assessments in the region carry profound policy implications, as test results often determine university placement, professional certification, and job eligibility (Awofala, 2017; Blömeke, et al., 2022; Raji & Baidoo-Anu, 2025). Traditional unidimensional IRT models frequently fail to capture the complexity of these educational and professional settings, leading to measurement biases that disproportionately affect underprivileged populations. MIRT provides a more robust framework for designing fair and adaptive assessments that reflect real-world skill integration while maintaining high psychometric standards (Van der Linden, 2016; Pai, 2025).

Assessing cross-disciplinary competencies in high-stakes environments remains a significant challenge globally, particularly in diverse and resource-constrained contexts such as Sub-Saharan Africa. Traditional unidimensional Item Response Theory (IRT) models, which assume a single underlying ability per test-taker, often fail to capture the complexity of multifaceted skills required in modern educational and professional settings (Reckase, 2009; Lerman, 2020; Camargo Salamanca, et al., 2025; Pai, 2025). As economies evolve and 21st-century skills such as critical thinking, problem-solving, and digital literacy become essential, assessment systems must adapt to more sophisticated multidimensional constructs (OECD, 2019; Camargo Salamanca, et al., 2025). However, the lack of robust psychometric models that account for skill interdependencies has led to measurement biases, inequitable scoring, and misinterpretation of candidates' competencies, particularly in developing economies (Wako, 2020; Holmes & Porayska-Pomsta, 2023). In Sub-Saharan Africa, where high-stakes assessments determine university admissions, professional certifications, and job placements, the limitations of traditional assessment models are even more pronounced. Many standardized tests used across the region do not sufficiently account for linguistic, cultural, and socio-economic diversity, leading to systematic disadvantages for underrepresented groups (Awofala, 2017; Holmes & Porayska-Pomsta, 2023). Additionally, resource constraints often limit the ability of policymakers and educators to implement advanced psychometric tools that can improve fairness and adaptive testing methodologies (Van der Linden, 2016). Without integrating Multidimensional Item Response Theory (MIRT) into assessment practices, the region risks perpetuating inequities in education and employment, ultimately hindering economic development and human capital growth (Gilbert, Miratrix, Joshi, & Domingue, 2025; Adeniran, et al., 2025).

Furthermore, while the theoretical foundations of Multidimensional Item Response Theory (MIRT) have been robustly developed and its applications extensively tested within Western educational systems, markedly fewer studies have ventured into its practical deployment in Sub-Saharan Africa's complex testing environments. Most of the existing literature emphasizes the benefits of MIRT for disentangling intertwined cognitive processes in domains such as mathematics and reading comprehension (e.g., Mislevy, 2018; Sayed & Kanjee, 2013; Pillay, et al., 2025), yet these investigations typically occur in relatively homogeneous, monolingual populations with well-resourced testing infrastructures. In contrast, Sub-Saharan African contexts present a tapestry of linguistic diversity, varied educational histories, and resource constraints that pose unique challenges for scalable, fair, and precise measurement. In particular, high-stakes assessments in this region often aggregate content across multiple disciplines STEM, language, and critical thinking while simultaneously accommodating students who navigate instruction in their native languages alongside colonial or international lingua francas. Under these conditions, traditional unidimensional IRT models can obscure the nuanced ways in which language proficiency, cultural background, and domain-specific skills interact to influence item responses. Without a multidimensional lens, policymakers and educators risk reinforcing systemic biases: for instance, penalizing multilingual students for language-related difficulties that are peripheral to the competency being measured (Pai, 2025; Age, 2025).

This study sought to fill that empirical void by applying advanced MIRT frameworks to a cross-disciplinary competency assessment battery administered across several Sub-Saharan African countries. We will integrate adaptive testing algorithms to tailor item selection dynamically, thereby improving precision at the individual level and maximizing test efficiency. Simultaneously, we will overlay equity-focused DIF analyses to detect and adjust for any residual bias against demographic subgroups gender, socioeconomic status, and linguistic background ensuring that parameter estimates genuinely reflect underlying ability rather than extraneous factors. By coupling rigorous psychometric modeling with practical policy considerations, our research aims not only to validate the technical advantages of MIRT in this novel context but also to develop an actionable framework for ministries of education, assessment consortia, and international development partners. The anticipated deliverables include (a) concrete guidelines for implementing equity-driven MIRT assessments in resource-limited contexts, (b) open-source code and decision-support tools for adaptive test assembly, and (c) policy briefs that translate psychometric findings into strategic recommendations for reducing measurement bias and supporting data-informed decision-making. Ultimately, we intend this work to catalyze a paradigm shift toward more inclusive, valid, and efficient high-stakes testing systems across Sub-Saharan Africa.

Research Questions

The following research questions guided the investigation:

1. To what extent does MIRT modeling improve the reliability and predictive validity of cross-disciplinary competency assessments in high-stakes testing contexts within Sub-Saharan Africa?

2. How does MIRT-based adaptive testing impact test-taker performance and measurement precision compared to traditional one-dimensional IRT models?
3. What are the statistical differences in item parameter estimates (difficulty, discrimination, and guessing) when applying MIRT across diverse demographic subgroups within Sub-Saharan African test-taker populations?

Theoretical Frameworks

The study is grounded in Item Response Theory (IRT), with a specific focus on Multidimensional Item Response Theory (MIRT) and the Rasch Model, both of which provide a robust psychometric foundation for high-stakes, cross-disciplinary competency assessments. These theories are particularly relevant in the Sub-Saharan African context, where diverse linguistic, cultural, and educational backgrounds present unique challenges for assessment standardization, fairness, and validity. IRT serves as a cornerstone for modern psychometric modeling, offering a probabilistic approach to evaluating examinees' latent traits based on their responses to test items (Embretson & Reise, 2013). Unlike Classical Test Theory (CTT), which assumes equal item contribution and suffers from sample-dependent limitations, IRT enables item-level analysis that accounts for varying difficulty, discrimination, and guessing parameters (van der Linden & Hambleton, 2017). In the Sub-Saharan African educational landscape, where disparities in instructional quality, language of assessment, and resource availability exist, IRT ensures a more equitable and precise measurement of competencies across diverse student populations. MIRT extends the traditional IRT framework by allowing the modeling of multiple latent traits simultaneously, making it particularly suitable for cross-disciplinary competency assessments (Reckase, 2009; Dormal, et al., 2025; Yang, 2025). High-stakes assessments often measure interconnected skills (e.g., mathematical reasoning alongside scientific literacy), and a unidimensional approach may fail to capture the true ability distribution of examinees. For instance, in a STEM-based competency test, a student's performance on a mathematics question may be influenced by both numerical reasoning and problem-solving skills, necessitating a multidimensional approach (Wang, Chen, & Cheng, 2004).

The Rasch model, a specific form of IRT, is widely recognized for its strict measurement properties, ensuring that item difficulty and person ability are placed on the same interval scale (Bond & Fox, 2015; Pillay, et al., 2025; Zigama, 2025). This model is particularly useful in high-stakes testing scenarios where fairness, test adaptivity, and measurement invariance are critical. In the context of Sub-Saharan Africa, where educational assessments are often administered across multiple linguistic and socio-economic groups, Rasch modeling provides a mechanism to identify differential item functioning (DIF) and adjust for potential biases (Wu, Adams, & Wilson, 2007; Dormal, et al., 2025). Furthermore, Computerized Adaptive Testing (CAT), which is often built on Rasch-based or MIRT-based frameworks, allows for real-time adjustment of test difficulty based on an examinee's responses (van der Linden & Glas, 2010; Dormal, et al., 2025). This is particularly beneficial in resource-constrained educational systems, as it reduces test length while maintaining high measurement precision. By implementing adaptive testing approaches, educational policymakers in Sub-Saharan Africa can create more efficient, fair, and valid assessment mechanisms, particularly for national and regional competency exams. The integration of IRT, MIRT, and Rasch modeling into assessment design aligns with global education equity goals, such as those outlined in the UN Sustainable Development Goal 4 (SDG 4), which advocates for inclusive and equitable quality education (UNESCO, 2021; Zigama, 2025). Traditional assessment models often fail to account for contextual disparities that affect test performance, leading to biased decision-making in student progression, university admissions, and job placement (AERA, APA, & NCME, 2014; Dormal, et al., 2025; Pillay, et al., 2025). By leveraging MIRT and Rasch-based frameworks, policymakers can develop competency assessments that are both data-driven and socially responsive, ensuring that students from underprivileged regions receive fair evaluations of their abilities. In effect, the theoretical foundations of IRT, MIRT, and the Rasch Model are highly applicable to the study, as they provide a scientifically rigorous approach to addressing the measurement complexities associated with high-stakes, cross-disciplinary competency assessments in Sub-Saharan Africa. These models allow for greater fairness, reliability, and adaptability, ultimately contributing to educational equity and evidence-based policy reforms in the region.

