frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 10 October 2025
DOI 10.3389/feduc.2025.1658971

OPEN ACCESS

EDITED BY
Maura Pilotti,
Prince Mohammad bin Fahd University,
Saudi Arabia

REVIEWED BY
Maryam Bojulaia,
Prince Mohammad bin Fahd University,
Saudi Arabia
Hassan Mubarik Iddrisu,
West African Examinations Council, Ghana

*CORRESPONDENCE
Nathaniel Quansah
✉ nquansah002@stu.ucc.edu.gh

RECEIVED 03 July 2025
ACCEPTED 05 September 2025
PUBLISHED 10 October 2025

CITATION
Quansah N, Quansah F and Dzakadzie Y (2025) Multiple-choice test development competencies of junior high school mathematics teachers in Ghana: a triangulation methodology.
Front. Educ. 10:1658971.
doi: 10.3389/feduc.2025.1658971

Multiple-choice test development competencies of junior high school mathematics teachers in Ghana: a triangulation methodology

Nathaniel Quansah[1]*, Frank Quansah[2] and Yayra Dzakadzie[2]

[1]Faculty of Educational Foundations, University of Cape Coast, Cape Coast, Ghana, [2]Department of Educational Foundations, University of Education Winneba, Winneba, Ghana

Background: Previous studies on test development competencies have used three distinct approaches, but none have combined these methods, leaving a gap in understanding teachers' test development skills comprehensively.
Purpose: Using a triangulation methodology, this study examined the multiple-choice test (MCT) development competencies of junior high school (JHS) mathematics teachers in the Sekondi-Takoradi Metropolis, Ghana.
Methods: A survey was first conducted with 218 mathematics teachers, followed by a documentary analysis of tests developed by purposefully sampled teachers. Finally, item analysis was performed using data from the teachers' tests. The collected data were analysed through confirmatory factor analysis, structured qualitative content analysis, item response theory (IRT), and Kendal's Tau-b.
Results: The study found that teachers generally reported multiple-choice test competencies in areas of test item assembling, content validity, and test option handling. An evaluation of sample items revealed that most of the developed MCTs lacked clear directions, contained ambiguous or irrelevant content, and had generally poor quality. The IRT analysis revealed generally poor psychometric properties of the test developed by the teachers. A moderate, statistically significant positive correlation was observed between adherence to recommended MCT development principles and the quality of items produced.
Conclusion: JHS mathematics teachers primarily developed low-quality MCT items, which is concerning because poor-quality tests undermine accurate assessments and sound educational decisions in Ghana.
Recommendations: The Ghana Education Service, NaCCA, and District Directors should prioritise sustained, hands-on professional development workshops in test construction and item analysis to strengthen in-service teachers' assessment literacy and improve classroom testing practices.

KEYWORDS

multiple-choice test, test construction competencies, triangulation research, test quality, item response theory

Introduction

Education is considered crucial to the development of economies. This is due to the connection between the quality of education provided to citizens and the development of nations. Based on this relationship, the education systems of nations are designed to meet economic needs and developmental priorities. Though education is seen to be critical to national development, Abreh et al. (2018) assert that science and mathematics education are even more vital in the expansion and management of economies. As a result, many governments have made mathematics and science core subjects at both the junior and senior high school levels. Again, the recognition of the role of mathematics and science education in achieving development goals has led governments to initiate and implement incentive schemes to increase participation and achievement in both mathematics and science. For example, some of the interventions undertaken by the Ghanaian government include the Secondary Education Improvement Project (SEIP) and the Mathematics, Science and Technology Scholarship Scheme (MASTESS).

Despite ongoing efforts by governments, improvements in mathematics achievement have been minimal, with consistently low performance among Junior High School (JHS) students (WAEC, 2016). The West African Examination Council (WAEC) has been the sole organisation evaluating students' performance after completing junior and senior high school in Ghana for many years. Recent trends show many students perform poorly in final exams (especially in mathematics) across various regions (Ansah et al., 2020; Fletcher, 2018; Abreh et al., 2018; Bosson-Amedenu, 2017; Mills and Mereku, 2016). This persistent underperformance has attracted scholarly attention, leading to the identification of several contributing factors, including teacher characteristics (Ansah et al., 2020; Abreh et al., 2018), student traits (Amoako et al., 2024; Iddrisu et al., 2023), and school-related factors (Ankoma-Sey et al., 2019; Gichuru and Ongus, 2016). Among these, a key factor often emphasised is teachers' test construction practices. Teachers need to have both the competence and confidence to design practical assessments (Oppong et al., 2023). They should be seen as knowledgeable individuals capable of accurately measuring learning objectives. Similarly, for teachers to effectively assess what they intend to, the test tools they develop must be precise and reliable. Achieving this requires teachers to be skilled in managing test tools, including tests and examinations (Fives and DiDonato-Barnes, 2013).

The issue of the validity and reliability of classroom tests in schools has attracted growing attention from researchers and educational stakeholders (Ansah et al., 2020; Baker, 2003). This concern has been sparked by the fact that teachers seem to lack the skill of constructing test items; the majority of their classroom-based achievement tests have low reliability and validity (Agu et al., 2013). Furthermore, it has been observed that some tests used for end-of-term examinations and continuous assessment in various schools contain ambiguous, construct-irrelevant variance, and misleading items (LeFevre and Dixon, 1986; Quansah et al., 2019). The implication is that the test developed by these teachers may be characterised by poor quality, which in turn affects the performance of learners due to their inadequate test construction skills.

Teachers' test construction competencies in Africa

In Africa, especially in Ghana and Nigeria, teachers' test construction competencies have been a contentious issue, with previous research showing mixed results. For example, a study by Ankomah (2020) reported a relatively high level of test construction skills among secondary school teachers teaching Mathematics, Science, and English Language. Generally, earlier studies found that teachers at the pre-tertiary level displayed limited skills and competence in test item development (Kissi et al., 2023; Ankomah and Nugba 2020; Quansah et al., 2019; Amedahe, 1993; Quaigrain, 1992). Other scholars argued that, despite teachers being trained in assessment and item writing, most had a negative attitude and did not follow recognised test development principles (Anhwere, 2009; Ebinye, 2001; Quansah, 2018; Quansah and Amoako, 2018). Findings from other significant studies indicated that teachers who had received training in educational assessment and test construction were better at following recommended principles than those who had not (Quansah and Ankoma-Sey, 2020; Anhwere, 2009; Oduro-Kyireh, 2008; Amedahe, 1993; Quaigrain, 1992).

An extensive study of previous research on teachers' test construction revealed three main approaches to assessing test development competencies. The first group of researchers, through surveys, asked teachers to respond to statements about test adherence principles, attitudes towards test development, and construction skills (see Amedahe, 1993; Anhwere, 2009; Ankomah and Nugba 2020; Oduro-Kyireh, 2008; Quaigrain, 1992; Quansah, 2018; Quansah and Amoako, 2018). This self-report method has been commonly used in earlier studies in this area, dating back to the 1990s. While the method is helpful in understanding teachers' perceptions, attitudes, and practices regarding their test construction, such responses may not accurately reflect their actual test construction practices, as most teachers would prefer not to be perceived as lacking test construction skills. This was demonstrated in a recent study where teachers rated their test construction skills, which were then compared to the quality of the test items they developed (Kissi et al., 2023). The study found that teachers' self-assessed skills did not align with their actual performance, as they failed to demonstrate the test development competencies that they claimed to possess in the tests they created. Asking teachers solely about how they construct MCT may not fully reveal their actual skills, and there is a strong likelihood that their responses may not accurately reflect their actual practices.

The second approach to understanding teachers' adherence to test principles and test construction skills comprised evaluating some test samples developed by the teachers themselves (Kissi et al., 2023; Quansah et al., 2019). Quansah et al. (2019), for example, evaluated end-of-term questions of school teachers teaching integrated science, mathematics and social studies subjects. Typically, such tasks were designed by the school's teaching staff before being administered. The authors clarified that most, if not all, pre-tertiary institutions have assessment committees that review the assessment tasks designed by these teachers. This was also the case for Kissi et al. (2023), where the end-of-term examination papers were assessed for two courses (i.e., Business Management and Core Mathematics). Given these instances, the final product of the assessment tasks evaluated may not accurately reflect individual teachers' test construction skills.

The third approach adopted by previous scholars involves conducting a psychometric analysis of assessment tasks developed by teachers (Kissi et al., 2023; Rao et al., 2023; Kumar et al., 2021; Shaheen et al., 2018; D'Sar and Visbal-Dionaldo, 2017; Rush et al., 2016; Tarrant et al., 2006; Downing, 2002). All the authors who studied teachers' MCT development skills and their adherence to assessment guidelines relied heavily on classical test theory (CTT) analysis (Kissi et al., 2023; Rao et al., 2023; Kumar et al., 2021; Shaheen et al., 2018; D'Sar and Visbal-Dionaldo, 2017; Rush et al., 2016; Tarrant et al., 2006; Downing, 2002). While the CTT approach forms the basis of most test theories, there is a significant challenge in using this method to evaluate item quality and test construction skills. A key concern is that item properties are linked to the characteristics of the examinees who sat the test, meaning such analyses may not accurately reflect teachers' test construction skills (Allen and Yen, 2001; Crocker and Algina, 1986; Gulliksen, 1950; Magnusson, 1967). Unlike CTT, which also centres on observed test scores and assumes measurement error is consistent across ability levels, the item response theory (IRT) models the probability of a person's response to each item based on their underlying latent trait (ability) and the item parameters (Baker, 2001; Morizot et al., 2007). Therefore, analysing the psychometric properties of a test using CTT may not necessarily reflect the item quality but the examinees' characteristics.

Why do multiple-choice test development competencies matter?

