UEJTL: Universal Education Journal of Teaching and Learning
E-ISSN: 3047-8235
2026, Vol. 3, No. 3, pp. 324~336
DOI: https://doi.org/10.63081/uejtl.v3i3.212

Original Research

Assessment of Model Fit and Item Invariance of Mock Mathematics Multiple-Choice Items in Nigeria: A Quest for Quality Test Items

Owor Effiom Owor
Department of Business Administration, Faculty of Business Management, Federal Polytechnic, Ugep, Calabar, Nigeria

Article Info

Article history:

Received 02 07, 2026
Revised 04 25, 2026
Accepted 11 12, 2026

Keywords:

Model fit
Test items
Differential item functioning
Chi-square goodness of fit
Item invariance
Mathematics
Quality test items
Multiple-choice items
Urban rural

ABSTRACT

The dearth of test experts in Nigerian's post-primary schools who have the knowledge and skills to write worthwhile multiple-choice items in different school subjects is a worrisome development to teachers, students, policymakers, government, and society. This development has led to imprecise placement and diagnostic decisions regarding students' true abilities. This problem prompted this study to assess the model fit and item invariance of 2025 mock mathematics multiple-choice test items in Cross River State, Nigeria. Two research questions were posed for the study. The 2025 mock mathematics multiple-choice item instrument was adapted and administered to a sample of 1015 senior secondary school students drawn from 48 public and private secondary schools in Cross River State. The items were analyzed using Bilog-Mg-3 software to determine chi-square goodness-of-fit and differential item functioning (DIF) using the 1PL, 2PL, and 3PL Item Response Theory (IRT) models. The results revealed that 10 items representing 20% under the 3-parameter logistic models showed significant misfit, indicative of bad items that cannot be trusted for measurement. These items were required to be dropped from the test or revised. The results also revealed that 30 items (60%) had no meaningful DIF, 11 items (12%) were harder for rural students than urban students, and three (3) items (6%) were harder for urban students than rural students, while six (6) items (12%) exhibited no DIF between the two groups of students. It was, therefore, recommended that, among other things, test developers carry out rigorous preliminary testing by administering the items to an illustrative sample before operational use.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Owor Effiom Owor
Department of Business Administration, Faculty of Business Management, Federal Polytechnic, Ugep, Calabar, Nigeria
Email: oworeffim@gmail.com

1. INTRODUCTION

The standard for assessment instruments for test purposes in Nigeria and globally, and the need to provide worthwhile test items for measuring students' overall abilities, has been a major concern for teachers, researchers, stakeholders, practitioners, and policymakers. To date, the method of determining students' abilities in a test is still dominated by the use of classical test theory, despite its shortcomings in revealing the true abilities of examinees in a particular test. Measuring students' ability using this method would only produce global/total scores, which, if applied in decision-making, would result in inaccurate selection, placement, or referral.

Measurement of students' ability in a test for decision-making purposes is a function of valid, reliable, and fair test items developed by a trained classroom teacher or a test expert with sufficient psychometric skills in designing fit items in such an instrument. Measurement, according to Dalati (2018) is defined in terms of the relationship to related factors and the illustration of variables and types of data provided, differentiating between scientific terms, including concepts, constructs, and operational definitions as the procedure for obtaining a mathematical explanation of the extent to which a person or object possesses a particular trait. Some of the fundamental ideologies of measurement in educational assessment are raw scores, derived scores, standard scores, etc., which are indirect, unobservable, and can be inferred. It is the analysis of data such as scores obtained from educational assessments to surmise the proficiencies and abilities of learners, with a focus on whether such tests are valid and reliable.

This inference of students' abilities can be achieved by the administration of a test. By definition, a test is a tool that calls for learners to complete an exercise by responding to a standard set of questions or tasks. It is intended to appraise the examinee's knowledge, skills, and ability in a particular subject matter (Johanns et al., 2017). These tests could be administered informally by a classroom teacher or formally by examination bodies such as the West African Examinations Council (WAEC), National Examination Council (NECO), and so on. The main thrust of these bodies centers on providing quality, reliable, and fair test items to examinees in a cost-effective and conducive atmosphere, as well as comparing tests that meet both the rationale and provision articulated by psychometricians based on standardization. Onyibe et al. (2015) noted that the primary challenge of any examination body is its duty to produce a well-calibrated instrument that is sensitive to the indicators of the psychological qualities which make up the objectives to be measured. This is equally supported by Siri and Freddana (2011), who noted that the results of a test may be misleading if the items in a test instrument do not meet the psychometric requirements.

Fischer (2021), while investigating the consequences of limited item pools and model misfit, noted that writing quality test items in diverse school subjects for measuring students' abilities is challenging, which manifests in inadequate item-writing skills, qualifications, and training of teachers, educational institutions, and examination bodies. These inadequacies, according to the authors, have resulted in making faulty decisions regarding students' abilities. Assessing students using the global/total scores may hide the weaknesses of individual items that make up the test, highlighting the need for item analysis of objective-format tests.

The judgment of students' abilities in Nigeria and other African countries based on the global/total score in a particular examination is rather misleading and does not give a true picture of students' ability strengths based on the specific items responded to in such a test. When items in objective-format tests are arranged in ascending order of difficulty by item developers/writers, it is expected that examinees would respond correctly to test items until a point when his/her ability diminishes. For instance, when a test has 1-10 items, it is expected that such an examinee must not respond correctly to items 6-10 when he cannot provide correct responses to items 1-5. If this occurs, it therefore implies that there is a violation of the consistency of examinees' response patterns in a test, which could be the manifestation of the incompetence of test developers or item writers during the process of item construction. Practically, examinees' responses in a test can be predicted by defining their traits, estimating scores for the examinee's traits, and using the scores to predict test performance. That is, the consistency examinees exhibit in their mode of response to independent test items arranged in order of difficulty and their abilities should be included among the parameters for judging their performance, and not the total/global scores gained after attending to the entire test.

Objective-format test items are fundamental to the success of education at all levels because of their objectivity in scoring, efficiency, and responsiveness to current psychometric models. However, the quality of a test is dependent on the proficiencies of item developers in writing worth-while, quality, and fit items capable of measuring ability invariance across various student groups.

Item misfit, according to the Item Response Theory (IRT) and Rasch modelling paradigm, designates items whose response patterns depart from expectations under the selected model, such as the 1PL (Rasch), 2PL, and 3PL. The 1PL (Rasch) refers to the examinee's likelihood of answering a test item correctly depends solely on the item difficulty (b). This parameter states that the success or failure of examinees on an item is caused solely by how difficult the item is relative to their abilities. In the 2PL model, it assumes that the possibility of answering a test item correctly depends on both the item difficulty (b) parameter and the discrimination (a)parameter. This model states that some items are good at distinguishing high-ability examinees from low-ability examinees. Lastly, the 3PL model states that the probability of answering a test item correctly depends on the item difficulty (b) parameter, discrimination (a) parameter, and guessing (c). This model assumes that even low-ability examinees can occasionally give a correct answer to an item by guessing. Bond (2015) noted that fit statistics such as infit/outfit mean-squares, item-fit chi-square, among others, are supportive but sensitive to sample size, multidimensionality, local dependence, and model misspecification. According to Zhao and Hambleton (2017), IRT model misfit can affect item equating, item scaling, and classification, especially in mixed-format tests and when scaling options are sub-optimal. Misfit,

according to the authors, violates standard item-writing rules and threatens the validity, reliability, and fairness of the items, thereby impeding students' cognitive ability precision and learning outcomes.