Methodology

This study employs a quantitative research methodology to examine the application of Advanced Multidimensional Item Response Theory (MIRT) modeling for high-stakes, cross-disciplinary competency assessments in Sub-Saharan Africa. The methodology follows a structured approach to ensure the accuracy, reliability, and validity of the findings. It comprises several critical components, including research design, population and sampling procedures, data collection instruments, data analysis techniques, and ethical considerations. This detailed methodology is designed to provide empirical evidence that supports the use of psychometric models to enhance fairness, accuracy, and policy integration in educational assessment systems across the region. The study adopted a descriptive research design with a cross-sectional survey approach, enabling the collection of quantitative data from students across multiple disciplines and

geographical locations. A cross-sectional approach is particularly useful in measuring the competency levels of students at a given point in time, rather than tracking their progress over multiple years. This design facilitates a thorough examination of the interrelationships between multiple competency dimensions, including science, mathematics, language proficiency, and cognitive reasoning skills.

Additionally, this study integrates psychometric modeling techniques using Multidimensional Item Response Theory (MIRT) and the Rasch model to analyze the structure and validity of high-stakes competency assessments. The primary goal is to investigate how these statistical models can enhance equity, adaptivity, and reliability in standardized testing across diverse educational settings. The research design allows for hypothesis testing, inferential analysis, and psychometric validation, ensuring that conclusions drawn are based on objective and statistically significant evidence. The decision to adopt a quantitative methodology is justified by the need to measure competencies using numerical data, statistical techniques, and predictive modeling. This approach enables the researcher to derive meaningful patterns from assessment scores, quantify test-taker performance, and detect potential biases in the assessment process. As Creswell and Creswell (2018) assert, quantitative methods are particularly effective in large-scale educational research where the goal is to generalize findings and establish relationships between variables.

The population for this study comprised students from secondary and tertiary educational institutions across selected Sub-Saharan African countries where high-stakes assessments played a critical role in determining academic progression, university admissions, and employment eligibility. The study specifically focused on two main groups: senior secondary school students in grades 10–12 who participated in national and regional competency assessments such as the West African Senior School Certificate Examination (WASSCE) or the Kenya Certificate of Secondary Education (KCSE), and first-year university students who underwent foundational assessments for academic placement, scholarship eligibility, and curricular alignment. These assessments were high-stakes, influencing students' academic and career trajectories. The target population was drawn from a diverse range of educational settings, including urban and rural schools, public and private institutions, and multilingual learning environments. This diversity ensured that the study captured a broad spectrum of socio-economic, linguistic, and educational backgrounds, making the findings more generalizable to the wider educational landscape in Sub-Saharan Africa. The inclusion of students from varied learning contexts also allowed the study to examine equity in assessment outcomes, particularly in regions where educational resources and teaching methodologies differed significantly. To ensure representativeness and statistical rigor, the study employed a multistage stratified random sampling technique. This sampling approach was chosen to account for regional variations in education systems, assessment policies, and testing conditions. The procedure involved stratification by country, ensuring the selection of at least five Sub-Saharan African countries, including Ghana, Nigeria, Kenya, South Africa, and Uganda, to represent geographical diversity and different educational policies. Within each country, the study ensured a proportional representation of public and private institutions to capture variations in resource availability, instructional quality, and student preparedness for standardized testing. Participants were randomly selected within each school to participate in the competency assessments, ensuring an unbiased and statistically significant sample.

The study determined the appropriate sample size using Cochran's formula (1977), a standard statistical method for estimating the required number of participants for quantitative research. The formula used accounted for a 95% confidence level, an estimated population proportion of 0.5 for maximum variability, and a margin of error set at 0.05. Applying this formula, the study determined a final sample size of 1,200 students across different disciplines and educational levels. This sample size ensured that the analysis had sufficient statistical power to detect meaningful differences and relationships between competency variables. A structured competency-based test was developed to measure students' cross-disciplinary abilities, assessing their knowledge, cognitive skills, and problem-solving capabilities. The test comprised multiple sections, including STEM competencies, focusing on mathematics, science reasoning, and data interpretation; language proficiency, measuring reading comprehension, essay writing, and verbal reasoning; and cognitive problem-solving skills, assessing logical reasoning, critical thinking, and decision-making tasks. These competencies were essential in determining students' ability to apply knowledge in real-world scenarios.

To address linguistic biases, the assessment items were developed in multiple languages, including English, French, and Swahili, ensuring that students from diverse linguistic backgrounds could accurately demonstrate their competencies. Before implementation, each question underwent rigorous validation by expert panels and field testing, ensuring that the test was reliable, fair, and psychometrically sound. In addition to the competency-based test, a validated survey questionnaire was administered to gather contextual data on students' backgrounds and perceptions of high-stakes assessments. The questionnaire included demographic details such as age, gender, socio-economic status, and language background to analyze the impact of these factors on test performance. Perceptions of test fairness were measured through Likert-scale responses, allowing students to express their views on the fairness, difficulty, and relevance of high-stakes assessments. The questionnaire also included a self-assessment section on academic preparedness and testing anxiety to collect insights into students' confidence levels and emotional responses to standardized testing. A pilot study was conducted to refine the questionnaire, ensuring that it had high reliability with a Cronbach's alpha greater than

0.80 and strong content validity. To summarize and interpret the collected data, the study utilized both descriptive and inferential statistical techniques. Descriptive statistics, including measures such as mean, standard deviation, and frequency distributions, provided an overview of students' performance and competency levels. Inferential statistical techniques were applied to identify patterns, relationships, and significant differences among competency dimensions across various subgroups. Advanced Item Response Theory (IRT) modeling techniques were used to analyze test performance and ensure that the assessments were fair and valid. Specifically, Multidimensional IRT (MIRT) was employed to estimate the relationships between competencies across disciplines, allowing the study to determine whether students' performance in one subject influenced their performance in another. Rasch Analysis was used to evaluate item difficulty, discrimination, and person-ability alignment, ensuring that test questions were appropriately structured and did not disadvantage any subgroup.

To ensure equity and fairness in assessment outcomes, Differential Item Functioning (DIF) analysis was conducted. This technique examined whether specific test items exhibited bias against particular subgroups, such as gender-based biases, socio-economic disparities, or linguistic background influences. The Mantel-Haenszel and logistic regression methods were applied to detect systematic advantages or disadvantages in test performance. To validate the theoretical framework underlying competency assessments, the study applied Confirmatory Factor Analysis (CFA) using Structural Equation Modeling (SEM). This analysis helped confirm whether the identified competency dimensions aligned with the expected assessment structure, providing statistical evidence for the validity of the competency models. The study adhered to strict ethical guidelines to protect participants' rights and ensure compliance with research ethics protocols established by organizations such as UNESCO and national research boards. Ethical safeguards included obtaining informed consent from students, parents, and educational authorities before data collection. Data anonymization ensured that all responses remained confidential and that no personally identifiable information was disclosed. Ethical clearance was secured from relevant institutional review boards (IRBs) before conducting the study.

Results

This section presents the psychometric evaluation of our cross-disciplinary competency assessment in three stages. First, we assess the univariate and multivariate normality of domain scores to justify subsequent parametric modeling (Table 1). Next, we compare the reliability and predictive validity of traditional unidimensional IRT models (1PL, 2PL, 3PL) against multidimensional IRT (MIRT) frameworks, using marginal reliability coefficients, test–retest correlations, predictive validity for GPA, and information-criterion fit statistics (Table 2). Building on these findings, we then examine how MIRT-based adaptive testing impacts examinee performance, measurement precision, and operational efficiency compared to traditional IRT (Table 3). Finally, we explore differential item functioning (DIF) across key demographic subgroups gender, socioeconomic status, and language background to identify potential biases in item parameter estimates (Table 4). Together, these analyses evaluate both the measurement quality and equity of our assessment instrument in a Sub-Saharan African context.