A multiple-choice item comprises two main components: (1) the stem which is an introductory statement or a question which presents a problem to the examinee; and (2) a list of options, usually between two to five, and contains the correct response (known as the key) and distractors (known as incorrect response) (Nitko, 2001a; Quansah, 2018). Each section of the multiple-choice item plays a critical role in measuring students' achievements in any learning area; thus, it is essential to follow the established principles in developing every aspect of the item.

This study's focus on assessing teachers' test development competencies regarding multiple-choice tests is based on notable reasons. First, the MCT is pervasive and has widespread use across local, national, and international assessments. For example, Ghanaian school-level assessments utilise multiple-choice items in their classroom and semester/term examinations. Further, national assessments organised by the WAEC across West African countries' pre-tertiary levels have dedicated section of the assessment where the multiple-choice items are used (Kissi et al., 2023; WAEC, 2016). It is important to highlight that the multiple-choice items are also used in international assessments like the Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) (The Organisation for Economic Co-operation and Development [OECD], 2019).

Secondly, compared to the other test formats, the MCT strongly aligns with item and psychometric analyses that evaluate item quality and associated contextual realities (Baker, 2001). Therefore, the multiple-choice items offer an objective platform for assessing the psychometric properties of the test items. Additionally, the MCT is relatively efficient in large-scale assessments (characterised by large class sizes in Ghanaian classrooms) and still promises much higher, reliable, and valid assessment outcomes (Nitko, 2001a). Given these compelling reasons, teachers' test development skills on MCT are critical and should be a skill every teacher should possess. Teachers without this specific skill may be limited to their pedagogical roles, which could negatively impact learning outcomes and overall teaching quality.

The study's focus

This study employed a triangulation methodology to evaluate mathematics teachers' multiple-choice test construction skills, integrating three procedures to provide comprehensive insights into the issue under investigation. The central research question that guided this study is "What are the multiple-choice test development competencies demonstrated by JHS mathematics teachers in Ghana?" The adoption of the triangulation methodology helped evaluate the teachers' test development competencies through the lens of a multiple sequential data collection approach. First, we surveyed mathematics teachers to gather their self-reports on their adherence to test construction principles. Next, the study deliberately selected and assessed test samples created by the mathematics teachers to examine how they applied test construction principles in test development. Additionally, these teachers were asked to design new multiple-choice tests and administered them to their pupils. The third aspect of the study involved conducting an IRT psychometric analysis of the tests prepared by the teachers, aimed at evaluating their ability to design psychometrically sound assessments. Finally, we evaluated the relationship between adherence to recommended test construction principles and the quality of tests developed by the teachers.

The study's methodology is grounded in three key theoretical perspectives in educational assessment, namely, the assessment literacy framework, measurement errors in test theory, and validity theory. First, the assessment literacy framework highlights that teachers should possess both technical knowledge and practical skills in educational assessment (DeLuca, 2012; Popham, 2009; Stiggins, 1991). Within the literacy assessment model, teachers should not only possess knowledge about assessment principles but should be able to exhibit competence in applying them to design highly valid and reliable assessment tools. In this context, designing MCT is a critical aspect of assessment literacy, and evaluating teachers' competencies reflects both their strengths and gaps in professional practice. The assessment literacy model, therefore, provides a link to the triangulated methodology (i.e., three sequenced phases) used in the study design: (1) knowledge domain—Do mathematics teachers possess knowledge on the principles of MCT design? (2) Skill domain—Can mathematics teachers apply the MCT principles in developing real test items? and (3) Impact domain—Do the MCTs that the teachers designed produce sound psychometric indicators when subjected to item analysis?

Secondly, the consideration of measurement error highlights the limitations of the CTT in assessing item psychometric properties (Allen and Yen, 2001) and offers strong support for the use of IRT. Although the CTT assumes that the observed score is an additive component of the actual score and the error, it fails to capture item-level error variances across ability levels. This measurement error limitation is rigorously addressed by the IRT, enabling the precise evaluation of mathematics teachers' competencies in MCT development. The third theoretical consideration is the validity theory,

which emphasises that validity is not merely a characteristic of a test, but rather reflects the extent to which evidence is provided to support the interpretation of test scores (Cronbach and Meehl, 1955). Using multiple data collection approaches in this research is a strategy for assessing the validity of the MCT development competencies of mathematics teachers from multiple angles, thereby serving as multiple sources of evidence (i.e., self-reports, documentary analysis, and psychometric analysis).

Methods and materials

As already indicated, the study employed a triangulation research methodology; therefore, the methods section is structured into three phases: the survey phase, the documentary analysis stage, and the item analysis phase.

First phase: survey

Purpose

The purpose of the survey was to explore JHS mathematics teachers' adherence to MCT construction principles.

Participants

The population consisted of all teachers in public JHSs within the Sekondi-Takoradi Metropolitan Assembly (STMA) in Ghana. The STMA comprises nine (9) circuits, namely Sekondi, Adiembra, Takoradi East, Takoradi West, Takoradi Central, Ketan, Essikado, Nkroful-Nkansaworado, and Kojokrom Circuits. Records from the Metropolitan Directorate of Education (2023) indicated a total of 96 public JHSs across these nine circuits. The study focused on JHS teachers who teach mathematics, as mathematics is one of the subjects where students frequently underperform (Chief Examiners' Report, 2021; Etsey, 2006). The total population of mathematics teachers was 218.

Considering the size of the target population of JHS mathematics teachers, we employed the census method to include all the mathematics teachers in the survey (Amedahe and Asamoah-Gyimah, 2015). This approach was suitable because it eliminated sampling error, as every member of the target population participated in the study. Consequently, the sample size was 218 mathematics teachers. Of these, 71.1% were males, while 28.9% were females, indicating that the majority of the mathematics teachers in the STMA schools were males. Additionally, the majority of the teachers (61.0%) were between 20 and 30 years old, 32.1% were between 31 and 40 years old, 5.0% were between 41 and 50 years old, and 1.8% were between 51 and 60 years old.

Instrumentation

The primary tool for data collection at the survey phase was a questionnaire. The instrument gathered information on respondents' demographic characteristics (i.e., gender and age). In addition to the demographic data, the modified version of the Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC) validated by Kissi et al. (2023) was used for the survey data collection. The TTCCQ-MC scale consists of 19 items with 4-point Likert-type agreement options. The scale is multidimensional, featuring a three-factor structure: content validity (6 items), item options handling (6 items), and test item assembling (7 items). The instrument demonstrates high validity and is suitable for use. The internal consistency scores for the subscales are as follows: content validity = 0.79, item options handling = 0.78, and test item assembling = 0.76.

Procedure

The survey phase allowed respondent to report on their level of adherence to the MCT guidelines in test construction. Ethical approval was obtained from the Ethical Review Board at the University of Cape Coast, Ghana, covering all phases of the study. Additional permission was secured from the heads of schools to facilitate data collection. Prior contacts were made with the mathematics teachers to establish rapport and explain the purpose of the research. The available times for teachers at each school were identified to ensure a smooth data collection process. On the day of data collection, eligible teachers were gathered in an office, and the questionnaires were administered to them. All other ethical considerations, including informed consent, anonymity, confidentiality, and protection from harm, were strictly observed. The data collection process lasted approximately 3 weeks.

Data analysis

Data collected for the study were coded and entered with the help of SPSS software version 25.0. The survey data were further screened for outliers, missing data and inconsistent responses or data entry errors. Identified issues from the screening were rectified by consulting the original completed questionnaires. The confirmatory factor analysis (CFA) was used to analyse the data to understand the mathematics teachers' adherence to test construction principles. Parameters from the CFA output included standardised factor loading (β) with their associated p-values, squared multiple correlation (R^2) and Average Variance Extracted (AVE). Higher β values with significant p-values indicated the predominant MCT adherence behaviours demonstrated under a specific adherence domain. For example, under the content validity adherence sub-domain, the item with the highest β value (preferably 0.5 and above) is the most demonstrated adherence behaviour in that sub-domain. Regarding the AVE values, the sub-domain (e.g., content validity, item options handling) with the highest value (preferably 0.5 and above) showed the dimension that the mathematics teachers predominantly adhered to when designing the MCT. Higher R^2 values reflected teacher behaviours consistently demonstrated on that adherence dimension and vice versa.

Second phase: documentary analysis

Documentary analysis is a systematic method for reviewing or evaluating documents (i.e., test samples in this context) (Creswell, 2014). The documentary analysis phase had two sub-stages. The first sub-stage focused on the documentary analysis of tests already designed by the teachers. The second sub-stage of the documentary analysis phase was limited to allowing teachers to redesign a set of MCT under specified conditions. The first sub-stage aimed to examine mathematics teachers' naturally produced MCT papers without blueprint support. The second sub-stage, on the other hand,

aimed at assessing mathematics teachers' MCTs when provided with a validated item specification table. The subsequent description of the methods for this phase is outlined to capture these two sub-stages.

Participants and documents collection

Sub-stage 1

Nine teachers were purposively sampled and asked to submit copies of test samples they had already administered, along with the associated scheme of work. For each of the nine circuits, teachers who achieved the highest score on the TTCCQ-MC scale (administered in the survey phase) were chosen. This was based on the current literature, which indicates that teachers with high scores on this measure are more likely to develop quality test items and create practical tests (Asamoah-Gyimah, 2022; Ankomah, 2020; Quansah and Amoako, 2018). The two most recent samples of question papers were presented by each selected teacher, totalling 18 examination papers. The scheme of work for each teacher was also obtained prior to examining the samples. Overall, the data collection process took approximately 4 weeks. We assured the teachers of data privacy, confidentiality, and anonymity. Consent from the respective schools' examination boards was obtained before data were requested.