Item misfit/non-fit refers to the inconsistency between observed item responses and model-predicted probabilities. According to Fährmann et al. (2022), item misfit should be interpreted not only as a statistical deviation but also as an indication of construct-related problems, such as multidimensionality or poorly designed items by item writers. This viewpoint reflects a shift in recent psychometric research toward assessing the applied significance of misfit, rather than relying solely on statistical implication. Von Davier and Bezirhan (2023) argue that traditional fit statistics may fail to detect complex forms of misfit in large-scale assessment data, especially in computer-based test situations where response behaviors are more dynamic. Their findings highlight the need for robust and context-sensitive approaches to identifying misfitting items. Schmucker and Moore (2025) in their study noted that item-writing errors are a key source of misfit. The authors stressed that several item-writing mistakes, such as implausible distractors, unclear wording, and negative phrasing, are associated with misfit. These flaws, according to the authors, reduce item discrimination and increase unpredictability in responses.

Measurement precision in IRT is represented by the test information function, which reflects the amount of information a test provides at different ability levels. Misfitting items contribute less meaningful information and introduce random error, resulting in increased standard errors of measurement. Fährmann et al. (2022) emphasize that misfitting items introduce construct-irrelevant variance, which decreases the total accuracy of measurement. This reduction in precision is challenging in high-stakes testing, where insignificant disparities in student scores can have substantial consequences. Item parameter estimation is fundamental to IRT-based studies. Misfitting items alter parameter estimations, leading to erroneous difficulty levels, defective discrimination statistics, and false guessing parameters.

Köhler and Hartig (2017) argued that item misfit should only be considered significant if it leads to real changes in test outcomes. They advised that teachers and test developers, including researchers, should shift the focus from traditional statistical item fit analyses - which regularly flag many items as misfitting in large populations- to assessing the practical consequences of item misfit. The authors further proposed that item misfit is practically important if the projected size of key statistical relationships, such as R-squared values, significantly changes when misfitting items are included in the statistical model versus when they are excluded. They highlighted that the practice of eliminating misfitting items is rash when the misfit is not important to the overall test outcome, particularly given the continuing argument about the validity and cut-off scores of numerous item fit statistics. In corroboration, Fährmann et al. (2022) contend that this approach highlights the real-world effect of misfit items on decision-making in pass rates and response consistency, rather than relying exclusively on p-values or single-fit thresholds. Some misfitting items reveal insignificant practical impact and might be retained for content coverage or equilibrium representation; others that implicitly shift outcomes require revision or removal. This realistic, impact-focused posture offers defensible decision rules for effective testing programmes, which are essential in making effective decisions regarding students' abilities in a test. As observed, most classroom teachers ignore the processes of item development, including planning and development, item writing, validation, deployment, and analysis. Some teachers are not interested in determining whether items are fit for testing, whether such items add to the quality of measurement, or whether the items are capable of differentiating the abilities of students who attend schools in town areas and those from rural areas.

Item Response Theory (IRT) experts rely on fit indices such as infit/outfit mean-squares, item-fit chi-squares, but researchers caution that fit statistics are dependent on sample size, scoring model, local dependence, and person-estimate bias. Müller (2020) highlights the complications associated with unconditional person-estimate-based statistics and recommends numerous conditional diagnostics. The author cautioned that item development experts should use numerous complementary diagnostic statistics such as classical item statistics, local dependence tests, model fit, and residual analyses, instead of only one threshold to make item decisions. Misfit items are those with keyed distractors, unclear stems, and double negatives. These items alter the cognitive demand of examinees and may measure unintended skills, rather than the target construct, thereby undermining content and construct validity capable of altering the validity of the entire test (Zhao & Hambleton, 2017).

The primary goal of IRT is the accurate estimation of examinee ability (θ). However, when items do not fit the model, which may be caused by a lack of psychometric skills among item developers, the ability estimates become biased and unstable. Fährmann et al. (2022) found that misfitting items significantly modify the relationship between latent traits and observed scores, leading to reduced interpretability of ability estimates. In support of the above, Zhao (2018) demonstrated that item misfit can hamper severity classification and decision consistency, mostly in applied environments where decisions involving pass or fail categorization are made. These findings show that the presence of non-fitting multiple-choice items can lead to misclassification of testees, thereby undermining the fairness, accuracy, and quality of assessments. In practice,

many multiple-choice items fail to conform to IRT model assumptions, leading to what is usually known as item misfit. Recent research has emphasized that item misfit is not only a statistical anomaly but a practical menace to measurement validity and fairness (Fährmann et al., 2022). When items do not fit any of the IRT models, the effect spreads beyond parameter estimation errors to include misrepresentations in ability measurement, compromised test validity, and flawed decisions based on test scores.

Similarly, Mullis et al. (2010) also conducted a study on the exploration of item model fit for items in the general reading passage for the Organization for Economic Cooperation and Development. The study aimed to detect passage-related local item dependencies in PIRLS 2001 cognitive data and approximate their association with national performance. The analysis was conducted using the three-parameter models. The item parameters were estimated on a dataset that made up 500 examinees randomly selected from the 23 PIRLS national samples. The results of the analysis revealed that ignoring the use of the three-parameter model in item construction could have a grave penalty for the goodness of fit of a model, for the parameter approximation and confidence levels. The finding revealed that the answer to any item is unconnected to any other item when the trait level is restricted. The items may be exceedingly intercorrelated in the entire sample; nevertheless, if the ability level is restricted, local independence signifies that no association remains between the items.

This signifies that examinees' achievement or failure on any item in a test is influenced by the likelihood of negligence in item construction without recourse to model fit considerations. A study by Jones (2019), on the differential item functioning and its relevance to measurement fairness, noted that if objective test items are not modeled to any of the parameters, the discrimination parameters of the mutually dependent items will be undervalued, stressing that the discrimination parameter and guessing parameter depend on the difficulty of the item it is intermingled with, but not on the difficulty of the item itself. Due to the effect of the difficulty parameter on the discrimination parameter, the negative local independence shrinks the item information, and the standard error of measurement is underestimated. No single model-data fit index detects all of the possible sources of fit or misfit.

In another study by Maric et al. (2023), a systematic literature review of psychometric trends in STEM education proposed a new model by incorporating probabilistic logic into longitudinal diagnostics modeling to provide refined learning tracking with longitudinal diagnostics assessment. It uses probabilistic features instead of binary features to model the latent variables that affect students' performance. The data were analyzed with item response theory statistical software. The results revealed that when differentiating student levels, the diagnostic results of the proposed model and the conventional longitudinal diagnostic model for binary characteristics displayed a high degree of consistency and quality model-data fit of the items.

Despite the above claims, Anagbogu and Owor (2021) found that the causes of students' inability to respond correctly to items in an instrument could not solely determine item misfit or invariance, but may be attributed to other factors such as teachers' commitment to covering syllabus content, school facilities, or class size.