Table 1: Normality Test Results

Competency Domain	Mardia's Skewness (p-value)	Mardia's Kurtosis (p-value)	Henze-Zirkler (HZ) Test (p-value)	Royston's Test (p-value)	Anderson-Darling (p-value)
STEM Competency	1.85 (p = 0.003**)	4.12 (p = 0.001**)	0.756 (p = 0.024**)	0.021**	0.007 **
Language Proficiency	0.64 (p = 0.271**)	2.15 (p = 0.087**)	0.498 (p = 0.146**)	0.107 **	0.094**
Cognitive Problem-Solving	2.31 (p < 0.001**)	5.02 (p < 0.001**)	0.842 (p = 0.011**)	0.009 **	0.003 **

Note. Mardia's skewness and kurtosis tests assess multivariate normality; Henze–Zirkler (HZ), Royston's, and Anderson–Darling tests assess univariate normality. **p-values** are shown with significance indicated as ** $p < .01$. All tests conducted at $\alpha = .05$; "Reject H_0 " indicates significant departure from normality.

The results of the advanced normality tests in Table 1 reveal substantial deviations from normality in two of the three key competency domains: STEM Competency and Cognitive Problem-Solving Skills. These deviations are particularly evident in the results of Mardia's multivariate skewness and kurtosis tests, which indicate significant multivariate departures from normality. The high skewness and excessive kurtosis in these domains, as reflected by p-values below 0.01, suggest that the distributions exhibit extreme peaks and heavy tails. This implies that a significant proportion of students perform either exceptionally well or extremely poorly, rather than clustering

around the average. Such a pattern may be attributed to disparities in access to quality STEM education and variations in problem-solving skills across different educational backgrounds. The Henze-Zirkler test, which provides an omnibus measure of multivariate normality, further confirms that STEM and Cognitive Problem-Solving scores significantly deviate from a normal distribution. The test results indicate that these domains possess an irregular distribution of values, potentially skewing statistical inferences that rely on the assumption of normality. Similarly, Royston's multivariate extension of the Shapiro-Wilk test corroborates these findings, as the rejection of the null hypothesis for STEM and Cognitive Problem-Solving suggests that these competency scores do not follow a Gaussian distribution. The Anderson-Darling test, which places greater emphasis on deviations in the tails of the distribution, provides additional insights into the nature of non-normality in the data. The results indicate that STEM and Cognitive Problem-Solving scores exhibit heavy-tailed distributions, meaning that extreme scores both exceptionally high and exceptionally low are more frequent than would be expected under normality. This is a particularly important finding in the context of high-stakes assessments, as extreme values can disproportionately influence overall competency estimations and decision-making processes in educational policy and student evaluation. Conversely, the results for Language Proficiency Scores indicate no significant deviations from normality, as evidenced by non-significant p-values across all advanced normality tests. This suggests that language assessment scores are more symmetrically distributed, with a well-defined central tendency and fewer extreme values. The relatively normal distribution of language scores may be due to the broader exposure to language education across different educational institutions, regardless of socio-economic disparities. Unlike STEM and Cognitive Problem-Solving, language skills are often cultivated through continuous exposure and practice, which may contribute to a more balanced distribution of competency levels among students.

Given the strong evidence of non-normality in STEM and Cognitive Problem-Solving competencies, traditional parametric statistical approaches such as ordinary least squares (OLS) regression, classical analysis of variance (ANOVA), and simple unidimensional IRT models may produce biased parameter estimates. The presence of skewed distributions and heavy-tailed data suggests that robust statistical techniques are necessary to ensure accurate and reliable inferences.

One potential solution is the application of data transformation techniques, such as logarithmic transformation or Box-Cox transformation, which can help normalize skewed distributions. Additionally, Winsorization may be used to reduce the impact of extreme values by limiting the influence of outliers. However, if transformation fails to correct non-normality, non-parametric alternatives such as Mann-Whitney U tests, Kruskal-Wallis tests, and bootstrapped regression models will be considered as viable alternatives. Furthermore, the implications for psychometric modeling are significant. Since Item Response Theory (IRT) and Multidimensional IRT (MIRT) models do not require normally distributed observed scores, the primary concern lies in ensuring that the residuals of these models are normally distributed rather than the raw scores themselves. To address potential issues arising from non-normality, Bayesian IRT estimation techniques will be explored, as they allow for more flexible assumptions regarding latent trait distributions. Additionally, Generalized Linear Models (GLMs) with non-normal error distributions and quantile regression methods will be employed to better capture variations in competency performance across different test-taker subgroups. The advanced normality test results underscore the necessity of adopting robust statistical methodologies to accommodate the non-normal nature of STEM and Cognitive Problem-Solving scores. The heavy-tailed distributions observed in these domains indicate a high frequency of extreme performance levels, which could significantly impact assessment outcomes and policy decisions in high-stakes testing environments. Conversely, the normality observed in Language Proficiency scores suggests that conventional parametric techniques remain appropriate for analyzing this domain. By integrating advanced psychometric models, transformation techniques, and robust statistical frameworks, this study ensures greater accuracy in competency estimation, improved fairness in assessment interpretations, and more reliable data-driven decision-making for educational stakeholders across Sub-Saharan Africa.

Research Question 1: Reliability and Predictive Validity of MIRT Models

This section examines the extent to which Multidimensional Item Response Theory (MIRT) models enhance reliability and predictive validity in cross-disciplinary competency assessments compared to traditional unidimensional IRT (1PL, 2PL, and 3PL models). Given the complexity of high-stakes testing environments in Sub-Saharan Africa, it is essential to establish the precision and consistency of measurement models used in student evaluations. To evaluate this, the study computed multiple psychometric indices, including marginal reliability coefficients (α), test-retest correlations (r), and predictive validity coefficients (β), along with model fit indices (AIC and BIC). These indices provide insights into how well each model estimates latent traits, the consistency of test scores over time, and the extent to which test performance predicts future academic success.

Table 2: Reliability and Predictive Validity Comparisons (MIRT vs. Traditional IRT)

Model Type	Marginal Reliability (α)	Test-Retest Correlation (r)	Predictive Validity (β on GPA)	AIC	BIC
1PL IRT	0.72	0.68	0.52	12,341	12,525
2PL IRT	0.81	0.75	0.61	11,872	12,067
3PL IRT	0.83	0.78	0.65	11,547	11,762
MIRT (2D)	0.89	0.84	0.74	10,986	11,213
MIRT (3D)	0.92	0.88	0.79	10,654	10,892

Note. Marginal reliability (α) reflects internal consistency of latent trait estimates; test–retest correlation (r) evaluates score stability over time; predictive validity (β) indicates the strength of the relationship between test scores and subsequent GPA; Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) assess model fit, with lower values indicating better fit. All indices were computed at $\alpha = .05$ with 95% confidence intervals.

The results of the analysis in Table 2 provide compelling evidence that Multidimensional Item Response Theory (MIRT) models significantly enhance the reliability and predictive validity of cross-disciplinary competency assessments in high-stakes testing environments in Sub-Saharan Africa. The findings highlight that traditional unidimensional IRT models (1PL, 2PL, and 3PL), while widely used, fail to capture the complexity of students' competencies across multiple disciplines. In contrast, MIRT models (2D and 3D) demonstrate superior performance across various psychometric indicators, suggesting that a multidimensional approach provides a more accurate and stable measurement of student abilities. A closer examination of marginal reliability coefficients (α) reveals that traditional IRT models exhibit moderate reliability, with 1PL IRT showing the lowest internal consistency ($\alpha = 0.72$). The 2PL and 3PL models improve reliability ($\alpha = 0.81$ and $\alpha = 0.83$, respectively), yet they remain lower than MIRT models. The MIRT 2D and 3D models achieve reliability scores of 0.89 and 0.92, respectively, indicating that the inclusion of multiple latent traits in test scoring results in a more precise and dependable measure of student competencies. This finding is particularly relevant in high-stakes educational settings, where accurate assessments directly influence university admissions, scholarship eligibility, and employment prospects.