Sub-stage 2

We further engaged the nine selected teachers to identify the topics they would cover for the term and those content areas they had already covered. The content areas were limited to JHS 2 (grade 8) level for two reasons: (1) all nine mathematics were currently teaching the JHS 2 students, with a few of them doubling as mathematics teachers for JHS 1; and (2) while the JHS 3 students were preparing for their final national examination (Basic Education Certification Examination, BECE), the JHS 1 students had not covered much content for this exercise. The content areas included Numbers and Numerations, Algebra, Geometry and Trigonometry, Measurement, and Statistics and Probability. Based on these content areas, the item specification table (see Supplementary material) was developed and validated by a team of experts comprising two measurement specialists and three mathematics education experts.

The nine teachers who were purposively sampled were tasked with designing 20 multiple-choice items and administering them to a minimum of 150 students. Although in Ghana, JHS examinations typically contain 40 multiple-choice items in addition to essay questions, the decision to limit the MCT to 20 items was based on some practical and logistical considerations. Using 20 items reduced the workload and burden on teachers in designing the MCT since each teacher had to administer the MCT to at least 150 examinees. With 20 items and four options per item, this produces 80 distractors per teacher, which is sufficient for assessing the teachers' proficiencies in designing multiple-choice items in relation to distractor plausibility, grammar, cueing, and ambiguity, among others (Nitko, 2001b). While there is no universal number of items for psychometric analysis, some previous studies have demonstrated that 20 multiple-choice items are sufficient for item analysis (Licona-Chávez et al., 2020; Quansah and Cobbinah, 2021). The nine teachers were given the item specification table to guide them to construct the items. This was done because the quality of items developed by teachers was compared, and therefore, the item specification table helped create comparable test forms.

Documentary review and analysis

Sub-stage 1

The naturally developed set of MCTs was reviewed based on a deductive codebook from established test construction criteria, including test format and test construction errors (Nitko, 2001a, 2001b). The test format error domain encompassed logical formats and patterns of options, time allocated, hinting and linking items, item arrangement and spacing, and instructions. The test construction error indicator assessed the objectivity of test items, the representation of thinking skills, and the practicality, ambiguity, content relevance, and fairness of the test. Two experts jointly evaluated the codebook and co-coded a pilot sample of 25% of the items. The first expert was a measurement specialist with a background in mathematics, and the second expert was a measurement specialist with an educational background and experience in gender equality and social inclusion. Based on the pilot coding, we computed the inter-rater reliability for each binary category (Cohen's κ), yielding an overall reliability coefficient of 0.71.

The two coders independently engaged in complete coding of the 18 papers by applying the agreed codebook. Upon completion of the coding, the results of the coding were compared item by item. We acknowledge that there were instances of disagreement, especially on issues of flagging an item as ambiguous. Through consensus, disparities in coding were resolved through the log rule classifications. Finally, we checked for potential coder drift by randomly double-coding 30% of the papers, which yielded an inter-rater reliability (Cohen's κ) of 0.79. The documentary analysis followed a structured qualitative content analysis approach proposed by Schreier (2012).

Sub-stage 2

In this sub-stage 2, the newly designed test documents were reviewed and analysed following the same approach adopted in sub-stage 1, with slight changes. The evaluation criteria for reviewing the developed MCTs were based on four hypothetical dimensions (Nitko, 2001b). The first dimension is identifiable patterns of answers which reflect a situation where the correct responses to multiple-choice items follow a systematic sequence (e.g., ABCD, ABCD, ABCD) or repeated correct responses to the items (e.g., AAAA, BBBB, CCCC). This trend of correct responses allows test-wise students to guess the correct answer without demonstrating content mastery. Ambiguity, which is the second evaluation criterion, reflects a situation where an item structure and wording depict multiple interpretations for the examinees. Such misinterpretations introduce confusion and unintended difficulty, causing task performance to reflect traits other than the intended construct, a phenomenon known as construct-irrelevant variance.

The third dimension is unclear instructions, where represent a situation instructions to a test or an item are misleading, incomplete, or missing important details (Amedahe and Asamoah-Gyimah, 2013). Such situations lead to a lack of clarity or misinterpretation of the expectations of the assessment tasks. The last evaluation dimension is fairness, which reflects the extent to which the assessment task(s) provide all test-takers with equal chances to demonstrate their knowledge, free from biases associated with language, gender, socioeconomic background, or culture (OECD, 2019).

The same two coders were engaged to review and rehearse the codebook using one of the newly designed MCT sets. Following this exercise, we computed the overall inter-rater reliability which yielded a reliability coefficient of 0.77. The two coders independently coded all the MCTs with an overall inter-rater reliability of 0.81. We double-checked the item specification table to ensure that identified issues related to ambiguity and fairness were wording flaws and not blueprint non-adherence. The analysis also followed Schreier's (2012) structured qualitative content analysis.

Third phase: item analysis

Purpose

This phase aimed to evaluate the psychometric qualities of teacher-developed MCT items using IRT, in order to determine how well these items discriminated between high- and low-ability students.

Participants

The participants for this phase included the nine mathematics teachers and 1,350 learners who took part in the test developed using the item specification table. The newly designed instruments were administered to the students in accordance with all standard protocols for test assembling, proctoring, and administration. All identifiers of teachers' identities were removed from the data to ensure anonymity. Additionally, the identity of the test evaluated was concealed, and pseudonyms were assigned. This process, known as data de-identification, involved assigning names to the teachers, such as A, B, C, D, E, etc. The developed tests were obtained in digital format and manually validated to confirm they aligned with the item specification table.

Statistical analysis

The MCTs administered by the teachers under the supervision of the researchers were collected and entered into Microsoft Excel, creating nine sheets to accommodate all nine teachers. Prior to the entries, the scripts were manually scored; therefore, the data were entered as 0 or 1 to represent incorrect and correct, respectively. The data were screened and cleaned to prevent errors in the entry. The psychometric analysis was conducted using the 2-Parameter Logistic Model (2-PLM). The choice of the 2-PLM was guided by three key reasons: (1) model fit analysis of 1-PLM, 2-PLM and 3-PLM showed that five of the datasets were more aligned to the 2-PLM compared with the 1-PLM and 3-PLM; (2) the sample size within each dataset was not sufficient to perform 3-PLM which demands large sample size due to the guessing parameter; and (3) the guessing parameter is more tied to the test-taking behaviours of the examinees rather than the item quality (Embretson and Reise, 2000).

The 2-PLM was, therefore, used to analyse the quality of MCT items developed by the nine JHS mathematics teachers. The 2-PLM produces the item difficulty and item discrimination. The discrimination index in IRT is denoted by "a" and refers to the slope or steepness of the item characteristic curve. It indicated how well an item differentiates between low and high abilities. Discrimination indices ranged from 0 to positive infinity. In this study, discrimination indices were interpreted based on the following criteria: a discrimination index below 0.50 was considered poor, and an index equal to or above 0.50 was deemed good (Baker, 2001). Difficulty represents the point on the ability scale where an item functions, specifically where the likelihood of a correct response is 0.5. This assumption applies to 1-PLM and 2-PLM. For 3-PLM, difficulty is expressed as (1 + c)/2. Difficulty ranged from positive to negative infinity, and it is denoted by "b." When the item is easy, the difficulty index became negative, suggesting that the item is suitable for low-ability groups (Embretson and Reise, 2000). The IRTPro computer software was used to perform the IRT analysis.

Kendall's Tau-b correlation coefficient was also used to assess the relationship between adherence to recommended MCT construction principles and the quality of MCT produced by mathematics teachers. The use of Kandall's Tau-b was justified by the small sample size ($n = 9$), which was used for the correlation analysis (Howell, 2013). Additionally, the data on teachers' ranked adherence and test quality had ties; thus, using Kendall's Tau-b correlation, adjusted for the tied ranks, provided accurate estimates (Gibbons and Chakraborti, 2011). The correlation analysis was conducted using SPSS software version 25.0.

Results

First phase: survey results

Mathematics teachers' adherence to recommended principles of MCT construction

The focus of the analysis was not to validate the scale but to find out how much variance each specific factor contributed to the level at which JHS teachers adhere to MCT construction principles. Adherence, as earlier mentioned, was segregated into three dimensions: test item assembling, content validity and item option handling. The result on the level of JHS mathematics teachers' adherence is shown in Table 1.

Table 1 shows results on adherence to multiple-choice test construction principles. Regarding test item assembling, respondents indicated that they mostly review test items for construction errors ($\beta = 0.690$, $p < 0.001$), ensure all parts of an item are on the same page ($\beta = 0.657$, $p < 0.001$), and correctly space test items for easy reading ($\beta = 0.656$, $p < 0.001$). However, the teachers showed lower adherence in the area of numbering test items sequentially ($\beta = 0.183$, $p < 0.001$). Concerning content validity, teachers consistently presented clear items ($\beta = 0.730$, $p < 0.001$) and allocated sufficient time for test completion ($\beta = 0.636$, $p < 0.001$). Nonetheless, respondents indicated a low level of adherence regarding including questions with varying difficulty ($\beta = 0.297$, $p < 0.001$). In terms of item options and handling, teachers mostly avoided using "none of the above" as an option when an item is the best answer type ($\beta = 0.771$, $p < 0.001$), but they do not typically ensure options are grammatically consistent with the stem or present options in a logical order when possible ($\beta = 0.143$, $p < 0.001$; $\beta = 0.291$, $p < 0.001$).

Generally, the mathematics teachers showed a moderate level of adherence to MCT construction in ensuring test item assembling, item option management, and content validity. The AVE for these dimensions (test item assembling, item option management, and content validity) was 0.089, 0.203, and 0.316, respectively, indicating that the teachers' adherence to the MCT construction principle was

TABLE 1 Adherence to summative test construction principles.