In support of the above assertion, Anagbogu et al. (2021), in their study on assessment of social media immersion and undergraduate students' academic performance, revealed that students' response inconsistencies in an examination may be a result of their overdependence and reckless use of social media platforms such as WhatsApp, Facebook, Instagram, 2go, MySpace, among others. This result is in tandem with Bisong et al. (2021), who argued that item response inconsistencies by students are not only attributed to test experts' inability to write quality items, but could also be caused by students' personal variables such as their gender and age. They noted that some students may not be mature enough to comprehend some constructs being measured as a result of their age. This contention was also affirmed by Anagbogu et al. (2021) in their study of a path analytical model on factors with the implementation of civic education curriculum among teachers in Cross River State, Nigeria, who noted that factors such as lack of teacher supervision, teachers' knowledge of subject matter and teachers' attitude, and the lack of implementation of subject curriculum could promote item misfit and invariance among school students. Item misfit affects the consistency of classification decisions, thereby weakening the relationship between test scores and external criteria (Zhao, 2018). This position is supported by Fährmann et al. (2022), who argue that item misfit is frequently associated with construct underrepresentation, both of which weaken the theoretical foundation of the test. As a result, test scores may not accurately reflect the intended latent trait. The author further noted that misfitting anchor items are particularly problematic because they propagate errors throughout the equating process. This compromises the comparability of scores across administrations.

Differential Item Functioning (DIF) happens when test-takers from diverse groups with the equivalent ability have different likelihoods of responding to an item correctly. The current study sees DIF as an abuse of measurement invariance, which is vital for fair assessment. DIF analysis assumes measurement invariance across groups of test-takers. Misfitting items disrupt the assumption of assessment non-Invariance, expand Type II Errors, provide erroneous mean judgments across student groups, and misrepresent group-level ability estimates, thereby presenting mysterious variability (Ouyang et al., 2025). Violations of standard item-writing

principles such as ambiguous stems, implausible distractors, cueing, negative or double-barreled phrasing yield measurable variations in item statistics and pass or fail outcomes. Disregarding these violations leads to inaccurate group assessments, mostly in large-scale international assessments. A study by Schmucker and Moore (2025) presented empirical evidence that item-writing flaws, such as implausible distractors and ambiguous wording, significantly affect item difficulty and discrimination parameters. Their study demonstrates that poorly constructed multiple-choice items often exhibit systematic misfit, which compromises the stability of item parameters.

Variances in item difficulty (b-parameter) between two groups of students are not only statistical objects but mirror deeper structural and educational inequalities. Features such as curriculum gaps, language barriers, contextual bias, and resource gaps boost DIF. A study by UNICEF (2021) found that access to educational materials improves learners' awareness of test content. Students who study in urban centers benefit more from improved access to reading aids such as textbooks, digital tools, and additional instruction, which lowers item difficulty. In contrast, inadequate access to these learning materials in rural schools enhances higher b-values. This study was supported by OECD (2022), which concluded that language barriers remain a crucial element influencing item difficulty. The study shows that students with limited exposure to academic language are more likely to misinterpret test items, enhancing perceived difficulty. This is mostly appropriate in rural contexts where learners may rely heavily on local mother tongues rather than the recognized instructional language.

Alterations in curriculum implementation and instructional delivery between teachers who teach in rural schools and their counterparts in urban schools remain a key factor in DIF. Studies by UNESCO (2020) and the World Bank (2018) revealed that schools in rural settings regularly experience incomplete syllabus coverage because of the shortages of teachers and infrastructural limitations, thus leading to poor academic performance and the rise of item difficulty for rural students (Owor, 2026). Therefore, items that are associated with under-taught content areas have the tendency to produce higher b-values for rural test-takers.

Halpin (2022) points out the status of robust DIF methods that minimize dependence on anchor items, especially in the presence of model misfit. The author suggested that incorporating process data from computer-based test systems can help separate the relationship between misfit and DIF by offering added insights into test-taker behavior (Chen et al., 2025). Item misfit has profound implications for differential item functioning (DIF), which examines whether items function equivalently across different subgroups such as gender, age group, school location, students from various socio-economic statuses, or school type (government and private). DIF analysis assumes that the underlying IRT model adequately represents the data; thus, violations of model fit can severely undermine the detection and interpretation of DIF (Von Davier & Bezirhan, 2023).

A study on the evaluation of incidences of examination malpractice and intervention measures among undergraduate students at the University of Calabar by Anagbogu et al. (2021) noted that item invariance among students could not be attributed to item construction skills only but also to the examination malpractices that go on in the school setting among some teachers and students alike. The study recommended the sincere implementation of disciplinary actions against culprits, irrespective of who they are. A similar study conducted by Anagbogu et al. (2021) on the evaluation of the provision and use of communication facilities in higher educational institutions in Cross River State reported that, in addition to determining students' abilities based on item model fit of item response theory, lack of available and functional office equipment, electricity, and so on could be added as factors inhibiting students' abilities in responding to the correct options in an instrument. It was, however, recommended that generators or sources of power generation be used for lighting systems to provide an alternative power supply for the effective functioning of office activities, as these measures could significantly influence students' abilities. If the item writers during item development did not take into consideration incomplete taught curricula in rural schools due to a lack of teachers, language barriers, students' socio-economic status, and lack of exposure and access to learning aids by rural students, it may cause the item to function differently, which may also affect the implementation of school goal attainment (Owor & Ogba, 2026).

Examiners or teachers who lack adequate skills in item development can unintentionally introduce differential item functioning (DIF) into their tests -an example of poor written items. Poorly written items frequently include ambiguous, difficult language structures, and diverse skills that are omitted from the intended construct. These defects are capable of making some items less difficult for some groups of students than others who have the same level of cognitive ability, thereby causing DIF and discouraging fairness. Gershenson and Langbein (2015) carried out a study on the impact of school location on student ability and observed that rural schools do not always have access to the equivalent level of federal funding as urban and suburban schools, and this could hinder the opportunity testers have for learning mathematics, which is evidence of the violation of local independence, stressing that students in urban areas demonstrate improved and consistent responses to test items than their rural equals in mathematics, reading, and sciences. Rural schools encounter challenges that lead to adverse educational outcomes for their students. One such challenge

is the lack of Information and Communication Technology (ICT) usage, which is usually deficient in rural areas, as these challenges could ultimately lead to violations in item consistency. Opoku-Asare and Siaw (2015), in their study on why students in the urban areas performed better than their counterparts in the rural areas, observed that the availability of fewer facilities in several rural schools than those in urban areas has been found to contribute to more inadequate curricula being made available at these rural schools. They stressed that since the curricula could not be implemented fully by the teachers in terms of content coverage, it will invariably affect students' response patterns, which is a violation of the local independence of the test items. The urban-rural difference in the response patterns of the students is responsible for factors such as an uneven distribution of resources and lack of functional facilities

Overall empirical evidence indicates that misfitting items hamper item quality and introduce construct-irrelevant variance, which decreases the total accuracy of measurement. However, the present research atmosphere on Computer-Based Testing (CBT) in Nigeria exposes substantial division, with a robust emphasis on infrastructure, approval, and user perceptions, while important parts of test development remain mostly underexplored. A key gap exists in the area of item development proficiency, as there is minimal empirical evidence on people who develop CBT items, their level of psychometric training, or their ability in modern assessment methods such as IRT, Computer-Adaptive Testing, and DIF. This shows a personnel capacity deficit within the CBT environment. In addition, while CBT has been extensively applied in high-stakes tests in Nigeria, there is negligible research assessing item-level statistics such as difficulty indices, distractor efficiency, local independence, and DIF using advanced models like item response theory. This depreciates the evidence base for the reliability and validity of CBT results. Moreover, the absence of structured training models for item writers represents a major gap.