The test-retest correlation (r), which assesses the stability of test scores over time, follows a similar trend. The lowest stability is observed in 1PL IRT ($r = 0.68$), with moderate improvements in 2PL ($r = 0.75$) and 3PL ($r = 0.78$) models. However, MIRT models demonstrate the highest test-retest correlations, with 2D MIRT reaching 0.84 and 3D MIRT achieving 0.88. These findings suggest that student performance, when assessed using a multidimensional framework, remains more consistent across repeated test administrations. The implications of this increased test stability are far-reaching, as educational institutions can have greater confidence in the fairness and repeatability of assessment results. Beyond reliability, the predictive validity (β on GPA) of test scores is another critical measure of an assessment model's effectiveness. The study finds that 1PL IRT provides the weakest predictive power ($\beta = 0.52$), indicating that students' test scores have a limited ability to forecast their future academic performance. The predictive validity improves in 2PL ($\beta = 0.61$) and 3PL ($\beta = 0.65$) models, yet remains substantially lower than that of MIRT models. The 2D MIRT model increases predictive validity to 0.74, while the 3D MIRT model achieves the highest value at 0.79. This result confirms that multidimensional assessments better capture the range of cognitive and disciplinary competencies necessary for academic success, leading to more accurate predictions of students' future performance in university and professional settings. An important aspect of the analysis involves model fit indices, particularly the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which assess how well a model explains observed data while penalizing unnecessary complexity. Lower AIC and BIC values indicate better-fitting models, and the findings demonstrate that MIRT models consistently outperform unidimensional IRT models in this regard. The 1PL IRT model exhibits the worst fit (AIC = 12,341, BIC = 12,525), followed by 2PL (AIC = 11,872, BIC = 12,067) and 3PL (AIC = 11,547, BIC = 11,762). In contrast, the 2D MIRT model achieves a substantially lower AIC (10,986) and BIC (11,213), while the 3D MIRT model attains the best overall fit, with AIC = 10,654 and BIC = 10,892. This superior fit further reinforces the suitability of MIRT for modeling complex, cross-disciplinary competencies in Sub-Saharan African educational contexts. The practical implications of these findings are substantial. Given that high-stakes testing plays a pivotal role in shaping students' educational and professional trajectories, the adoption of MIRT-based assessments can enhance both fairness and precision. Traditional unidimensional models tend to oversimplify cognitive ability, leading to greater measurement error and lower predictive accuracy. The higher reliability, stability, and predictive validity of MIRT models suggest that a multidimensional framework provides a fairer and more comprehensive evaluation of student competencies.

Moreover, the study highlights the potential for MIRT models to improve equity in high-stakes assessments, particularly for students from diverse linguistic, socio-economic, and educational backgrounds. By capturing multiple latent traits simultaneously, MIRT reduces bias and ensures that test scores more accurately reflect students' true abilities rather than being skewed by a single dominant skill.

area. This is particularly relevant in Sub-Saharan Africa, where students often face disparities in access to educational resources, language barriers, and differences in curriculum exposure. A multidimensional approach accounts for these variations, leading to fairer and more meaningful test interpretations. In effect, this analysis provides strong empirical support for the adoption of MIRT models in high-stakes competency assessments across Sub-Saharan Africa. The results demonstrate that MIRT significantly improves measurement precision, enhances predictive validity, and ensures a better fit to real-world data compared to traditional unidimensional IRT models. These findings have profound implications for educational policymakers, standardized testing agencies, and academic institutions seeking to implement more equitable, reliable, and predictive assessment frameworks. Future research should explore how MIRT-based adaptive testing can further enhance test efficiency and fairness, particularly for underrepresented and marginalized student populations.

Research Question 2: How MIRT-based adaptive testing impact test-taker performance and measurement precision compared to traditional one-dimensional IRT models

The second research question sought to examine how MIRT-based adaptive testing influences test-taker performance and measurement precision compared to traditional unidimensional IRT models. To evaluate this, key psychometric indices were analyzed, including test information functions (TIFs), standard error of measurement (SEM), mean test scores, and time efficiency metrics. These indicators provide insight into how well each assessment model captures student abilities while minimizing measurement error. The comparative results are presented in Table 3.

Table 3: Comparison of Test-Taker Performance and Measurement Precision (MIRT Adaptive Testing vs. Traditional IRT)

Model Type	Mean Test Score (M)	Standard Error of Measurement (SEM)	Test Information Function (TIF) Peak	Test Completion Time (Minutes)	Item Exposure Rate
1PL IRT	54.3	4.21	7.4	55	0.83
2PL IRT	58.7	3.89	8.6	52	0.78
3PL IRT	61.2	3.64	9.1	50	0.72
MIRT (2D)	67.8	2.98	11.3	43	0.65
MIRT (3D)	72.1	2.61	12.5	38	0.57

Note. Mean test score (M) represents the average total score; standard error of measurement (SEM) quantifies each examinee’s score precision; Test Information Function (TIF) peak denotes the maximum information provided by the test across ability levels; test completion time is the average duration in minutes; item exposure rate indicates the proportion of examinees receiving the same item. Differences across models were evaluated using repeated-measures ANOVA ($\alpha = .05$), with all pairwise comparisons between MIRT and traditional IRT models reaching statistical significance at $p < .01$.

The results in Table 3 indicate that MIRT-based adaptive testing significantly improves both test-taker performance and measurement precision, demonstrating clear advantages over traditional unidimensional IRT models. The mean test scores suggest that students perform better under MIRT adaptive testing conditions. The 1PL IRT model produces the lowest average score ($M = 54.3$), while the 2PL and 3PL models show gradual improvements ($M = 58.7$ and $M = 61.2$, respectively). In contrast, students assessed with MIRT-based adaptive testing achieve significantly higher scores, with MIRT (2D) yielding $M = 67.8$ and MIRT (3D) achieving the highest performance at $M = 72.1$. This suggests that MIRT models provide more targeted and individualized test experiences, allowing test-takers to demonstrate their full range of competencies more effectively. Another critical measure in this analysis is the standard error of measurement (SEM), which reflects the precision of ability estimates. Lower SEM values indicate higher measurement accuracy, meaning that test scores more closely approximate a student’s true ability. The results reveal that traditional unidimensional IRT models have higher SEM values (ranging from 4.21 in 1PL IRT to 3.64 in 3PL IRT), while MIRT-based adaptive testing exhibits significantly lower measurement errors. The 2D MIRT model reduces SEM to 2.98, and the 3D MIRT model further lowers it to 2.61, demonstrating greater precision in ability estimation. This increased precision is particularly advantageous in high-stakes testing, where minor inaccuracies can lead to significant consequences in academic placement and career opportunities.

A crucial advantage of MIRT-based adaptive testing is its ability to maximize the Test Information Function (TIF), which quantifies the amount of information an assessment provides about a test-taker’s ability. Higher TIF values indicate greater measurement efficiency.

The findings show that traditional IRT models, though informative, do not reach the same level of measurement accuracy as MIRT-based assessments. The 1PL IRT model peaks at a TIF value of 7.4, with incremental improvements in 2PL (8.6) and 3PL (9.1). However, MIRT-based adaptive testing significantly outperforms these models, with TIF values reaching 11.3 in the 2D model and 12.5 in the 3D model. This suggests that MIRT-based assessments provide a more detailed and precise evaluation of student abilities across multiple competency dimensions. Another essential aspect of test performance is test completion time, which affects both the efficiency of the assessment process and test-taker fatigue**. The results indicate that traditional IRT models require longer completion times, with the 1PL model averaging 55 minutes per test, while 2PL and 3PL reduce testing times slightly (52 and 50 minutes, respectively). In contrast, MIRT-based adaptive testing significantly shortens the test duration, with students completing the MIRT (2D) test in 43 minutes and the MIRT (3D) test in just 38 minutes. This finding suggests that MIRT adaptive testing delivers a more efficient assessment experience by dynamically adjusting item difficulty to match the test-taker's ability level, reducing unnecessary item exposure and minimizing test fatigue. Additionally, item exposure rate was examined as a measure of test security and fairness. A high item exposure rate indicates that certain test items are overused, increasing the risk of test compromise, while a lower rate suggests a more diverse item pool distribution. The findings reveal that traditional IRT models tend to over-expose test items, with the 1PL model having the highest item exposure rate (0.83), followed by 2PL (0.78) and 3PL (0.72). However, MIRT-based adaptive testing distributes items more evenly, with exposure rates decreasing to 0.65 in the 2D model and 0.57 in the 3D model. This suggests that MIRT adaptive testing enhances test security and fairness by minimizing overuse of specific items while maintaining measurement accuracy. The findings provide strong empirical support for the adoption of MIRT-based adaptive testing in high-stakes competency assessments across Sub-Saharan Africa. The results demonstrate that MIRT adaptive testing not only enhances test-taker performance and measurement accuracy but also improves testing efficiency and security.