Statements *When constructing multiple-choice tests, I.....*	β	*p*-value	R^2	AVE
Test item assembling	**0.535**	**<0.001**	**0.286**	**0.089**
give specific instructions on the test	0.537	<0.001	0.288	
use the appropriate number of test items	0.466	<0.001	0.217	
number the test items one after the other	0.183	0.019	0.033	
appropriately assign page numbers to the test	0.553	<0.001	0.305	
properly space the test items for easy reading	0.656	<0.001	0.430	
keep all parts of an item (stem and its options) on the same page	0.657	<0.001	0.431	
review test items for construction errors	0.690	<0.001	0.476	
Content validity	**0.538**	**<0.001**	**0.289**	**0.316**
match test items to instructional objectives (intended outcomes of the appropriate difficulty level)	0.536	<0.001	0.288	
make sure each item deals with an important aspect of the content area	0.470	<0.001	0.221	
prepare a marking scheme while constructing the items	0.564	<0.001	0.318	
pose unambiguous items	0.730	<0.001	0.533	
include questions of varying difficulty	0.297	<0.001	0.088	
give appropriate time for completion of the test	0.636	<0.001	0.404	
Item option handling	**0.399**	**<0.001**	**0.157**	**0.203**
include in the stem any word(s) that might otherwise be repeated in each option	0.337	<0.001	0.113	
make the options grammatically consistent with the stem	0.143	0.004	0.021	
make options independent of each other	0.386	<0.001	0.149	
avoid the use of "none of the above" as an option when an item is of the best answer type	0.771	<0.001	0.550	
make options approximately equal in length	0.467	0.003	0.218	
present options in some logical order (e.g., chronological, most to least, alphabetical) when possible	0.291	<0.001	0.084	

β-Factor Loadings, AVE, Average Variance Extracted.

more related to content validity than to test item assembling and item option management.

Second phase: documentary analysis results

Sub-stage 1—naturally designed MCTs without a blueprint

Test format errors

After carefully analysing samples of submitted tests, several test formats and construction errors were identified. The errors included test format issues, such as alternatives not presented in a logical order or a detectable pattern, incorrect answers, and test forms without an indicated completion time. *For instance, in one of the MCT samples, the options alternated between numerical values and words ("12; twelve, 14; fourteen)*. In another paper, *there was a clear pattern of answers (A, B, C, D repeated across the first 11 items)*.

Other problems involved wrong keys, clueing and linking items, implausible distractors, mixing horizontal with vertical options, page numbers not assigned, poor item spacing and arrangement, construct irrelevant variance, ambiguous items with more than one correct answer, instructional-related issues like missing or incomplete instructions, and the use of k-type items. For example, one of the test samples had "all the above" as the key, while not all preceding options were true, possibly causing ambiguity. While in some cases, the time allocation for task completion was not indicated, other test items had mixed vertically and horizontally arranged alternatives.

Test construction errors included thinking skills, content relevance, and fairness to different learners.

Thinking skills represented: Careful analysis of the MCT samples indicates that the majority of the items measured lower-level skills. Most of the items only required examinees to recall facts without demonstrating deeper understanding. One of the items asked "*What is the value of π (pi) to two decimal places?"* which is basically a recall question. Only a few measured understanding and application. For example, one of the teachers wrote a question, "*If the perimeter of a square is 24 cm, what is the length of one side?*" Tasks of this nature, which required application, were rarely found. Given these types of questions, it is likely that most of the items may favour test-wise examinees, failing to discriminate between examinees who have mastered the content and those who have not. This weakness aligns with Nitko's (2001a) observation that measuring different types of thinking skills (knowledge, understanding, applying, analysis, synthesis, and evaluation) increases test validity. Ultimately, the limited representation of higher-order skills is a threat to content validity.

Content relevance and fairness to different learners: Analysis of the MCT instruments showed that the teachers did not adequately sample

from all the test content areas as listed in the scheme of work for the relevant term. In one of the test samples, the greater proportion of the items focused on algebra manipulation; meanwhile, no task was developed on statistics, which formed part of the scheme of work. This imbalance could potentially reduce representativeness, thereby affecting content and construct validity (Nitko, 2001b). Aside from issues such as font size and style, as well as clueing and linking items, and detectable patterns, the assessment tasks did not provide a particular group of examinees with an advantage over others. With regards to fairness, no explicit issues of fairness were detected with none of the items showing gender-stereotyped and culturally-biased wording.

Sub-stage 2—newly designed MCTs by the teacher with blueprint support

Identifiable Patterns: Evaluation of the newly developed MCTs showed that most test items did not exhibit any identifiable answer patterns. However, exceptions were noted for Teachers A, D, and G, where identifiable answer patterns were observed. For example, Teacher D's scoring rubric indicated that about 15 out of 20 items had option "B" as the correct answer. Similarly, Teacher G's test had 11 items with option "A" as the correct answer. In this case, test takers selecting option "A" for items 1–20 would score 11 points. Moreover, choosing option "A" for items 1–10 would earn test takers a score of 8 out of 10. These situations illustrate how identifiable answer patterns could threaten test validity, allowing students to guess the correct option based on position instead of competence (Haladyna, 2004).

Ambiguity: Evaluation of the newly developed MCTs revealed that ambiguous items were widespread across all the mathematics teachers (i.e., Teacher A–I). The ambiguity took diverse forms, including distractors that could be interpreted differently among examinees, unclear mathematical expressions, and incomplete or unclear stems. For example, an item asked "*Find the value of x in the equation* $2x + ? = 10$" where the missing value (labelled,?) makes the stem incomplete and unclear. In one instance, an item had an option as "it depends," which does not reflect a mathematically valid answer. Other items had more than one correct answer (e.g., both $x = 2$ and $x = -2$ satisfy the equation), yet only one was marked as correct. These flaws could introduce construct-irrelevant variance, affecting students' achievement (Nitko, 2001a).

Unclear directions: Evaluation of the newly designed MCTs showed that while some teachers provided clear directions, most tests contained incomplete or missing instructions. For example, Teachers E, G, and C included directions specifying the class/grade level, subject area, type of test items, and how to select the correct answer. However, they omitted crucial details such as the test duration, which left examinees without guidance on how much time was allocated. By contrast, Teacher A's test lacked almost all essential instructions: there was no indication of the class/grade level, subject area, type of test items, answering procedures, or duration. The absence of such information can confuse students, disadvantage less test-wise examinees, and ultimately reduce the fairness and validity of the assessment (Amedahe and Asamoah-Gyimah, 2013; Nitko, 2001b).

Fairness to different students: Analysis of sampled tests showed that, overall, the newly developed MCT items did not systematically favour any particular group of examinees. Generally, the items were fair to candidates regardless of their gender, ethnic background, or socioeconomic status. However, some format and construction flaws (e.g., detectable answer patterns, poor spacing, and unclear fonts) could create disadvantages for students who struggle with reading (Nitko, 2001b). Gender-sensitive language was also an issue. Two teachers consistently used male names in their items. For example: "*John and Joseph are baking bread. The recipe calls for 3/4 cup of flour. They only have 3/8 cups of flour left. How many more cups of flour do they need to bake the bread?*" Although gender equality and social inclusion policies encourage using male names in fields like catering and sewing to promote male engagement, the exclusive use of male names does not reflect gender-sensitive assessment practices. Another item read: "*Kofi bought 36 banners. He bought an equal number of blue banners and gold banners. How many banners of each colour did Kofi buy?*" This kind of consistent male representation could inadvertently marginalise female students. Despite these concerns, most items provided all examinees with equal opportunities to demonstrate mathematical competencies. Issues such as font style, size, spacing, attractiveness, and uniformity were also noted, but these were judged to have minimal influence on fairness.

Third phase: item analysis results

Quality of MCT items developed by JHS mathematics teachers

Table 2 presents the results on the discrimination and difficulty indices of MCT developed by JHS teachers.

From the 2-PLM results, clear patterns emerge across teachers' item development. Teachers E, B, and I produced the highest-quality items, with at least half of their MCT items showing good discrimination between high- and low-ability learners (see Table 2). For example, teacher B's test had 50% of the items (i.e., items 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20) demonstrating good discrimination. In contrast, Teachers D, G, A, and C developed the poorest-quality items, with the majority of their items failing to discriminate effectively. Taking Teacher C's test, for example, twelve (12) of the items (items 1, 3, 5, 7, 9, 10, 11, 13, 15, 17, 18, 19, and 20) had poor discrimination. Teachers F and H showed mixed results, but overall, their items leaned towards poor discrimination. Broadly, the results indicate wide disparities in item-writing skills among JHS mathematics teachers, with only a few demonstrating a strong capacity to construct high-quality items.

In applying the 2-PLM, some items produced extremely large discrimination estimates (e.g., above 30), which are unlikely to reflect true item performance. In IRT, such values often indicate model misfit, sparse response data at certain ability levels, or potential coding/administration errors, rather than genuinely high discriminative power. The perceived causes of these extreme values included: (a) sparse data at certain ability levels, where too few students responded in a way that allowed stable estimation of the parameter, and (b) item flaws, such as poor wording, ambiguity, or guessability, which may have resulted in response patterns that do not conform to the expected IRT model. In this study, all parameter estimates, including the extreme values, were reported to maintain transparency. However, these extreme values were interpreted with caution, as they may have artificially inflated or distorted conclusions about item quality. Additionally, items with implausibly high discrimination values were flagged for further review rather than assuming that they represent high-quality test items.

10.3389/feduc.2025.1658971

TABLE 2 Item discrimination and difficulty index.