While insufficient skills are often recognized, there is minimal research on resourceful capacity-building methods, such as continuous professional development, institutional alliances, or the integration of AI-assisted item generation tools. Furthermore, Nigeria's readiness for Computer-Adaptive Testing (CAT) remains mostly unexamined. Changing to CAT requires calibrated item pools, advanced processes, and psychometric expertise, yet research in this area is almost non-existent. Also worthy to note is the absence of inclusive and context-sensitive item calibration outlines. Current research recognizes disparities in digital access and literacy but does not satisfactorily explore how item design can accommodate different learner groups, including students in rural areas and their counterparts in urban centers who have little exposure to technology. Lastly, there are noteworthy procedural constraints in the current research. Most research depends on descriptive and perception-based methods, with minimal application of experimental or advanced psychometric approaches. This limits the depth of understanding and the capacity to establish causal relationships. It is against this background that this research was conducted to examine the 2025 Mathematics multiple-choice items in Cross River State.

2. PROBLEM STATEMENT

Explaining Insufficient item-writing skills by teachers and personnel from some examination bodies can seriously impact the quality and fairness of both classroom and standardized tests. When test items are poorly written, they are more likely to misalign with measurement models like IRT/Rasch, resulting in low discrimination and irregular response patterns. This directly affects the validity and reliability of test scores. Moreover, flaws in item construction—such as unclear wording, overly complex language, cultural biases, gender stereotypes, and misalignment with learning objectives—significantly increase the risk of differential item functioning (DIF). This implies that some examinee subgroups may be unfairly favored, even when they share the same underlying cognitive abilities. Consequently, students may end up with misleading scores, be inaccurately classified regarding their proficiency, and face unsuitable decisions regarding placement, selection, or remedial support. Over time, being exposed to confusing and biased test items can diminish students' motivation, increase test-related anxiety, and create negative feelings toward the subject matter and schooling as a whole. Not only does this hinder the accurate assessment of students' true achievements, but it also perpetuates educational inequity. Despite this, many educational systems, particularly at the classroom level, lack empirical evidence regarding the specific effects of poor item-writing skills on item fit, DIF, and the resulting academic outcomes for students. There is a pressing need for systematic research to explore how deficiencies in item development practices lead to misfitting and DIF-prone items and how these issues impact students' performance, classification, and overall learning experiences. It is against this backdrop this research was commissioned to investigate 2025 Mathematics multiple-choice items in Cross River State. This is supported by the alarming increase in the use of Computer-Based Tests (CBT), as objective test items are, in most cases, more reliable, valid, and fair.

3. RESEARCH QUESTIONS

RQ1. Which items of the 2025 mock Mathematics multiple-choice test fit the Rasch, two-parameter, and three-parameter logistic models?

RQ2. Does school location affect students' item response consistency in the mock test?

4. METHODOLOGY

A comparative–causal research design was adopted for the study. It is a method that compares two or more naturally occurring groups to determine whether differences in the independent variable (school location – urban and rural) may have caused differences in the research outcome. The rationale for using this type of design is to accurately investigate the cause-and-effect relationships of the variables without manipulating them. The study covered the entire Cross River State, Nigeria, which comprised 18 local government areas. The population of the study comprised all Senior Secondary Three (SS3) students in the 272 public and 277 private secondary schools in the state, totaling 10,145 in the 2024/2025 academic session who sat for mock Examinations. Forty-eight (48) secondary schools (24 private and 24 public) were selected through a stratified sampling method, while a sample of 1,015 was randomly selected from the population. Two main conditions were used to form the stratification: (1) School Ownership – public and private schools, and (2) School Location – rural schools and urban schools. These principles were combined to form four separate strata (stratum 1: Public–Urban; stratum 2: Public–Rural; stratum 3: Private–Urban; and stratum 4: Private–Rural). This process guarantees balanced representation across both groups. To determine the sample, the researcher applied the formula as follows:

$$n_h = N_h/N * n$$

Where:

n_h = sample from each stratum

N_h = Population in each stratum

N = Total population

n = total sample size

In substituting the data into the formula

Stratum	Population(N_h)	Sample (n_h)
Public-urban	3,000	300
Public-rural	2,500	250
Private-urban	2,800	280
Private-rural	1,845	185
Total	10,145	1,015

The state mock examination was the instrument used for data collection with written permission from the state examination body. The instrument consisted of fifty (50) multiple-choice test items with four options A-D. The adapted instrument consisted of demographic characteristics such as gender, school location (rural and urban), and school type (government-owned and privately-owned schools), as well as items measuring topics such as geometry, probability, algebraic processes, calculus, latitude and longitude, trigonometry, Construction and lucid, statistics, logarithms, indices, arithmetic, Mensuration, and surds. The research instrument (Mathematics multiple-choice items) was submitted to two measurement experts at Federal Polytechnic, Ugep, Cross River State, Nigeria, for face and content validation. The face validity was used to evaluate whether the test appears valid to users(students). For content validation, the experts identified irrelevant items, poorly worded items, missing content areas, and assessed whether the items satisfactorily aligned with the cognitive level.

A letter of approval was obtained from the State Ethical Committee Review Board for the collection and use of data for the study. To guarantee that the research instrument was reliable, a trial test was administered to 50 senior secondary school three students (SSS 3) of Uyo High School, Uyo, Akwa Ibom State, Nigeria, who were not members of the study area. Kuder-Richardson (K-R 20 reliability was adopted to check the internal consistency of the instrument, which had a reliability coefficient estimate of 0.824. The responses to the instrument were used for statistical analysis. A key was developed in which all the responses to the test items were coded. The research instrument was made up of two segments, viz., section "A" and section "B". Section "A" consisted of the test taker identification in terms of gender and school location, while section "B" of the instrument was made up of the responses to the mathematics multiple-choice items. A score sheet was provided in a person-by-item matrix format where all the responses were coded with one (1) for the right response and zero (0) for the wrong responses. After scoring, the responses were arranged and coded in Notepad using a person-by-item matrix format with no records of missing data. Bilog-MG3 computer software was applied to analyze the data. However, prior to the analysis, the assumption of unidimensionality of Item Response Theory (IRT) was tested. Chi-square goodness-of-fit of Rasch (1PL), 2-parameter, and 3-parameter

logistic Item response theory (IRT) models were adopted to determine the results of the item fit statistics, while the Rasch/1-parameter logistic (1PL) model was used to determine how the items function among the two groups of students (DIF).