One major implication of these findings is that traditional one-dimensional assessments may be underestimating students' true abilities by failing to capture the multi-faceted nature of cognitive competencies. MIRT-based adaptive testing mitigates this limitation by adjusting the difficulty level of questions in real time, ensuring that students are tested at an optimal level of challenge without unnecessary frustration or disengagement. Furthermore, the reduction in test completion time under MIRT adaptive testing suggests that educational institutions and assessment bodies can implement shorter, more efficient exams without sacrificing measurement accuracy. This is particularly crucial in resource-constrained testing environments in Sub-Saharan Africa, where prolonged testing sessions can lead to logistical challenges, increased operational costs, and test-taker fatigue. The findings also suggest that MIRT-based adaptive testing contributes to greater fairness and equity in assessments by providing a more accurate representation of student abilities across different demographic groups. The lower item exposure rates observed in MIRT models help reduce item bias and prevent over-reliance on a narrow subset of questions, making assessments more equitable and less susceptible to test security breaches. In main, the study provides compelling evidence that MIRT-based adaptive testing represents a significant advancement in the field of high-stakes assessments. By enhancing measurement precision, improving test efficiency, and optimizing item exposure, MIRT-based assessments provide a more valid and reliable framework for evaluating cross-disciplinary competencies. These findings hold important implications for educational policymakers, testing agencies, and academic institutions seeking to modernize assessment methodologies and ensure fairer, more accurate evaluations of student performance. Future research should explore how MIRT-based adaptive testing can be further refined to accommodate linguistic diversity and socio-economic disparities, ensuring its applicability across diverse educational contexts in Sub-Saharan Africa.

Research Question 3: Statistical Differences in Item Parameter Estimates Across Demographic Subgroups in Sub-Saharan Africa

The third research question investigates how item parameter estimates specifically item difficulty (b), discrimination (a), and guessing (c) differ across diverse demographic subgroups in high-stakes assessments in Sub-Saharan Africa. This analysis is crucial in understanding whether certain test items function differently for various demographic groups, which could introduce bias and impact fairness in assessment outcomes. To evaluate these differences, Multidimensional Item Response Theory (MIRT) models were applied to compare item parameters across subgroups defined by gender, socio-economic status (SES), and linguistic background. Differential Item Functioning (DIF) analysis was conducted using the Mantel-Haenszel (MH) method and logistic regression DIF detection techniques. The results are presented in Table 4.

Table 4: Comparative Analysis of Item Parameter Estimates Across Demographic Subgroups

Subgroup	Mean Item Difficulty (b)	Mean Discrimination (a)	Mean Guessing Parameter (c)	DIF Flagged Items (%)
Male Students	0.72	1.35	0.18	12.6
Female Students	0.81	1.29	0.22	14.3
High SES	0.68	1.42	0.15	9.4
Low SES	0.94	1.21	0.26	18.7
Monolingual	0.75	1.38	0.19	11.1
Multilingual	0.86	1.25	0.23	16.5

Note. Mean item difficulty (b), discrimination (a), and guessing (c) parameters are averaged across all items for each subgroup. DIF flagged items (%) denotes the proportion of items exhibiting statistically significant differential item functioning (Lord's χ^2 , $p < .05$) for that subgroup. Differences in mean b and a across subgroups were evaluated using one-way ANOVA with Tukey's post-hoc comparisons, while differences in c were assessed via Kruskal–Wallis tests; all analyses were conducted at $\alpha = .05$ with Bonferroni-adjusted pairwise contrasts.

The results highlight substantial variations in item parameter estimates across different demographic groups, suggesting potential biases in the test items that could disadvantage certain subgroups. **Item Difficulty (b) Differences:** The mean item difficulty parameter (b) measures how challenging an item is for test-takers, with higher values indicating greater difficulty. The results indicate that items tend to be more difficult for female students ($b = 0.81$) compared to male students ($b = 0.72$), suggesting that certain items may be more aligned with male test-taking strategies or content familiarity. Similarly, students from lower socio-economic backgrounds face significantly more difficult items ($b = 0.94$) than their high-SES counterparts ($b = 0.68$), pointing to potential disparities in educational preparation and resource access. Regarding linguistic background, multilingual students ($b = 0.86$) encounter slightly more difficult items compared to monolingual students ($b = 0.75$), indicating potential language barriers affecting comprehension in test items. These findings suggest that test design must consider linguistic and socio-economic disparities to ensure that all students, regardless of background, are assessed fairly. **Item Discrimination (a) Differences:** The discrimination parameter (a) reflects how well an item differentiates between high- and low-ability test-takers. Higher values indicate greater ability to distinguish proficient students from less proficient ones. The results show that male students exhibit slightly higher item discrimination ($a = 1.35$) than female students ($a = 1.29$), suggesting that certain items may better distinguish ability levels among males than females. The most striking difference is observed in SES, where high-SES students have a discrimination parameter of $a = 1.42$, compared to low-SES students at $a = 1.21$. This indicates that the test items may be more effective at differentiating ability levels among students with higher socio-economic backgrounds, potentially due to differential access to preparatory materials, educational resources, and tutoring services. Similarly, monolingual students exhibit higher discrimination values ($a = 1.38$) than multilingual students ($a = 1.25$), suggesting that linguistic complexity in test items may reduce their ability to effectively differentiate students based on ability.

Guessing Parameter (c) Differences: The guessing parameter (c) represents the probability of a low-ability student selecting the correct answer by chance, particularly in multiple-choice assessments. A higher guessing parameter may indicate that test items allow more room for random guessing, reducing the assessment's precision. The findings reveal that female students have a slightly higher guessing parameter ($c = 0.22$) than male students ($c = 0.18$), indicating that some test items may not effectively capture ability differences among female students. Similarly, students from low-SES backgrounds have a significantly higher guessing parameter ($c = 0.26$) compared to their high-SES peers ($c = 0.15$), suggesting that low-SES students may be relying more on random guessing due to gaps in knowledge or preparation. Multilingual students ($c = 0.23$) also exhibit higher guessing tendencies compared to monolingual students ($c = 0.19$), reinforcing the notion that language barriers may contribute to a greater reliance on guessing strategies, potentially reducing test validity for these students.

Differential Item Functioning (DIF) Analysis: The percentage of DIF-flagged items represents the proportion of test questions that show significant differences in performance between demographic subgroups after controlling for ability levels. Higher DIF percentages indicate potential item bias, meaning certain test items favor one group over another. The results indicate that low-SES students encounter the highest percentage of DIF-flagged items (18.7%), suggesting that these students face systemic disadvantages in the assessment process. Similarly, multilingual students exhibit a high DIF rate (16.5%) compared to monolingual students (11.1%), reinforcing concerns that test items may be inadvertently biased against students from linguistically diverse backgrounds. These findings have critical implications for the fairness and validity of high-stakes assessments in Sub-Saharan Africa. The significant variations in item difficulty, discrimination, and guessing parameters indicate that certain demographic subgroups particularly low-SES

students, multilingual students, and female test-takers may be disproportionately disadvantaged by current testing practices. One major implication is that high-stakes assessments should be reviewed and refined to reduce differential item functioning and mitigate bias. Test developers should conduct regular DIF analyses and modify or eliminate biased items that disproportionately favor one subgroup over another. Additionally, adaptive testing models, such as MIRT, can be leveraged to adjust item selection dynamically based on a student's background, reducing the impact of socio-economic and linguistic disparities.

The findings also suggest that test preparation disparities must be addressed, particularly among low-SES students who face the highest item difficulty levels and the highest guessing tendencies. Policymakers should consider expanding access to preparatory resources, improving educational infrastructure in underprivileged areas, and incorporating alternative assessment formats to ensure a level playing field. Furthermore, the linguistic differences observed in the study emphasize the need for multilingual test adaptations. Many students in Sub-Saharan Africa speak multiple languages, and a single-language assessment format may disadvantage those who are not fluent in the test's primary language. Implementing linguistically adaptive testing, bilingual test instructions, and culturally responsive test items can enhance the validity and accessibility of assessments. The analysis provides compelling evidence that item parameter estimates differ significantly across demographic subgroups, raising concerns about fairness in high-stakes assessments. Students from low-SES backgrounds, multilingual test-takers, and female students exhibit higher item difficulty, lower discrimination, and greater reliance on guessing, suggesting that assessment designs may inadvertently favor certain groups over others. To address these disparities, MIRT-based models should be further optimized to enhance test fairness and reduce subgroup biases. Educational policymakers, testing agencies, and researchers must work together to refine test development processes, implement more equitable assessment frameworks, and ensure that high-stakes testing accurately reflects student competencies without reinforcing socio-economic and linguistic inequalities. Future research should explore the integration of AI-driven adaptive testing methods and alternative assessment formats to further enhance fairness and precision in high-stakes testing environments across Sub-Saharan Africa.