Item	1st Teacher (A)		2nd Teacher (B)		3rd Teacher (C)		4th Teacher (D)		5th Teacher (E)		6th Teacher (F)		7th Teacher (G)		8th Teacher (H)		9th Teacher (I)	
	a	*b*	*a*	*b*	*A*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	0.67	−5.96	0.08	0.38	−1.64	−0.55	0.33	2.77	3.81	−0.82	0.82	−0.39	100.2	0.37	−35.01	−1.31	0.64	−0.13
2	0.15	−15.9	1.10	0.73	1.00	−0.44	−1.25	−0.85	2.59	−0.59	1.94	−0.30	−1.05	−0.71	35.01	−1.31	2.58	0.28
3	0.95	−0.14	−1.27	−0.82	−0.50	0.15	1.09	−0.14	0.22	−0.60	−0.30	0.58	0.09	4.96	2.23	−0.91	0.63	0.99
4	−1.26	−1.33	1.04	0.61	0.96	−2.37	0.14	−0.10	−0.63	−1.40	0.16	2.89	0.58	0.69	0.03	−117.2	0.82	0.48
5	1.59	0.91	−1.59	−0.70	−1.21	−0.36	1.79	0.09	−0.29	−2.44	−1.14	−0.55	−0.31	−0.49	−2.60	1.72	0.27	2.15
6	−0.05	24.70	0.83	0.23	1.32	0.09	−0.45	−1.38	1.06	−0.50	1.45	−0.00	0.64	−0.77	0.57	0.11	0.44	0.12
7	−0.82	−0.15	−1.28	−0.70	−3.98	−0.55	0.63	0.93	0.21	1.11	−0.11	−2.61	−0.38	−1.78	0.69	−3.42	−0.41	−0.70
8	3.04	−0.66	2.38	0.46	1.57	0.18	−1.88	−0.95	2.33	−0.02	1.37	−0.10	1.04	0.48	0.66	−3.69	0.12	0.40
9	−0.37	5.29	−0.84	−0.93	−0.61	−0.86	0.80	0.85	1.75	−0.74	0.05	−8.87	−0.37	0.12	98.50	0.36	0.04	3.14
10	−0.53	0.87	1.53	0.41	0.60	0.34	−1.19	−0.64	0.29	−0.09	0.89	0.36	0.38	0.27	−0.86	0.22	0.09	0.54
11	−1.56	−0.80	−1.68	−0.55	−0.03	19.77	0.38	0.66	−0.53	−1.12	−1.08	−0.12	0.23	0.60	35.01	−1.31	0.47	0.61
12	−0.61	2.26	1.28	0.11	0.50	1.08	0.08	1.20	0.84	−0.96	1.27	−0.25	100.2	0.37	−31.24	1.38	1.17	0.26
13	−22.97	1.93	−1.09	0.12	−0.03	−19.78	0.15	6.06	2.80	−0.73	0.10	−22.4	−1.05	−0.71	−0.24	−5.22	0.41	1.03
14	2.30	−1.40	0.87	0.22	0.43	1.69	−0.16	−7.12	8.23	−0.63	3.55	−0.21	0.66	0.14	−0.63	0.82	0.47	−0.11
15	0.47	1.85	−1.46	−0.06	−0.02	−16.68	0.28	1.42	−0.33	−2.07	−0.83	−0.94	−0.01	−11.4	0.52	1.88	0.36	2.24
16	1.06	−1.23	1.51	0.11	0.87	0.40	−0.10	−11.62	1.78	−0.26	2.02	0.08	0.20	2.74	0.53	−2.36	0.62	0.03
17	71.80	−0.57	−1.60	−0.66	−0.38	−1.91	0.34	0.13	0.77	−0.78	−0.02	12.76	0.40	2.36	−1.64	0.92	12.36	0.02
18	−0.61	−1.87	1.76	0.51	0.20	2.98	−0.28	−3.52	0.19	5.57	0.12	5.21	−0.22	−3.11	−0.35	2.50	0.76	0.26
19	0.89	−2.02	−1.44	−0.51	−0.72	−1.09	0.20	3.76	1.78	−0.44	0.40	−0.72	0.80	0.77	0.35	3.32	1.37	−0.33
20	37.23	−1.04	1.33	−0.07	0.38	1.98	0.12	7.21	2.34	−0.59	0.69	−1.03	−0.09	−7.31	−1.23	0.03	1.00	0.22

The total information curves from the teachers indicated that nearly all the tests developed by the teachers provided little to no information about students' mathematics achievement for most of the pupils. It was evident from all the curves that the standard errors related to measuring students' ability were largely high across different ability levels (Figure 1). These elevated standard errors suggest that the tests created by teachers generally have low reliability.

The test characteristic curves also demonstrated that a non-definitive and counterintuitive relationship exists between ability and the likelihood of answering an item correctly. For instance, Teacher A's examinee with an underlying ability of 1.0 had a higher probability of answering an item correctly than examinees with an ability of 3.0 (Figure 2). This illustrates that, according to this model, very low-ability examinees can achieve a test score simply because the items are problematic (exhibiting low or negative discrimination). At a certain point, examinees with an underlying ability around −1.5 to −3.0 had the same probability of answering an item correctly. Additionally, test takers with theta levels of −1.4 and +1.9 had equal chances of answering an item correctly. Regarding the information function, the test items created by Teacher A provided little information for test takers with theta levels between −1.3 and +1.2, as well as those with theta levels between +1.7 and +2.2. Although the test characteristic curves for Teachers B and C did not reveal a direct relationship between actual scores and the ability scale, these test items offered valuable insights for test takers with a range of abilities.

Table 3 presents the summary results of the total number of quality items developed by JHS mathematics teachers. It was found that 75 (41.67%) of the items developed by JHS mathematics teachers were of good quality. However, about 105 (58.33%) of the items developed by JHS mathematics teachers were of poor quality. To better understand the characteristics of items developed by mathematics teachers, the frequency and percentage of items with negative discrimination were examined. It was further found that out of the one hundred and five (105) poor items, sixty-three (63) had negative discrimination. This indicates that, in such negative items, the probability of correct responses decreases as the ability increases from low to high.

Relationship between adherence to test construction principles and the quality of test items developed by JHS mathematics teachers

The quality of the item was conceptualised as the number of good and poor items created by the nine selected JHS mathematics teachers. The aim was to rank teachers' scores based on adherence, including several good and poor items, and to explore the relationship between these variables. Figure 3 displays descriptive statistics for the variables in question. Table 4 shows the results on the relationship between adherence to MCT construction principles and the quality of MCT developed by JHS mathematics teachers.

Figure 3 presents the results of the descriptive summary of the level of adherence and the number of quality items. The majority of teachers with higher average adherence scores developed a fairly good number

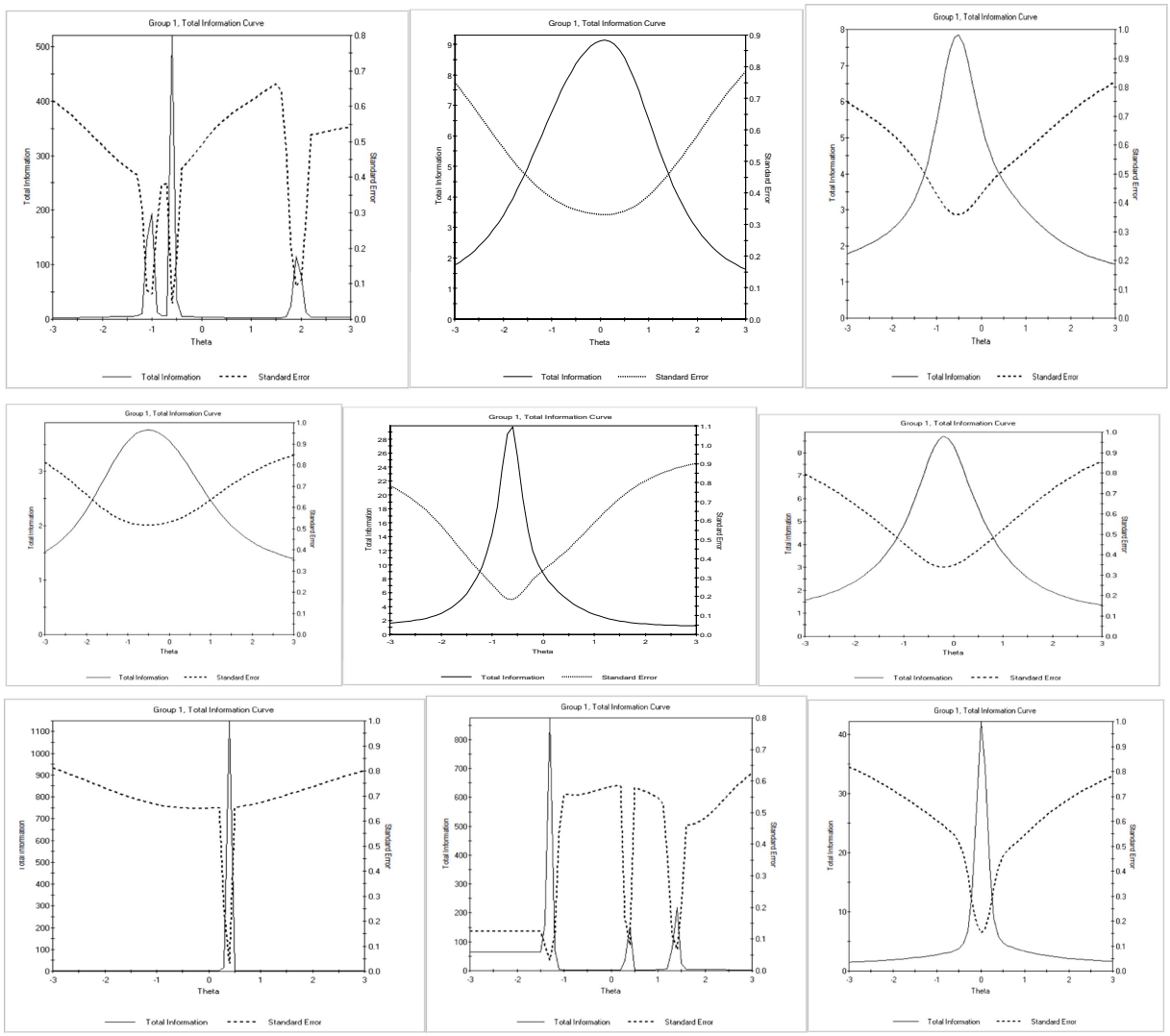

FIGURE 1
Total information curve of the test developed by Teacher 1 to Teacher 9.

of items. Interestingly, Teacher "F," who had the lowest adherence score, developed more good items as compared to Teachers G, D, C, and A, who obtained adherence scores of 3.38, 3.29, 3.31, and 3.25, respectively.