5. RESULTS

5.1. Item Fit Analysis of the 2025 Mathematics Mock Test Across the Rasch, 2PL, and 3PL Models

Item fit results were evaluated using Chi-square statistics generated from Bilog-Mg to answer the research question (see Table 1). Items with p-values less than or equal to 0.05 were considered misfitting any model, while items whose p-values are greater than 0.05 are considered to fit the model. Under the Rasch (1PL) model, the results revealed that 30 out of 50 items representing 60% exhibited good fit, while 20 items representing 40% significantly misfit the model. From the 1PL model, the misfit items are as follows: item 1 (χ^2 = 53.9, df = 27, p = 0.001), item 11 (χ^2 = 36.3, df = 24, p = 0.031), item 13 (χ^2 = 41.1, df = 24, p = 0.016), item 15 (χ^2 = 39.6, df = 24, p = 0.023), item 17 (χ^2 = 40.9, df = 23, p = 0.012), item 18 (χ^2 = 37.0, df = 23, p = 0.032), item 19 (χ^2 = 56.4, df = 22, p = 0.000), item 20 (χ^2 = 42.5, df = 22, p = 0.007), item 21 (χ^2 = 42.3, df = 24, p = 0.011), item 22 (χ^2 = 47.5, df = 23, p = 0.001). Other misfit items under the Rasch (1-Parameter Logistics) model are item 24 (χ^2 = 45.2, df = 22, p = 0.002), item 25 (χ^2 = 43.8, df = 26, p = 0.015), item 42 (χ^2 = 40.1, df = 27, p = 0.020), item 43 (χ^2 = 50.9, df = 28, p = 0.005), item 44 (χ^2 = 42.1, df = 27, p = 0.032), item 45 (χ^2 = 53.4, df = 27, p = 0.032), item 47 (χ^2 = 89.3, df = 29, p = 0.000), item 48 (χ^2 = 53.1, df = 28, p = 0.002), item 49 (χ^2 = 69.9, df = 29, p = 0.000), and item 50 (χ^2 = 76.7, df = 28, p = 0.000). Meanwhile, 30 fitted items (60%) are marked with an asterisk (*) sign. These 20 misfit items exhibited significant misfit, demonstrating a discrepancy between observed and expected response patterns under the 1PL model. These misfit items are attributed to multidimensionality (items measuring more than one ability level) and unequal discrimination across items (the items were too easy to guess, or too sensitive to the differences in examinees' abilities). Other reasons for these misfit items were unclear wording, many probable answers, weak distractors, and examinees' rush to answer the last items. These items cannot be trusted for Rasch-based measurement.

Similarly, under the 2-parameter Logistics model, which accounts for both item difficulty and discrimination, Chi-square statistics indicated that 33 out of 50 items representing 66% fit the model, while 17 items representing 34% misfit the model. The fit items include 1, 2, 3, 4, 7, 8, 10, 14, 16, 23, 24, 26, 27, 28, and 30. Other items are 31, 32, 33, 34, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, and 50. These items showed that their respective p-values were all greater than the 0.05 significance level. Meanwhile, the misfit items under the 2PL models include item 5 (χ^2 = 50.8, df = 32, p = 0.018), item 6 (χ^2 = 52.9, df = 32, p = 0.011), item 9 (χ^2 = 42.3, df = 26, p = 0.022), item 11 (χ^2 = 40.7, df = 26, p = 0.033), item 12 (χ^2 = 40.2, df = 24, p = 0.020), item 13 (χ^2 = 58.5, df = 28, p = 0.000), item 15 (χ^2 = 52.1, df = 24, p = 0.000), item 17 (χ^2 = 48.2, df = 27, p = 0.007), item 18 (χ^2 = 59.4, df = 25, p = 0.000), item 19 (χ^2 = 56.7, df = 26, p = 0.000), item 20 (χ^2 = 55.0, df = 22, p = 0.000), item 21 (χ^2 = 50.1, df = 29, p = 0.008), and item 22 (χ^2 = 55.2, df = 23, p = 0.000). Other 2PL misfit items are item 22 (χ^2 = 49.3, df = 25, p = 0.030), item 29 (χ^2 = 51.5, df = 34, p = 0.030), item 35 (χ^2 = 57.3, df = 35, p = 0.010), and item 40 (χ^2 = 52.0, df = 36, p = 0.041). These 17 items also exhibited significant misfit, demonstrating a disagreement between observed and expected response patterns under the 2-parameter logistic models. The misfit was caused by violations of unidimensionality (items measuring several abilities), monotonicity (lack of consistency in the probability of correct response from examinees with superior abilities), and local independence (responses to items depend on other items) assumptions of item response theory (IRT). Other reasons were miskeyed answers, trick questions, and extreme item difficulty.

Correspondingly, for the 3-parameter Logistics model, which incorporates a guessing factor, Chi-square goodness-of-fit analysis revealed that 40 items representing 80% fit the model, while 10 items representing 20% revealed a significant misfit. For example, the 10 misfit items comprised 20, 22, 25, 27, 36, 44, 47, 48, 49, and 50. These 10 items also exhibited significant misfit, demonstrating a disagreement between observed and expected response patterns under the 3-parameter logistic models, and this is attributed to the multidimensionality of the test items. This misfit is attributed to the violation of unidimensionality (items measuring several abilities), monotonicity (lack of consistency in the probability of correct response from examinees with superior abilities), and local independence (responses to items depend on other items) assumptions of item response theory (IRT). Other reasons for the misfit were cheating and overly high or overly low behaviours.

However, several items were identified as misfitting based on Chi-square statistics, the overall pattern indicated improved fit with more complex (3PL) models. This shows that the assumptions of equal discrimination (1-Parameter Logistics model) did not hold for the examination, while the presence of discrimination and guessing factors (2-Parameter Logistics and 3-Parameter Logistics models) offers an improved representation of the item response behavior of the examinees.

Table 1. Chi-Square Goodness of Fit results of the 2025 Mathematics Mock Test using Rasch, 2PL, and 3PL IRT Models

Item	Rasch Model χ^2	df	p	2PL Model χ^2	df	p	3PL Model χ^2	df	p
1	53.9	27	0.001	26.6	38	.917*	29.2	34	.703*
2	25.8	27	.530*	30.8	36	.715*	33	33	.469*
3	24.8	26	.531*	42.4	34	.152*	21	31	.911*
4	19.8	26	.800*	38.8	34	.262*	30.7	30	.430*
5	17.1	26	.904*	50.8	32	0.018	24.4	30	.751*
6	19.4	27	.854*	52.9	32	0.011	29.9	29	.418*
7	15	25	.941*	30.1	29	.411*	18	27	.904*
8	19.5	26	.815*	27.8	30	.580*	22.3	25	.621*
9	14	25	.961*	42.3	26	0.022	20.7	26	.755*
10	17.1	24	.845*	32.9	30	.327*	21.3	25	.675*
11	36.3	24	0.031	40.7	26	0.033	18.8	24	.762*
12	17	24	.850*	40.2	24	0.02	25.5	21	.224*
13	41.1	24	0.016	58.5	28	0	22.5	24	.550*
14	18.4	25	.825*	38.5	27	.069*	11	23	.982*
15	39.6	24	0.023	52.1	24	0	31.4	21	.067*
16	31	22	.096*	35.3	24	.063*	25.5	22	.275*
17	40.9	23	0.012	48.2	27	0.007	24.2	21	.283*
18	37	23	0.032	59.4	25	0	29.9	20	.072*
19	56.4	22	0	56.7	26	0	25.8	21	.214*
20	42.5	23	0.007	55	22	0	38.4	21	0.011
21	42.3	24	0.011	50.1	29	0.008	34.9	23	.052*
22	47.5	23	0.001	55.2	23	0	39.3	20	0.006
23	37	25	.057*	32.1	27	.229*	34.4	25	.099*
24	45.2	22	0.002	23.3	27	.670*	33.7	23	.069*
25	43.8	26	0.015	49.3	30	0.014	46.4	24	0.003
26	33.7	26	.141*	31.8	31	.426*	23.4	28	.714*
27	35.6	26	.100*	45.6	32	.056*	41.7	28	0.046
28	28.1	26	.351*	38.5	30	.136*	29.6	26	.283*
29	22.6	25	.601*	51.5	34	0.027	30.5	30	.438*
30	26.2	25	.398*	35.9	33	.334*	28.6	29	.484*
31	24	27	.632*	46.2	33	.062*	37	29	.145*
32	22	24	.580*	35.8	33	.337*	26.4	29	.602*
33	21.3	26	.726*	45.2	35	.115*	39	29	.101*
34	22.5	24	.546*	26.4	34	.821*	31.2	28	.307*
35	26.9	26	.412*	57.3	35	0.01	39.8	31	.133*
36	22.2	27	.728*	35.9	35	.426*	48.6	33	0.038
37	34.3	27	.156*	47.8	35	.073*	41.2	33	.154*
38	17.8	28	.931*	32	35	.612*	30	34	.661*
39	39.5	28	.073*	38	34	.293*	28.8	36	.798*
40	26.1	29	.618*	52	36	0.041	41.2	33	.155*
41	40.3	28	.062*	51.4	37	.058*	24.6	34	.881*
42	40.1	27	0.02	33.1	34	.513*	38.6	36	.354*
43	50.9	28	0.005	31.7	39	.792*	41.3	37	.287*
44	42.1	27	0.032	45	36	.145*	48.7	33	0.038
45	53.4	29	0.003	40.7	42	.528*	49.9	39	.112*
46	39.3	28	.076*	50.1	40	.130*	45.9	39	.208*
47	89.3	29	0	35.7	42	.741*	58.2	39	0.024
48	53.1	28	0.002	27.7	40	.930*	58.5	39	0.023
49	69.9	29	0	39.7	42	.571*	58.4	39	0.023
50	76.7	28	0	51.3	41	.130*	84.2	39	0