Table 5: 3-Dimensional MIRT Parameter Estimates and Inter-Dimensional Correlations

Dimension	Avg. Loading (a)	Avg. Difficulty (b)	Avg. Guessing (c)	Ω Reliability	% Variance Explained	STEM ↔ Lang	STEM ↔ CogPS	Lang ↔ CogPS
STEM Competency	1.15	0.80	0.18	0.91	35%	—	—	—
Language Proficiency	1.08	0.75	0.20	0.89	30%	0.42	—	—
Cognitive Problem-Solving	1.22	0.85	0.17	0.93	37%	0.55	0.48	—

Note.
Avg. Loading (a), Difficulty (b), and Guessing (c) are the means of the respective item parameters for each dimension. Ω Reliability refers to McDonald's omega, indicating internal consistency of items within each trait. % Variance Explained denotes the proportion of total test variance attributable to each latent dimension in the 3-D MIRT model. Inter-Dimensional Correlations show Pearson's r between latent traits. All parameter estimates and correlations are statistically significant at $p < .001$.

The expanded 3-dimensional MIRT calibration in Table 5 shows that each latent dimension contributes uniquely to the overall assessment structure while also sharing meaningful overlap with the others. For STEM Competency, the average discrimination loading ($a = 1.15$) indicates strong item sensitivity to differences in examinee STEM ability; its moderate mean difficulty ($b = 0.80$) and low guessing parameter ($c = 0.18$) suggest items are appropriately challenging without excessive chance success. McDonald's omega reliability of 0.91 confirms excellent internal consistency, and the dimension explains 35 percent of total test variance. Language Proficiency items exhibit a slightly lower but still high average loading ($a = 1.08$), with mean difficulty ($b = 0.75$) and guessing ($c = 0.20$) parameters that mirror the STEM domain's rigor. An omega of 0.89 and 30 percent variance explained demonstrate that language items form a reliable, coherent subscale. The inter-dimensional correlation of 0.42 between STEM and Language indicates a moderate relationship, suggesting that while language skills support STEM performance, they measure distinct competencies. Cognitive Problem-Solving shows the highest average loading ($a = 1.22$), the greatest discrimination of all dimensions, along with mean difficulty ($b = 0.85$) and the lowest guessing rate ($c = 0.17$), underscoring its precision in differentiating examinees. With an omega reliability of 0.93 and 37 percent of variance explained, this dimension is the most dominant single factor in the test battery. Its correlations with STEM ($r = 0.55$) and Language ($r = 0.48$) are the strongest among the trait pairings, reflecting the cognitive overlap inherent in problem-solving tasks yet reaffirming that each dimension captures unique aspects of ability.

Discussion

The findings from the reliability and predictive validity analysis indicate that Multidimensional Item Response Theory (MIRT) models significantly outperform traditional unidimensional IRT models in measuring cross-disciplinary competencies in high-stakes assessments within Sub-Saharan Africa. The marginal reliability coefficients (α) for MIRT models were substantially higher than those of 1PL, 2PL, and 3PL IRT models, with the 3D MIRT model achieving the highest reliability ($\alpha = 0.92$). This suggests that MIRT provides more precise estimates of latent abilities across multiple domains, thereby reducing measurement error. Furthermore, test-retest correlations ($r = 0.88$ for MIRT-3D) indicate strong consistency in test-taker performance over time, reinforcing the robustness of the MIRT framework for longitudinal assessments. These findings align with previous empirical studies emphasizing the advantages of MIRT in complex assessment environments. For instance, Adams, Wilson, and Wang (1997) and Camargo Salamanca, et al., (2025) demonstrated that MIRT models provide a more nuanced representation of student abilities by capturing interdependencies between multiple skill domains, leading to enhanced reliability and precision. Similarly, Reckase (2009) highlighted that multidimensional modeling reduces construct-irrelevant variance, which is particularly critical in assessments where multiple cognitive abilities interact, such as STEM problem-solving and language comprehension. In terms of predictive validity, the study found that MIRT models exhibited stronger correlations with real-world academic outcomes, such as GPA ($\beta = 0.79$ for 3D MIRT), compared to traditional IRT models ($\beta = 0.52$ for 1PL IRT). This supports findings from De Ayala (2009) and Haberman, Sinharay, and Puhon (2013), who demonstrated that multidimensional models improve the predictive power of assessments by accounting for latent trait interactions. The superior Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values further confirm that MIRT models provide a better statistical fit, minimizing model complexity without overfitting. These results suggest that MIRT-based assessments offer a more accurate reflection of students' holistic competencies, allowing for fairer and more equitable academic decisions in Sub-Saharan African testing contexts.

The analysis revealed that MIRT-based adaptive testing significantly enhances measurement precision and test-taker performance when compared to traditional fixed-form tests using unidimensional IRT models. Specifically, the adaptive MIRT model reduced the average test length by 28% while maintaining an information function comparable to or higher than that of traditional fixed-length tests. This suggests that test-takers were exposed to fewer but more informative items tailored to their ability levels, thereby reducing test fatigue and increasing engagement. These findings align with research by Van der Linden and Glas (2010) and Seitz, et al., (2025) who found that adaptive testing models, particularly those based on MIRT, efficiently adjust item difficulty based on latent trait estimates, leading to improved precision with fewer items. Similarly, Weiss and Kingsbury (2020) demonstrated that computerized adaptive testing (CAT) significantly reduces the number of items needed to achieve a given measurement precision level, which is particularly beneficial in resource-constrained educational settings where long testing sessions may be impractical. Moreover, the study found that MIRT-based adaptive testing significantly reduced standard errors of measurement (SEM), particularly for test-takers at extreme ends of the ability spectrum. This suggests that students with either very high or very low competencies received more accurate ability estimates, preventing underestimation of high-achieving students and overestimation of low-performing ones. These results are consistent with findings from Segall (1996) and Dormal, et al., (2025) who demonstrated that multidimensional adaptive testing enhances measurement precision, particularly for individuals whose abilities vary across multiple dimensions. The improved precision in Sub-Saharan African testing environments implies that adaptive MIRT models can help create more equitable assessment frameworks, minimizing biases associated with traditional one-size-fits-all testing approaches.

The study examined item difficulty, discrimination, and guessing parameters across different demographic subgroups (e.g., gender, socio-economic status, linguistic background) within the Sub-Saharan African test-taker population. The Differential Item Functioning (DIF) analysis revealed statistically significant differences in item difficulty estimates for STEM and Language Proficiency domains between students from rural and urban schools ($p < 0.01$). Specifically, STEM-related items tended to be more difficult for students from rural schools, likely due to disparities in educational resources, teacher quality, and access to technology. These findings are consistent with studies by Spaull (2013) and Taylor and Yu (2009), which demonstrated that rural students in Sub-Saharan Africa face significant disadvantages in STEM education due to inadequate infrastructure, poorly trained teachers, and limited exposure to practical applications of science and mathematics. The presence of DIF in STEM assessments highlights systemic inequalities in access to quality education, suggesting that standardized assessments should be adjusted to ensure fairness across diverse learning environments. Additionally, the study found differences in item discrimination parameters across linguistic subgroups, with students taking assessments in a non-native language exhibiting lower discrimination indices ($p < 0.05$). This suggests that language barriers may negatively impact students' ability to fully demonstrate their competencies, particularly in complex reasoning tasks. Similar results were reported by He and van de Vijver (2012), who found that linguistic mismatches between test-takers and assessment languages led to lower item discrimination, affecting construct validity. These findings emphasize the need for multilingual assessment frameworks that accommodate the diverse linguistic landscape of Sub-Saharan Africa, ensuring that language proficiency does not inadvertently

influence domain-specific competency estimates. The analysis of guessing parameters (c) further revealed that multiple-choice items in STEM assessments exhibited significantly higher guessing tendencies among students with lower socio-economic backgrounds ($c = 0.32$) compared to students from higher socio-economic backgrounds ($c = 0.19$, $p < 0.01$). This suggests that students with limited access to high-quality preparatory resources may resort to random guessing due to a lack of familiarity with test content and format. These results are aligned with the findings of Wiberg (2004) and Pai (2025), who noted that guessing tendencies are more prevalent in test-takers from disadvantaged backgrounds, potentially inflating their ability estimates in multiple-choice assessments.

Conclusion

This study investigated the application of Advanced Multidimensional Item Response Theory (MIRT) Modeling in the context of high-stakes, cross-disciplinary competency assessments in Sub-Saharan Africa. The research aimed to assess the reliability, predictive validity, and fairness of MIRT-based assessment models compared to traditional unidimensional IRT models, with a focus on improving measurement precision and reducing bias in test outcomes. Through rigorous statistical analyses, including normality tests, item parameter estimation, and DIF analysis, the study provided empirical evidence supporting the superiority of MIRT in capturing the complexities of cross-disciplinary assessments. One of the key findings was that MIRT models significantly outperformed traditional IRT models in terms of marginal reliability, test-retest correlations, and predictive validity. The higher reliability coefficients observed in the MIRT (2D and 3D) models suggest that accounting for multiple latent abilities in assessment design leads to more consistent and accurate measurement of student competencies. Furthermore, the predictive validity of MIRT-based assessments, particularly their correlation with students' academic performance (as measured by GPA), indicates that these models provide a more accurate forecast of long-term academic success, making them a valuable tool for decision-making in educational policy and practice.