The result in Table 4 indicates that there exists a moderate statistically significant positive correlation between adherence to MCT construction principles and the quality of MCT developed by JHS mathematics teachers ($r = 0.569$, $n = 9$). This implies that high levels of adherence to recommended principles of MCT construction are associated with an increased number of quality items developed by JHS mathematics. It can also be said that a low level of adherence to MCT construction principles leads to a moderate to significant decrease in the number of quality items developed by JHS mathematics.

Discussion

The study assessed mathematics teachers' MCT development competencies using a triangulation methodology. We began with a survey of mathematics teachers' adherence to MCT construction principles, followed by documentary reviews of test samples, and concluded with psychometric analyses of newly developed MCTs. The discussion section has been structured to follow the three main phases of the research.

Survey of mathematics teachers' adherence to MCT construction principles

The findings from the survey phase showed that the teachers demonstrated a moderate level of adherence to MCT construction. Notably, the teachers, to some extent, ensured adherence to MCT principles in relation to test item assembling, content validity, and test option handling. The average variance extracted for each dimension also revealed that teachers adhere to recommended principles on ensuring content validity more than to other dimensions. While this appears positive from a theoretical perspective, failing to strictly

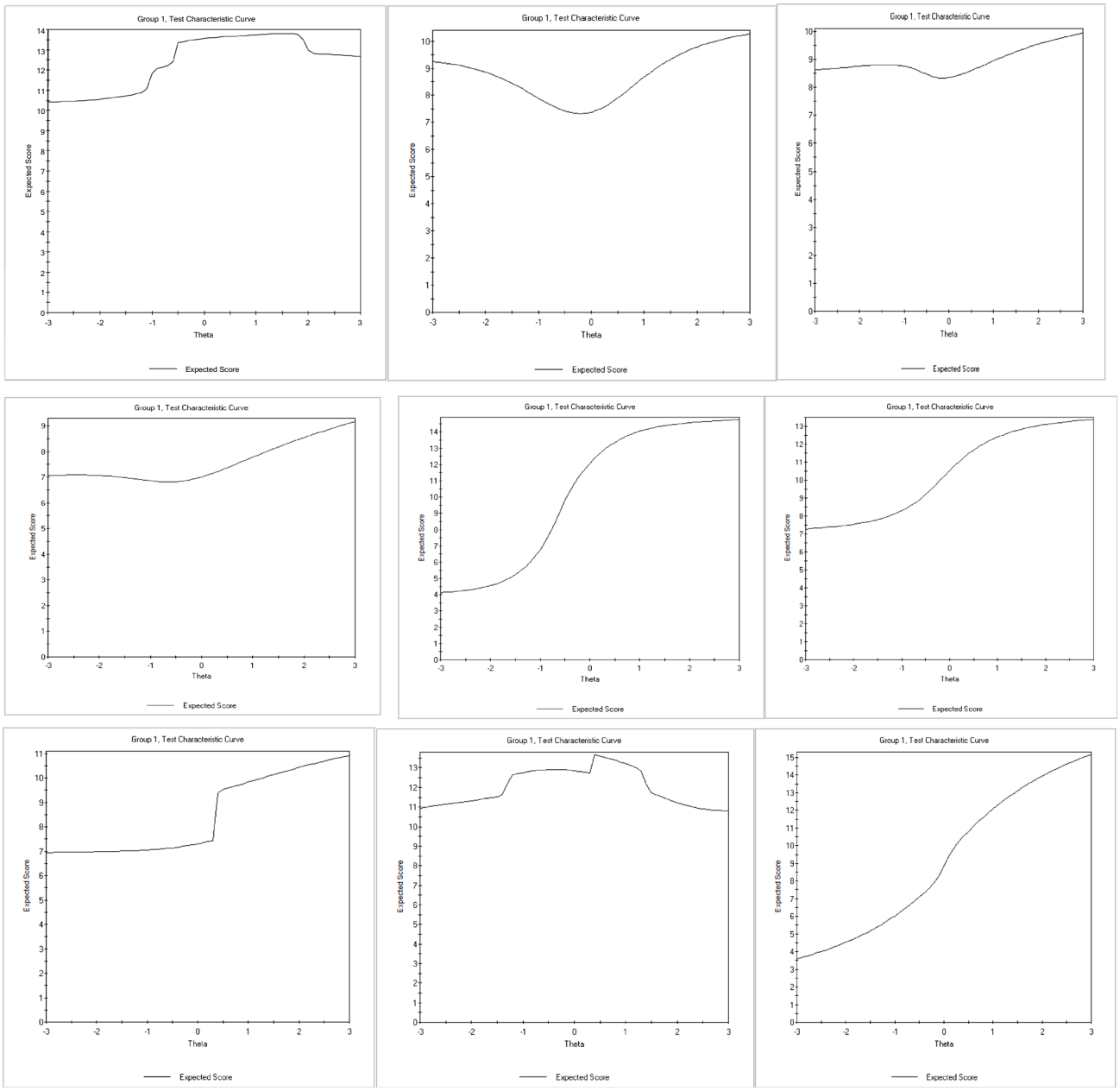

FIGURE 2
Test characteristic curve for test developed by Teacher 1 to Teacher 9.

follow all recommended principles in MCT construction could seriously lead to errors in MCT. Similar findings have been reported in previous studies, which found moderate to high levels of adherence to test construction principles (Kissi et al., 2023; Asamoah-Gyimah, 2022; Ankomah, 2020; Armah, 2018; Garba et al. (2015); Kinyua and Okunya, 2014; Agu et al., 2013; Oduro-Kyireh, 2008; Etsey, 2006). This has some implications for teachers' assessment of learners' learning and instructional effectiveness. High test adherence to recommended principles of MCT construction reduces errors in test development and administration that could affect the test's reliability and validity (Nitko, 2001b).

Documentary analysis of both existing and newly developed test samples of the mathematics teachers

The documentary analysis highlighted persistent weaknesses in the mathematics teachers' MCT development that directly affect validity and fairness. One recurring concern was the presence of identifiable answer patterns in some tests. Although not widespread, such systematic sequences of correct options can enable testwise examinees to guess answers without engaging with the content. Haladyna (2004) warns that answer patterns reduce item

TABLE 3 Summary of the number of quality items developed by JHS mathematics teachers.

Teacher	Good item, *N* (%)	Poor item, *N* (%)	*f*/% of Poor items with negative dis.
A	7 (35.0)	13 (65.0)	9 (69.23)
B	10 (50.0)	10 (50.0)	9 (90.0)
C	7 (35.0)	13 (65.0)	10 (76.92)
D	4 (20.0)	16 (80.0)	7 (43.75)
E	12 (60.0)	8 (40.0)	4 (50.0)
F	9 (45.0)	11 (55.0)	6 (54.55)
G	7 (35.0)	13 (65.0)	8 (61.54)
H	9 (45.0)	11 (55.0)	9 (81.81)
I	10 (50.0)	10 (50.0)	1 (10.0)
	75 (41.67)	105 (58.33)	63 (59.76)

f, Frequency; %, Percentage.

independence and ultimately compromise validity. The fact that these flaws occurred even when teachers used a blueprint suggests that knowledge of principles alone does not guarantee error-free test construction.

A more pervasive issue was ambiguity and construct-irrelevant variance. Ambiguous stems, vague distractors, and overly complex sentence structures made the items difficult to interpret, leaving them open to multiple possible answers. Previous studies have confirmed that grammatical and structural flaws confuse learners and lower test validity (Amedahe and Asamoah-Gyimah, 2013). The persistence of such problems across teachers underscores that many mathematics teachers have not yet mastered the linguistic and technical precision required for effective item writing. As Nitko (2001b) observes, attention to sentence structure, punctuation, and clarity is central to producing high-quality items. Without this, MCTs risk measuring reading or guessing ability rather than mathematical competence.

Another critical concern was the lack of clear test directions. While some teachers included partial guidance, many omitted key information, such as test duration or answering procedures. Directions are not mere formality; they shape how examinees approach test items. When omitted, they disadvantage less testwise students and threaten fairness (Amedahe and Asamoah-Gyimah, 2013). Although some examinees may ignore or scan through test directions, LeFevre and Dixon (1986) stressed that their consistent inclusion is essential, especially when new or unfamiliar formats are introduced. This indicates that teacher competence must extend beyond item-level construction to encompass test-level organisation.

In terms of fairness to diverse learners, the findings showed that most tests did not intentionally privilege one group over another in relation to gender, socio-economic status, or ethnicity. However, some teachers consistently used male names in contextual problems, which departs from gender-sensitive assessment practice. In addition, superficial formatting choices, such as inconsistent font size, spacing, or layout, though less critical than content or structure, can influence test validity by affecting readability and student engagement (Gabuyo, 2012).