Results at a glance
20 items (40%) misfit the 1PL model.
17 items (34%) misfit the 2PLmodel
10 items (20%) misfit the 3PLmodel

5.2. Effect of School Location on Students' Item Response Consistency in the 2025 Mathematics Mock Test

Differential item functioning (DIF) was examined using the Rasch one-parameter logistic (1PL) model implemented in BILOG-MG software. The analysis compared item difficulty estimates between students attending rural schools (the focal group) and those attending urban schools (the reference group), as shown in Table 2. Since the 1PL model assumes equal discrimination, DIF was evaluated using differences in item difficulty (b) parameters. Items with a change in the difficulty parameter greater than or equal to 0.50 ($\Delta b \geq 0.50$) were identified for DIF. The extent of the DIF was categorized based on the difference in item difficulty parameters change (Δb) between rural and urban groups. Items with (Δb) less than 0.50 were classified as

having low DIF, those between 0.50 and 0.99 as moderate DIF, and those with (Δb) of 1.00 or greater as high DIF.

The results indicated that 30 items representing 60% had no meaningful DIF. This is because the items exhibited low DIF, signifying minimal ability differences between students who attended schools in urban areas and their counterparts in rural areas. For example, item 1 (Δb = 0.057), item 2 (Δb = -0.152), item 3 (Δb = 0.071), item 5 (Δb = 0.057), item 6 (Δb = -0.111), item 8 (Δb = -0.119), item 10 (Δb = -0.112), item 11 (Δb = 0.026), item 13 (Δb = -0.010), item 16 (Δb = -0.040), item 18 (Δb = -0.030), item 20 (Δb = -0.056), item 21 (Δb = 0.111), item 23 (Δb = 0.122), and item 25 (Δb = 0.126). Other items include 27, 28, 29, 30, 31, 33, 34, 35, 37, 39, 41, 44, 45, 47, and item 48.

Similarly, the results further revealed that 11 items representing 12% demonstrated moderate DIF, which indicated that students who studied in rural areas required approximately 0.5 to 1.0 logits higher ability than their counterparts in urban schools to achieve the same probability of a correct response, indicating significant differential functioning, which made the items harder for Rural students. These 11 items showed a negative (-) value in the ability change between both groups of students. These items are item 4 (Δb = -0.598), item 7 (Δb = -0.623), item 9 (Δb = -0.699), item 12 (Δb = -0.584), item 14 (Δb = -0.551), item 24 (Δb = -0.789), item 26 (Δb = -0.628), and item 32 (Δb = -0.576). Other items are 36, 38, and 42.

Correspondingly, the results also showed that only three (3) items representing 6% also demonstrated DIF, in this case showing positive signs, which indicated that students who studied in urban areas required approximately 0.5 to 1.0 logits higher ability than their counterparts in rural schools to achieve the same probability of a correct response, indicating significant differential functioning, which made the items harder for urban students. These 11 items showed a positive value in the ability change between both groups of students. These items are item 22 (Δb = 0.613), item 43 (Δb = 0.613), and item 46 (Δb = 0.585). However, six (6) items representing 12% exhibited no DIF between the two groups of students, indicating that both groups of students have an equal chance of answering the items correctly. These items are item 15 (Δb = 0.000), item 17 (Δb = 0.000), item 19 (Δb = 0.000), item 40 (Δb = 0.000), item 49 (Δb = 0.000), and item 50 (Δb = 0.000).

Table 2. Differential Item Functioning (DIF) of the 2025 Mock Mathematics multiple-choice test items responses based on student's school location

Item	Urban (b)	Rural (b)	Δb	Ability Difference	Remarks	Item	Urban (b)	Rural (b)	Δb	Ability Difference	Remarks
1	-0.878	-0.82	0.057	Low	NMDIF	26	-1.015	-1.643	-0.628	Moderate	HRS
2	-0.865	-1.017	-0.152	Low	NMDIF	27	-1.082	-1.016	0.065	Low	NMDIF
3	-1.001	-0.93	0.071	Low	NMDIF	28	-1.046	-1.071	-0.025	Low	NMDIF
4	-0.404	-1.002	-0.598	Moderate	HRS	29	-1.082	-0.979	0.103	Low	NMDIF
5	-1.04	-0.984	0.057	Low	NMDIF	30	-1.05	-1.031	0.02	Low	NMDIF
6	-0.95	-1.061	-0.111	Low	NMDIF	31	-1.03	-0.988	0.042	Low	NMDIF
7	-1.098	-1.721	-0.623	Moderate	HRS	32	-0.44	-1.016	-0.576	Moderate	HRS
8	-0.972	-1.091	-0.119	Low	NMDIF	33	-1.02	-0.939	0.082	Low	NMDIF
9	0.204	0.903	-0.699	Moderate	HRS	34	-1.015	-1.031	-0.015	Low	NMDIF
10	-1.001	-1.113	-0.112	Low	NMDIF	35	-1.071	-0.891	0.18	Low	NMDIF
11	-1.082	-1.056	0.026	Low	NMDIF	36	1.006	1.646	-0.64	Moderate	HRS
12	-1.051	-1.635	-0.584	Moderate	HRS	37	-1.006	-0.9	0.106	Low	NMDIF
13	-1.087	-1.097	-0.01	Low	NMDIF	38	-0.349	-1.017	-0.668	Moderate	HRS
14	-1.04	-1.591	-0.551	Moderate	HRS	39	-0.982	-0.913	0.069	Low	NMDIF
15	-1.098	-1.098	0	Equal ability	—	40	-0.979	-0.979	0	Equal ability	—
16	-1.03	-1.071	-0.04	Low	NMDIF	41	-0.958	-0.948	0.011	Low	NMDIF
17	-1.093	-1.093	0	Equal ability	—	42	-0.458	-0.998	-0.54	Moderate	HRS
18	-1.035	-1.066	-0.03	Low	NMDIF	43	0.826	0.213	0.613	Moderate	HUS
19	-1.061	-1.061	0	Equal ability	—	44	-0.963	-0.988	0.026	Low	NMDIF
20	-1.046	-1.102	-0.056	Low	NMDIF	45	-0.847	-0.816	0.031	Low	NMDIF
21	-1.077	-0.965	0.111	Low	NMDIF	46	-0.897	-0.312	0.585	Moderate	HUS
22	0.93	0.317	0.613	Moderate	HUS	47	-0.83	-0.74	0.09	Low	NMDIF
23	-1.114	-0.993	0.122	Low	NMDIF	48	-0.813	-0.939	0.126	Low	NMDIF
24	-1.035	-1.824	-0.789	Moderate	HRS	49	-0.775	-0.775	0	Equal ability	—
25	-1.143	-1.016	0.126	Low	NMDIF	50	-0.816	-0.816	0	Equal ability	—