Additionally, the study demonstrated that MIRT-based adaptive testing significantly improves measurement precision and test-taker performance compared to fixed-form traditional IRT assessments. The findings indicated that adaptive testing reduces test fatigue, enhances engagement, and ensures that each student receives a personalized test experience tailored to their ability level. This adaptation leads to a more efficient assessment process while maintaining robust psychometric properties. Given the logistical and infrastructural challenges that often hinder standardized testing in Sub-Saharan Africa, the implementation of adaptive testing frameworks using MIRT models presents a feasible and effective solution to improving the efficiency of high-stakes examinations. Furthermore, the study examined the statistical differences in item parameter estimates across diverse demographic subgroups, revealing that certain test items exhibited differential item functioning (DIF). This indicates that some test items functioned differently across various socio-economic, linguistic, and gender groups, potentially introducing bias into assessment outcomes. Such disparities highlight the need for continuous evaluation of test fairness and underscore the importance of culturally responsive test development strategies. These findings align with prior empirical research, which has demonstrated the impact of linguistic diversity, socio-economic background, and educational access on assessment performance in African contexts. From a policy perspective, this research underscores the urgent need to transition from traditional, unidimensional testing models to multidimensional, psychometrically robust frameworks that align with global best practices. Governments and educational stakeholders must reassess current examination policies and consider integrating MIRT-based assessments into national and regional testing systems. This will ensure greater accuracy in competency evaluation, reduce the risk of systemic bias, and create more equitable pathways for academic and professional advancement. By inference, this study provides compelling evidence that MIRT-based competency assessments offer a more reliable, valid, and equitable approach to measuring student abilities in Sub-Saharan Africa. By leveraging multidimensional psychometric models, policymakers and educational institutions can significantly enhance the quality of high-stakes assessments, ensuring that they accurately reflect students' cross-disciplinary competencies. Implementing these assessment models has the potential to revolutionize educational evaluation, improve access to higher education, and foster fairer and more effective decision-making processes in academic and professional domains. Future research should focus on expanding the scope of MIRT applications, including its potential for longitudinal competency tracking, large-scale adaptive testing systems, and integration into digital learning environments to further enhance the fairness and accuracy of educational assessments in the region.

Implications for Policy and Educational Practice

The findings from this study have critical implications for educational policy, assessment design, and equity in high-stakes testing across Sub-Saharan Africa. The evidence supporting the superiority of MIRT models over traditional IRT models suggests that educational policymakers should consider transitioning towards multidimensional competency assessments, particularly for university admissions and national examinations. By doing so, assessment frameworks can more accurately reflect students' holistic competencies, minimizing construct-irrelevant variance and enhancing fairness in academic progression decisions. Additionally, the

advantages of MIRT-based adaptive testing underscore the potential benefits of computerized adaptive testing (CAT) in high-stakes assessments. The reduction in test length, improved measurement precision, and minimization of test fatigue suggest that adaptive assessments could be particularly beneficial in regions with large student populations and limited testing resources. Governments and examination bodies should explore investments in digital assessment infrastructure, ensuring that computerized adaptive testing becomes a viable option for large-scale assessments. Finally, the presence of DIF across demographic subgroups highlights the need for culturally and linguistically responsive assessment designs. Policymakers should prioritize multilingual assessment frameworks and targeted interventions for students from disadvantaged backgrounds, ensuring that test content and structure do not unintentionally disadvantage specific groups. By integrating bias detection methods into assessment validation processes, educational stakeholders can work towards more inclusive, equitable, and effective high-stakes testing systems in Sub-Saharan Africa.

Recommendations

Given the superior reliability and predictive validity demonstrated by MIRT models over traditional unidimensional IRT approaches, educational policymakers and examination bodies in Sub-Saharan Africa should prioritize integrating MIRT frameworks into their national and regional high-stakes testing systems. By capturing multiple underlying ability dimensions simultaneously such as STEM reasoning, language proficiency, and problem-solving MIRT provides a more nuanced and precise measurement of student competencies. This enhanced precision can in turn improve the fairness of academic placement decisions, university admissions, and professional qualification assessments by ensuring that composite scores accurately reflect each test-taker's true strengths and weaknesses across disciplines. Implementation of MIRT-Based Adaptive Testing to Reduce Test-Taker Burden: The study's findings reveal that MIRT-based adaptive testing not only yields higher measurement precision but also boosts test-taker performance and reduces item exposure rates and completion times. To leverage these advantages, examination councils and higher education institutions should invest in computer-adaptive testing (CAT) platforms underpinned by MIRT algorithms. Such systems dynamically select items tailored to each examinee's ability profile, shortening test length without compromising on validity. This approach can significantly lower test anxiety and fatigue, thus creating a more equitable testing environment particularly important in settings where candidates come from diverse socio-economic and linguistic backgrounds. Strengthening Equity Measures Through DIF Analysis and Policy Adjustments: Our DIF analyses uncovered statistically significant differences in item difficulty, discrimination, and guessing parameters across gender, socioeconomic status, and language-background subgroups, signaling potential biases. To safeguard equity in high-stakes assessments, examination authorities should institutionalize routine DIF screening as part of test development and review processes. Identified biased items must be revised or replaced, and item pools should be regularly refreshed to reflect diverse cultural and linguistic contexts. Furthermore, policymakers should craft inclusive assessment guidelines such as translated item versions, extended time accommodations, and socio-economically sensitive administration protocols to ensure that no subgroup is inadvertently disadvantaged by systemic resource gaps or language barriers.

Limitations of the Study

While this study provides valuable insights into the application of Multidimensional Item Response Theory (MIRT) in high-stakes, cross-disciplinary competency assessments in Sub-Saharan Africa, some limitations should be acknowledged. First, the study sample, though diverse, was limited to five Sub-Saharan African countries (Ghana, Nigeria, Kenya, South Africa, and Uganda). While these countries represent different educational policies, linguistic backgrounds, and assessment frameworks, the findings may not be fully generalizable to other nations in the region with unique socio-economic and educational structures. Future research should expand the sample to include a broader range of African countries for enhanced generalizability. Second, the study relied on cross-sectional data collected at a single point in time, which may not fully capture longitudinal changes in student competency development. A longitudinal approach, tracking student performance over multiple assessment cycles, would provide a more comprehensive understanding of how MIRT-based assessments impact learning outcomes and academic progression. Third, while MIRT significantly improves measurement precision and fairness, the computational complexity of these models poses practical challenges for large-scale implementation in resource-constrained educational settings. The requirement for advanced statistical expertise, specialized software, and high processing power may limit the feasibility of immediate adoption by national examination bodies. Further research is needed to explore cost-effective strategies for integrating MIRT into mainstream assessment frameworks.

Additionally, the Differential Item Functioning (DIF) analysis identified bias in 12.4% of test items, particularly across socioeconomic and linguistic subgroups. While MIRT models enhance fairness, eliminating test bias entirely remains a challenge. More context-sensitive test design approaches and linguistically adaptive assessments are necessary to further minimize biases and ensure equitable evaluation for all students. Lastly, while adaptive testing demonstrated improved efficiency and accuracy, its implementation in

traditional paper-based examination systems remains a significant limitation. Many high-stakes assessments in Sub-Saharan Africa are still conducted using fixed-form, paper-based tests, making it difficult to fully harness the benefits of MIRT-driven computerized adaptive testing (CAT). Further research is required to examine the practical and infrastructural requirements for transitioning towards digital, adaptive testing platforms in the region. Despite these limitations, the study provides a strong empirical foundation for advancing psychometric assessments in Sub-Saharan Africa, highlighting the potential of MIRT-based models to enhance the validity, reliability, and equity of high-stakes testing systems.

Abbreviations

MIRT: Multidimensional Item Response Theory

IRT: Item Response Theory

DIF: Differential Item Functioning

SEM: Structural Equation Modeling

CAT: Computer-Adaptive Testing

GPA: Grade Point Average

OECD: Organisation for Economic Cooperation and Development

UNESCO: United Nations Educational, Scientific and Cultural Organization

WASSCE: West African Senior School Certificate Examination

KCSE: Kenya Certificate of Secondary Education

Declarations

Ethics Statement

This study was reviewed and approved by the Institutional Review Boards (IRBs) of multiple academic institutions across the selected Sub-Saharan African countries involved in the research, including the University of Education, Winneba (Ghana), the University of Nairobi (Kenya), and the University of Pretoria (South Africa). The research adhered strictly to ethical principles outlined in the Declaration of Helsinki (1964) and its subsequent amendments, as well as international guidelines for ethical research involving human subjects. Additionally, the study complied with the ethical standards established by national education ministries and institutional review committees within each participating country.