Comparing the documentary analysis outcome for the existing test sample and the newly designed test samples revealed some interesting revelations. The item specification intervention used improved the alignment between test items and curriculum content. It was also observed that the recurring errors related to ambiguity and unclear directions reduced when the teachers were given the item specification table to guide the MCT development. Despite these improvements, the test development problem persisted, suggesting that even with structured support, some teachers struggled to translate guidelines into practice.

Psychometric qualities of the teachers' newly developed MCT

The psychometric analysis of the newly developed MCTs revealed that most mathematics teachers produced items of poor quality, with low or negative discrimination indices. Very few mathematics teachers achieved relatively better results, with about half of their items showing acceptable discrimination. These observed deficiencies mirror earlier studies documenting poor assessment practices and test quality among teachers (e.g., Tani, 2021; Friatma and Anhar, 2019; Hamafyelto et al., 2015; Adedoyin and Mokobi, 2013) and are consistent with Ghanaian evidence pointing to persistent gaps in classroom assessment competence (e.g., Akayuure, 2021; Quansah and Ankoma-Sey, 2020; Quansah et al., 2019; Garba et al., 2015; Kankam et al., 2014; Asare and Nti, 2014; Sofo et al., 2013; Quargrain, 1992).

A large proportion of the poor items exhibited negative discrimination, meaning that high-ability students were less likely to answer them correctly than low-ability students. This phenomenon is particularly damaging to test validity, as it reverses the intended relationship between ability and performance. Previous literature has attributed this to common item-writing problems such as implausible distractors, use of "all of the above," incorrect answer keys, ambiguous stems, or poor formatting (Downing, 2002; Kissi et al., 2023; Tarrant et al., 2006; Case and Swanson, 2002; Haladyna, 2004; Kumar et al., 2021; D'Sar and Visbal-Dionaldo, 2017). These weaknesses point to insufficient teacher expertise in constructing psychometrically sound items.

The findings also resonate with broader critiques of teacher preparation and professional development. Despite government initiatives, in-service training often emphasises lesson planning, pedagogy, and classroom management rather than assessment literacy (Quansah and Ankoma-Sey, 2020). As Mannion et al. (2018) and Odukoya et al. (2017) observed, many teachers perceived MCT development as time-consuming and burdensome, and few had received formal training in item analysis. Consequently, teachers may rely on recycled items or poorly constructed new ones (Onyechere, 2000; Ebinye, 2001), confirming the gap between theoretical knowledge of test construction and practical competence (Asamoah-Gyimah, 2022).

The test characteristic curves (TCCs) and information functions provided further insight into item quality. For some teachers, the TCCs did not follow the expected relationship between ability and actual score, reflecting the influence of negatively discriminating items. In some cases, examinees with low ability had equal or greater probabilities of answering items correctly compared with high-ability students, which is psychometrically illogical. The information functions also revealed that many tests provided

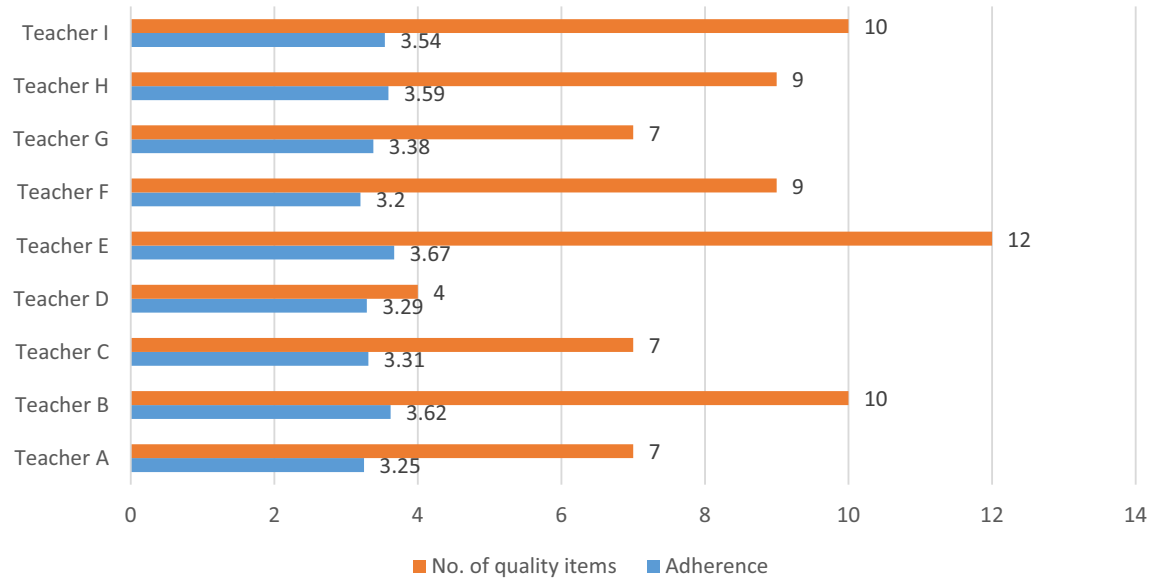

FIGURE 3
Level of adherence and number of quality items.

TABLE 4 Relationship between adherence to MCT construction principles and the quality of MCT developed by JHS mathematics teachers.

Variables	*n*	Kendall's *r*	*p*-value
Adherence—no. of quality items	9	0.569	0.041**

**Correlation is significant at the 0.01 level.

limited measurement precision across ability levels, though a few offered more useful information for decision-making. These findings are consistent with earlier studies that linked poor item discrimination to ineffective distractors, vague stems, and construct-irrelevant variance (Adedoyin and Mokobi, 2013; Frey, 2007).

Triangulation insight across findings from the three phases

The triangulated design of this study provided a comprehensive understanding of teachers' competence in MCT development. While the survey findings (Phase 1) suggested that teachers were aware of the principles underlying quality test construction, the documentary analysis (Phase 2) and psychometric evidence (Phase 3) demonstrated persistent flaws in actual practice. Errors, such as identifiable patterns, ambiguity, and unclear directions, were reflected in the poor psychometric performance of items, many of which displayed low or negative discrimination. This convergence highlights the disparity between teachers' self-perceived competence and their actual ability to construct valid tests, a challenge well-documented in Ghana and beyond (Quansah and Amoako, 2018; Akayuure, 2021). At the same time, the phases complemented each other: the survey provided insight into teachers' awareness of test development principles, the documentary analysis captured observable practices, and the psychometric analysis statistically confirmed item weakness. Although a small group of teachers produced higher-quality items, the overall pattern highlights the need for targeted professional development that links theoretical knowledge with practical skills in test construction.

The findings from the correlational analysis provided an important, second-layer point of integration between teachers' reported adherence to MCT development principles and the quality of their developed items. A moderate, significant positive correlation indicated that teachers who adhered more closely to recommended test construction principles tended to produce a greater number of quality items. This finding is consistent with Garba et al., (2015), who demonstrated that good items are contingent upon systematic adherence to established guidelines such as clarifying assessment purposes, defining content and objectives, using unambiguous stems, ensuring appropriate item formats, and conducting item analysis. At the same time, individual cases in this study complicate the pattern; one of the teachers, despite a low adherence score, developed more quality items than most colleagues with higher adherence levels. This divergence suggests that while adherence is generally associated with better outcomes, other factors such as prior experience, test-taking context, or individual skill may also influence item quality. Generally, the correlation strengthens the conclusion that systematic adherence to test construction principles enhances item validity and reliability. However, it also underscores the need for practical training that ensures knowledge translates into consistent practice.

The assessment literacy framework provides a meaningful lens for understanding the triangulated findings from the study. The framework emphasises that teachers should not only possess knowledge of assessment principles but also be able to demonstrate competencies in applying them to design highly valid and reliable

assessment tools (DeLuca, 2012; Popham, 2009; Stiggins, 1991). The triangulated findings portray a consistent gap between teachers' conceptual knowledge of MCT development and their practical competence in developing psychometrically sound MCT. The positive association between adherence to principles and item quality further reinforces the framework's central claim that without deep assessment literacy, teachers cannot reliably generate valid evidence of student learning. This connection highlights the importance of embedding assessment literacy more explicitly into teacher education, professional development, and education policy.

Implications for policy and practice

The findings of this study have several important implications for teacher education, professional development, and assessment policy in Ghana and similar contexts:

1. The consistent presence of flawed multiple-choice items highlights the need to embed assessment literacy, including practical test construction and item analysis, into both pre-service and in-service teacher training programmes. Teacher preparation institutions should revise their curricula to include more comprehensive assessment courses. At the same time, the National Teaching Council (NTC) should incorporate classroom assessment competence evaluation into teacher licensing and certification requirements. This action is necessary, as current professional development often emphasises pedagogy and classroom management (Quansah and Ankoma-Sey, 2020), but dedicated modules on item writing and psychometric evaluation are urgently needed.
2. The gap between teachers' self-reported knowledge (survey findings) and their demonstrated practices (documentary and psychometric analyses) indicates that theoretical awareness alone is insufficient. The District Directors of Education, with support from the Ghana Education Service (GES) and the National Council for Curriculum and Assessment (NaCCA), should prioritise hands-on workshops where teachers design items, receive structured feedback, and conduct item analysis.
3. The item specification table improved content coverage but did not entirely prevent flaws such as ambiguity or unclear directions. GES and NaCCA should mandate the use of test blueprints in school-based assessments, embedding important exemplars in the curriculum. Additionally, subject panels, district-level assessment officers, and examination boards should provide expert validation of teacher-developed items to enhance both content representativeness and technical accuracy.
4. Although most tests were broadly fair, gender-insensitive language and weak formatting were observed. GES and NaCCA should issue clear guidelines on gender-sensitive language (including linguistic flaws) and inclusive assessment practices. Headteachers and school assessment committees should monitor the appearance and formatting of tests to ensure consistency in font, spacing, and layout, which directly affects readability and fairness (Gabuyo, 2012).
5. The Ministry of Education and other key educational policymakers should integrate classroom assessment standards into teacher appraisal systems, school inspection frameworks, and continuing professional development (CPD). Headteachers should be tasked not only with counting the number of assessments administered but also with monitoring the quality of items. This would promote accountability and break the cycle of weak assessment practices.