Result summary, Abbreviations and meaning

Δ*b* = Ability change;
NMDIF (No meaningful DIF) = 30;
HRS (Harder for Rural students) = 11;
HUS (Harder for urban students) = 3;
Equal ability = 6;
TOTAL = 50

6. DISCUSSION

The assessment of model fit of the 2025 mock mathematics multiple-choice test revealed that 20 items (40%) misfitted the Rasch (1PL) model, 17 items (34%) misfitted the 2PLmodel, and 10 items (20%) misfitted the 3PLmodels. This misfit was caused by the violation of unidimensionality, monotonicity, and local independence assumptions of IRT. Other reasons for the misfit were tricky questions, unclear item wording, many possible answers, negative phrasing, weak distractors, and examinees' rush to answer the last items in the test. Items under the 3PL are indicative of improved-fit items that are supposed to be trusted and banked for measurement, while misfit items are indicative of bad items, expected to be reviewed or removed from the test. This is because the items cannot be trusted in measurement. This result agrees with Schmucker and Moore (2025), who remarked that item-writing errors are a major cause of item misfit. The authors e stressed that several item-writing flaws such as implausible distractors, unclear wording, and negative phrasing are linked with misfit. These errors, according to the authors, reduce item discrimination and increase unpredictability in responses. Misfitting items introduce construct-irrelevant variance, which decreases the overall accuracy of measurement. This reduction in accuracy is problematic in high-stakes testing, where small differences in scores can have substantial consequences (Fährmann et al., 2022).

The results of the DIF analysis revealed that 30 items representing 60% had no meaningful DIF. This is because the items exhibited low DIF, signifying minimal ability differences between students who attended schools in urban areas and their counterparts in rural areas. Meanwhile, 11 items representing 12% demonstrated moderate DIF, and the items were harder for students in rural schools than their counterparts in urban schools, while three (3) items were harder for the urban students than those from rural schools, with six (6) having equal ability among the two groups of students. These differences in ability among the two groups of students were caused by changes in curriculum implementation and instructional delivery by teachers in rural schools, incomplete syllabus coverage, language barriers, and improved access to reading aids such as textbooks, digital tools, and additional instruction. Students in rural schools have limited exposure to academic language, shortages of teachers to effectively implement the curriculum in later, under-taught content areas, and infrastructural limitations. These factors contributed to the rise of b-parameter values for rural examinees. This result is in tandem with UNICEF (2021), which noted that access to educational materials improves learners' awareness of test content. They stressed that students who study in urban centres benefit more from improved access to reading aids such as textbooks, digital tools, and additional instruction, which lowers item difficulty. In contrast, inadequate access to these learning materials in rural schools enhances higher b-values. This result was affirmed by OECD (2022), which concluded that language barriers remain a crucial element influencing item difficulty. The study shows that students with limited exposure to academic language are more likely to misinterpret test items, enhancing perceived difficulty. This is mostly appropriate in rural contexts where learners may rely heavily on local mother tongues rather than recognised instructional language. The study by UNESCO (2020) and the World Bank (2018) also supported the result. Their studies revealed that schools in rural settings regularly experience incomplete syllabus coverage because of the shortages of teachers and infrastructural limitations, thus leading to the rise of item difficulty for rural students. Therefore, items that are associated with under-taught content areas have the tendency to produce higher b-values for rural test-takers.

7. CONCLUSION

From the results of the study, it could be concluded that 10 of the items that did not fit the 3-parameter logistic model were indicative of bad items and cannot be trusted for measurement. These items were required to be dropped from the test or revised. These misfit items affect the consistency of classification decisions, thereby weakening the relationship between test scores and external criteria. These items may also cause construct underrepresentation, and the scores may not accurately reflect the intended latent trait measured. In a nutshell, the results concluded that a lack of item development experts, inadequate psychometric training in modern assessment frameworks such as Item Response Theory (IRT), and a lack of or inadequate research on the processes of item development are associated with poor item writing skills by item developers who are teachers.

This result further concluded that the differences in the DIF among the two groups of students (urban-rural) were not about item writing skills, but other factors such as incomplete syllabus coverage, language barriers, under-taught content areas, infrastructural limitations, and inadequate access to learning materials in rural schools. These factors caused a change in the ability parameter (b) among the two groups of students.

8. RECOMMENDATIONS

To improve the quality, reliability, and fairness of future test items for measurement and item banking, several recommendations are proposed based on the findings of this study. First, test developers should conduct rigorous preliminary testing by administering newly developed items to an appropriate pilot sample before their operational use, allowing items with poor fit statistics to be identified and subsequently revised or removed. They should also ensure that all items are aligned with the intended construct, avoiding the inclusion of items that measure unrelated abilities, such as combining mathematics items with reading comprehension items in a test designed to measure mathematical ability. Furthermore, item writers should conduct factor analysis during the development stage to assess the unidimensionality assumption underlying IRT-based calibration, carefully review ambiguous item wording, and revise or eliminate items that are excessively easy or difficult for the target population. To strengthen these practices, the government should provide specialized training for item-writing experts in IRT-based item development, calibration, and DIF detection, enabling them to develop and evaluate high-quality items that function appropriately across different student groups. In addition, item writers should consider inequalities in educational resources when developing test items by avoiding questions that assume students have prior experience with specialized equipment or digital technologies, such as ICT tools, laboratory facilities, and textbooks, which may be more readily accessible to urban students. Finally, test items should reflect curriculum content that is realistically taught across different regions, regardless of school location, rather than placing disproportionate emphasis on advanced topics that are more likely to be covered in urban schools. These measures are expected to support the development of more valid, equitable, and contextually appropriate assessment instruments for future testing and item banking.

ACKNOWLEDGEMENTS

I sincerely appreciate the contributions of all individuals who supported the successful completion of this study. Special appreciation is extended to the two research assistants for their dedication, commitment, and invaluable assistance during the data collection process. Their cooperation and diligence contributed significantly to the successful administration and retrieval of the research instruments.

I also express my profound gratitude to the experts in item development and educational measurement at the Faculty of Education, University of Calabar, who reviewed the research instrument and provided valuable professional insights during its face and content validation. Their constructive comments, observations, and recommendations were vital in enhancing the clarity, relevance, appropriateness, and overall quality of the instrument.