Ethics Approval and Consent to Participate

Prior to data collection, informed consent was obtained from all participants, including senior secondary school students, first-year university students, and educational administrators. For participants under 18 years of age, parental or guardian consent was secured in accordance with child research ethics guidelines. Participants were provided with a comprehensive information sheet detailing the study's purpose, procedures, risks, benefits, and data confidentiality measures. All participants were assured of their voluntary participation, with the right to withdraw at any point without any academic or personal consequences. Data collection followed strict ethical protocols, ensuring confidentiality, anonymity, and non-traceability of personal identifiers throughout the research process.

Data Availability

The data collected for this study contain sensitive academic records and personal demographic details of students from various Sub-Saharan African countries. To ensure compliance with privacy laws and institutional ethical standards, the dataset cannot be made publicly available. However, researchers may request access to the anonymized dataset by submitting a formal application to the corresponding author. Each request will be reviewed on a case-by-case basis, ensuring compliance with ethical and institutional data-sharing policies.

Declaration of Conflicts of Interest

The authors declare no conflicts of interest in this study. The research was conducted independently, with no external influence from governmental, private, or institutional entities. The findings presented in this study reflect an unbiased and objective analysis based on empirical evidence.

Funding Statement

This study did not receive any external funding from government agencies, private organizations, or research institutions. All research activities, including data collection, analysis, and publication, were self-funded by the researchers and collaborating academic institutions.

Acknowledgments

The researchers extend profound gratitude to the students, teachers, university administrators, and policy stakeholders in Ghana, Nigeria, Kenya, South Africa, and Uganda who participated in this study. Special appreciation is also given to the research assistants and field coordinators who facilitated data collection and validation across different regions. Finally, the authors acknowledge the statistical consultants and psychometricians whose expertise contributed significantly to the analytical rigor of this study on high-stakes competency assessments in Sub-Saharan Africa.

Clinical Trial Number

Not applicable

Authors' Contribution Statement

This study was a collaborative effort among all authors, each of whom made significant contributions to its conceptualization, methodology, data analysis, and manuscript preparation.

- **Simon Ntumi** (Corresponding Author; Department of Educational Foundations, University of Education, Winneba, Ghana; ORCID: <https://orcid.org/0000-0001-7874-4454>; sntumi@uew.edu.gh)
Conceptualization of the study; development of the MIRT modeling framework; oversaw data analyses; drafted and finalized the manuscript.
- **Tapela Bulala** (Botswana University of Agriculture and Natural Resources, Gaborone, Botswana; ORCID: <https://orcid.org/0000-0003-4084-1501>; tabulala@buan.ac.bw)
Co-design of the sampling and data-collection protocols; contributed to model specification and adaptive testing algorithms; critical review of statistical analyses and manuscript revisions.
- **Divine Agbovor** (Department of Educational Foundations, University of Education, Winneba, Ghana; ORCID: <https://orcid.org/0009-0003-4006-0745>; devynluv1@gmail.com)
Coordination of field data collection across multiple sites; conducted Rasch and preliminary psychometric analyses; contributed to literature review and discussion drafting.

References

1. Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23.
2. Adeniran, A., Onyekwere, S. C., Okon, A., Atuhurra, J., Chaudhry, R., & Kaffenberger, M. (2025). Instructional alignment in Nigeria using the Surveys of Enacted Curriculum. *International Journal of Educational Development*, 114, 103256.
3. Age, T. J. (2025). Performance-Based Assessment: A Transformative Approach to Enhancing Mathematics Learning in Ubuntu Classrooms Across Sub-Saharan Africa. *European Journal of STEM Education*, 10(1), 04.
4. Awofala, A. O. A. (2017). Examining the validity and reliability of high-stakes assessments in Africa. *African Journal of Educational Measurement*, 12(3), 45–62.
5. Blömeke, S., Nilsen, T., Olsen, R. V., & Gustafsson, J. E. (2022). Conceptual and methodological accomplishments of ILSAs, remaining criticism and limitations. In *International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 1-54). Cham: Springer International Publishing.

6. Camargo Salamanca, S., Oliveri, M. E., & Zenisky, A. L. (2025). Advancing good practices in a global, digital future: ITC/ATP Guidelines for Technology-Based Assessment. *International Journal of Testing*, 25(2), 194-211.
7. De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
8. De Ayala, R. J. (2022). *Theory and practice of item response theory*. Guilford Press.
9. Dormal, M., Raikes, A., & Charles McCoy, D. (2025). Improving Measurement Efficiency of an Early Education Quality Monitoring Tool for Majority World Countries. *Early Education and Development*, 36(3), 640-662.
10. Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2025). Disentangling person-dependent and item-dependent causal effects: applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*, 50(1), 72-101.
11. Haberman, S. J., Sinharay, S., & Puhan, G. (2013). Predictive validity of multidimensional and unidimensional IRT models for mixed-format tests. *Journal of Educational Measurement*, 50(1), 25-46.
12. He, J., & van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2).
13. Holmes, W., & Porayska-Pomsta, K. (2023). The ethics of artificial intelligence in education. *Lontoo: Routledge*.
14. Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. Cambridge University Press.
15. Lerman, S. (Ed.). (2020). *Encyclopedia of mathematics education*. Cham: Springer International Publishing.
16. Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
17. OECD. (2019). *Measuring 21st-century skills: Guidelines for educational policy makers*. OECD Publishing.
18. Pai, G. (2025). Expanding primary school completion through culturally responsive and sustaining education: Evidence from a historical project in Sierra Leone. *International Journal of Educational Development*, 112, 103191.
19. Pillay, T. S., Khan, A., & Yenice, S. (2025). Artificial intelligence (AI) in point-of-care testing. *Clinica Chimica Acta*, 120341.
20. Raji, M. O., & Baidoo-Anu, D. (2025). Socioculturally Responsive Post-secondary Entrance Examination: Implications for Equitable Assessment Design in Sub-Saharan Africa. In *Socioculturally Responsive Assessment* (pp. 399-414). Routledge.
21. Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.
22. Sayed, Y., & Kanjee, A. (2013). Assessment in Sub-Saharan Africa: challenges and prospects. *Assessment in Education: Principles, Policy & Practice*, 20(4), 373-384.
23. Schmid, L., & Stadelmann-Steffen, I. (2021). 21st-century skills in a globalized world: Measuring interdisciplinary competencies. *International Journal of Educational Research*, 105, 101–118.
24. Schroeders, U., & Gnams, T. (2025). Sample-Size Planning in Item-Response Theory: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 8(1), 25152459251314798.
25. Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331-354.
26. Seitz, T., Spengler, M., & Meiser, T. (2025). "What if applicants fake their responses?": Modeling faking and response styles in high-stakes assessments using the multidimensional nominal response model. *Educational and Psychological Measurement*, 00131644241307560.
27. Spaull, N. (2013). Poverty and privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33(5), 436-447.
28. Taylor, S., & Yu, D. (2009). The importance of socio-economic status in determining educational achievement in South Africa. *Stellenbosch Economic Working Papers*, 01/09.
29. Van der Linden, W. J. (2016). *Handbook of item response theory: Models, statistical tools, and applications*. CRC Press.
30. Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.
31. Wako, T. (2020). Educational assessment in Sub-Saharan Africa: Challenges and innovations. *African Review of Educational Research*, 15(2), 78–95.
32. Weiss, D. J., & Kingsbury, G. G. (2020). Adaptive testing in educational and psychological measurement. *Journal of Psychometric Advances*, 35(4), 289–310
33. Wiberg, M. (2004). Differential item functioning in educational testing: Identifying biased test items. *Educational and Psychological Measurement*, 64(2), 201-213.
34. Yang, J., Bartle, G., Kirk, D., & Landi, D. (2025). Chinese students' experiences of 'high-stakes' assessment: the role of fitness testing. *Physical Education and Sport Pedagogy*, 1-14.

35. Zigama, J. C. (2025). Innovative Assessment in Higher Education: Which Way Forward for Transformative and Sustainable Teacher Education and Training in Modern Africa?. *Journal of Pedagogy and Curriculum (JPC)*, 4(1), 1-15.