Furthermore, my sincere appreciation goes to the classroom teachers who assisted in coordinating and organising the students throughout the data collection exercise. Their cooperation, support, and effective coordination facilitated the smooth administration of the instrument and ensured minimal disruption to regular classroom activities.

I am deeply grateful to everyone whose contributions, in one way or another, facilitated the successful conduct of this study.

DECLARATION OF INTEREST

The author declare no conflicts of interest.

RESEARCH FUNDING

This research did not receive any financial support.

ETHICAL STATEMENT

Ethical approval for this study was obtained from the State Ethical Committee Review Board, Nigeria. The study involved human participants, specifically Senior Secondary School Three (SS3) students, who participated in the administration of the mathematics multiple-choice test.

AI USE STATEMENT

The authors declare that no generative artificial intelligence (AI) tools were used in the preparation, analysis, or writing of this manuscript.

REFERENCES

Anagbogu, G. E., & Owor E. O. (2021). School factors as correlates to examination malpractice among secondary school students in Cross River State, Nigeria. *IAR Journal of Humanities and Social Science, 2*(4), 1-4. https://doi.org/10.47310/iarjhss.2021.v02i02.005

Anagbogu, G. E., Onu, E., & Owor, E. O. (2021). Evaluation of incidences of examination malpractices and intervention measures among undergraduate students in university of calabar, Nigeria: Implication for assessment and certification. *International Journal of Arts and Commerce, 10*(8), 16-32

Anagbogu, G. E., Akpan, S.M., & Owor, E. O. (2021). An evaluation of the provision and use of communication facilities in higher educational institutions in Cross River State. *IAR Journal of Humanities and Social Science, 2*(5), 98-103. https://doi.org/10.47310/IARJHSS.V02I05.014

Anagbogu, G. E., Dada, O.A., Petters, J.S., & Owor, E. O. (2021). An assessment of social media immersions and undergraduate student's academic performance in research methods in University of Calabar, Nigeria. *IAR Journal of Humanities and Social Science, 2*(4), 45-53. https://doi.org/10.47310/iarjhss.2021.v02i02.009

Anagbogu, G. E., Ofoegbu, J.U., Ovat, S,V., & Owor, E. O. (2021). A path analytical model on factors with the implementation of civic education curriculum among teachers in Cross River State, *Nigeria. International Journal of Education, 9*(3), 13-24. https://doi.org/10.5121/IJE2021.9302

Bisong, A. E., Oko, B. A., John, E. I., Achu, R. A., & Owor, E. O. (2021). Personal variables and students' academic performance of virtual Classroom learners of social studies in secondary schools in Uyo Local Government Area, *Akwa Ibom State Review of European Studies, 13*(3), 11-19. https://doi.org/10.5539/res.v13n3p1

Bond, T. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.).* Routledge.

Chen, L., Zhang, S., & Liu, J. (2025). Reducing differential item functioning via process data. https://doi.org/10.48550/arXiv.2504.00136

Dalati, S. (2018). Measurement and measurement scales. *Journal of Measurement, 18*(7), 79-96

Fährmann, K., Köhler, C., Hartig, J., & Heine, J.-H. (2022). Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies. *Large-scale Assessments in Education, 10*(1), 7-28. https://doi.org/10.1186/s40536-022-00122-8

Fischer, L. (2021). Consequences of limited item pools and model misfit. *Frontiers in Psychology, 12(*6), 68-81. https://doi.org/10.3389/fpsyg.2021.633896.

Halpin, P. F. (2022). Differential item functioning via robust scaling. https://doi.org/10.48550/arXiv.2207.04598

Johanns, B., Dinkens, A., & Moore, J. (2107). A systematic review comparing open-book and closed-book examinations: Evaluating effects on development of critical thinking skills. *Nurse Education in Practice, 27*, 89–94. https://doi.org/10.1016/j.nepr.2017.08.018

Jones, R. N. (2019). Differential item functioning and its relevance to measurement fairness: A review. *Journal of Educational Measurement, 56(*2), 123–140. https://doi.org/10.1111/jedm.12123.

Köhler, M,, & Hartig , T. (2017). Effects of mindfulness training and nature exposure on psychological well-being. *Journal of Environmental Psychology, 12*(7), 438-447

Maric, D., et al. (2023). A systematic literature review of psychometric trends in STEM education. *International Journal of Education, 9*(2), 37-50 https://doi.org/10.1186/s40594-023-00430-x

Mullis, 1. V. S., Martin, M. O., Kennedy, A. M., & Flaherty, C. I. (Eds.) (2010). *PIRL 2001 encyclopedia: A reference guide to reading education in the countries participating in IEA's Progress International Reading Literacy Study (PIRLS). Chestnut Hill, MA: Boston College.* TIMSS & PIRLS International Study Center.

Müller, M. (2020). Item fit statistics for Rasch analysis : can we trust them?. *Journal of Statistical Distributions and Applications, 7*(5). https://doi.org/10.1186/s40488-020-00108-7

Opoku-Asare, N. A. A., & Siaw, A. O. (2015). Rural–urban disparity in students' academic performance in visual arts education: Evidence from six senior high schools in Kumasi, Ghana. *SAGE Open, 5*(4), 1–14. https://doi.org/10.1177/2158244015612523

OECD (2022). PISA 2022 Results (Volume I).

Onyibe, C. O., Uma, U. U., & Ibina, E. (2015). Examination malpractice in Nigeria: Causes and effects on National development. *Journal of Education and Practice, 6*(26), 56-73

Owor, E. O. & Ogba, U. F. (2026). Factors inhibiting goal attainment in public secondary schools in Cross River State, Nigeria: A path analytical model. *International Journal of Sub-Saharan African Research, 4*(1), 472-484. https://doi:10.5281/zenodo.18957398

Owor, E. O. (2026). A path model of academic performance among part-time students in Nigerian Tertiary institutions: The role of facilities, peer influence, and teacher-student relationships. Review of European Studies, 18(1), 27-41. https://doi.org/10.5539/res.v18n1p41

Ouyang, J., Chen, Y., Li, C., & Xu, G. (2025). Statistical analysis of large-scale item response data under measurement non-invariance. https://doi.org/10.48550/arXiv.2505.16608

Schmucker, R., & Moore, S. (2025). The impact of item-writing flaws on difficulty and discrimination in item response theory. https://doi.org/10.48550/arXiv.2503.105333

Gershenson, S., & Langbein, L. (2015). The Effect of Primary School Size on Academic Achievement. *Educational Evaluation and Policy Analysis, 37*(1S), 135S-155S. http://www.jstor.org/stable/43773501

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia - Social and Behavioral Sciences, 29*, 188–197. https://doi.org/10.1016/j.sbspro.2011.11.224

UNESCO (2020). *Global Education Monitoring Report.*

UNICEF (2021). *The State of the World's Children.*

von Davier, M., & Bezirhan, U. (2023). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement, 83*(4), 740–765. https://doi.org/10.1177/00131644221105819

World Bank (2018). *World Development Report: Learning to Realize Education's Promise.* World Bank. http://hdl.handle.net/10986/28340

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology, 8*(4), 484-499. https://doi.org/10.3389/fpsyg.2017.00484.

Zhao, Y. (2018). The impact of IRT item misfit on score estimates and classification decisions. *Quality of Life Research, 26*(3), 555–564. https://doi.org/10.1007/s11136-016-1467