Limitations

The study's strength lies in the triangulation methodology adopted, which combines three major approaches to exploring the mathematics teachers' MCT development competencies in a municipality in Ghana. Despite this strength, some limitations exist. First, the survey phase relied on self-reported adherence behaviours of mathematics teachers, which could be influenced by social desirability bias and may lead to overestimating competence. While triangulation with content and item analysis helped reduce this risk, self-report bias cannot be eliminated. Additionally, using purposive sampling with only nine teachers limits the range of perspectives and might not fully represent the diversity of practices among JHS mathematics teachers in Ghana (Dzakadzie and Quansah, 2023).

Moreover, since the study is context-specific, the findings might not be fully applicable to other subject areas, educational levels, or regions. These limitations should be kept in mind when interpreting the results, and future studies could include primary school teachers or teachers of other JHS subjects to broaden the sample. With the 2-PLM, some items yielded tremendous discrimination values, which are unlikely to reflect actual item quality. Such estimates are often symptomatic of underlying problems rather than genuine high discriminative power. Future studies should address this limitation by setting plausible parameter bounds during estimation, exploring alternative IRT models (e.g., 3PL to account for guessing), or re-examining raw data for errors.

Conclusion

This study demonstrates the value of employing a triangulated methodological approach integrating different methodologies to gain a holistic understanding of mathematics teachers' competence in MCT development. The findings collectively highlight a critical paradox: while teachers demonstrated awareness of MCT construction principles, this knowledge often failed to translate into practice. The study, therefore, raises an urgent call for systematic investment in assessment literacy. Poorly developed MCTs are not trivial errors; they distort measurement, compromise fairness, and lead to flawed educational decisions that affect students, families, and the broader education system. Given the widespread use of MCTs in Ghana and beyond for high-stakes and classroom-based decisions, building teacher capacity in test development and item analysis is both a pedagogical and policy imperative. Without targeted reforms in teacher education, professional development, and assessment policy, the potential of MCTs as objective, scalable, and psychometrically robust tools will remain unrealised, perpetuating cycles of weak assessment practice and undermining the quality of educational outcomes.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by College of Education Studies, Ethical Review Board, University of Cape Coast, UCC, Ghana. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

NQ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. FQ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. YD: Conceptualization, Data curation, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1658971/full#supplementary-material

References

Abreh, M. K., Owusu, K. A., and Amedahe, F. K. (2018). Trends in performance of WASSCE candidates in the science and mathematics in Ghana: perceived contributing factors and the way forward. *J. Educ.* 198, 113–123. doi: 10.1177/0022057418800950

Adedoyin, O. O., and Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple-choice examination test items. *Int. J. Asian Soc. Sci.* 3, 992–1011.

Agu, N. N., Onyekuba, C., and Anyichie, A. C. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: need for a test construction skill inventory. *Educ. Res. Rev.* 8:431.

Akayuure, P. (2021). Classroom assessment literacy levels of mathematics teachers in Ghanaian senior high schools. *Contemp. Math. Sci. Educ.* 2:ep21013. doi: 10.30935/conmaths/11286

Allen, M. J., and Yen, W. M. (2001). Introduction to measurement theory. Long Grove, IL: Waveland Press.

Amedahe, F. K. (1993). Test construction practices in secondary schools in the central region of Ghana. *Oguaa Educator* 2, 52–63.

Amedahe, F. K., and Asamoah-Gyimah, K. (2013). Introduction to measurement and evaluation. Cape Coast, Ghana: UCC Printing Press.

Amedahe, F. K., and Asamoah-Gyimah, K. (2015). Introduction to educational research. Cape Coast: UCC Printing Press.

Amoako, S. O. A., Suraj, N., Boamah, J. Y., and Rashid, B. (2024). Exploring the perceptions and interest of senior high school students' achievement in mathematics at Adansi-North, Ghana. *Int. J. Novel Res. Phys. Chem. Math.* 11, 33–49. doi: 10.5281/zenodo.10893967

Anhwere, Y. M. (2009). Assessment practices of teacher training college tutors in Ghana. University of Cape Coast, Cape Coast: (Unpublished Master's thesis). doi: 10.5281/zenodo.10893967

Ankomah, F. (2020). Predictors of adherence to test construction principles: the case of senior high school teachers in Sekondi-Takoradi Metropolis (Master's dissertation). Cape Coast, Ghana: University of Cape Coast.

Ankomah, F., and Nugba, R. M. (2020). Validation of Test Construction Skills Inventory through the lens of Item Response Theory (IRT). *American Journal of Creative Education*, 3, 86–100.

Ankoma-Sey, V. R., Asamoah, D., Quansah, F., and Aheto, K. S. (2019). Factors affecting junior high school pupils' performance in mathematics in Cape Coast Metropolis, Ghana. *Staff Educ. Dev. Int.* 24, 128–138.

Ansah, J. K., Quansah, F., and Nugba, R. M. (2020). 'Mathematics achievement in crisis': modelling the influence of teacher knowledge and experience in senior high schools in Ghana. *Open Educ. Stud.* 2, 265–276. doi: 10.1515/edu-2020-0129

Armah, C. (2018). Test construction and administration practices among lecturers and staff of the examinations unit of the University of Cape Coast (Doctoral dissertation). Cape Coast, Ghana: University of Cape Coast.

Asamoah-Gyimah, K. (2022). Influence of knowledge of assessment on test construction skills among tutors of College of Education in Ghana. *Am. J. Educ. Pract.* 6, 60–71. doi: 10.47672/ajep.1177

Asare, K. B., and Nti, S. K. (2014). Teacher education in Ghana: a contemporary synopsis and matters arising. *SAGE Open* 4, 1–8. doi: 10.1177/2158244014529781

Baker, F. B. (2001). The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation Original work published in 1985.

Baker, J. O. (2003). Testing in modern classrooms. London: George Allen and Urwin Ltd.

Bosson-Amedenu, S. (2017). Predictive validity of mathematics mock examination results of senior and junior high school students' performance in WASSCE and BECE in Ghana. *Asian Res. J. Math.* 3, 1–8. doi: 10.9734/ARJOM/2017/32328

Case, S. M., and Swanson, D. B. (2002). Constructing written test questions for the basic and clinical sciences. *3rd* Edn. Philadelphia: National Board of Medical Examiners, 31–66.

Chief Examiners' Report. (2021). 2021 BECE mathematics. The West African Examinations Council. Available online at: https://kuulchat.com/bece/chief_examiners_report/2021%20Mathematics.pdf

Creswell, J. W. (2014). A concise introduction to mixed methods research. Thousand Oaks, CA: SAGE Publications.

Crocker, L., and Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

DeLuca, C. (2012). Preparing teachers for the age of accountability: toward a framework for assessment literacy in teacher education. *Action Teach. Educ.* 34, 576–591. doi: 10.1080/01626620.2012.730347

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make any difference? *Acad. Med.* 77, S103–S104. doi: 10.1097/00001888-200210001-00032

D'Sar, J. L., and Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discri-mination Index and Distractor Efficiency. *Internation-al Journal of Nursing Education*, 9, 109–114.

Dzakadzie, Y., and Quansah, F. (2023). Modelling unit non-response and validity of online teaching evaluation in higher education using the generalizability theory approach. *Front. Psychol.* 14:1202896. doi: 10.3389/fpsyg.2023.1202896

Ebinye, P. O. (2001). Problems of testing under the continuous assessment programme. *J. Qual. Educ.* 4, 12–19.

Embretson, S. E., and Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Etsey, Y. K. A. (2006). Assessment in education. (Unpublished). Cape Coast, Ghana: University of Cape Coast.

Fives, H., and DiDonato-Barnes, N. (2013). Classroom test construction: the power of a table of specifications. *Pract. Assess. Res. Eval.* 18:n3. doi: 10.7275/cztt-7109

Fletcher, J. (2018). Performance in mathematics and science in basic schools in Ghana. *Acad. Discourse Int. J.* 10, 1–18.

Frey, B. B. (2007). Coming to terms with classroom assessment. *J. Adv. Acad.* 18, 402–423. doi: 10.4219/jaa-2007-495

Friatma, A., and Anhar, A. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *J. Phys. Conf. Ser.* 1387:012063. doi: 10.1088/1742-6596/1387/1/012063

Gabuyo, Y. A. (2012). Assessment of learning I. Manila, Philippines: Rex Book Store. Inc.

Garba, A. G., Musa, D. J., and Salihu, K. T. (2015). A scheme for assessing technical teachers' competencies in constructing assessment instruments in Technical Colleges in Gombe State. *ATBU Journal of Science, Technology and Education*, 3, 1–8. Available at: https://www.atbuftejoste.com.ng/index.php/joste/article/view/84

Gibbons, J. D., and Chakraborti, S. (2011). Nonparametric statistical inference. *5th* Edn. Boca Raton, FL, United States: Taylor & Francis Ltd. 10: 9781439896129.

Gichuru, L. M., and Ongus, R. W. (2016). Effect of teacher quality on student performance in mathematics in primary 6 national examination: a survey of private primary schools in Gasabo District, Kigali City, Rwanda. *Int. J. Educ. Res.* 4, 237–260.

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika* 15, 259–269. doi: 10.1007/BF02289042

Haladyna, T. M. (2004). "MC formats" in Developing and validating multiple-choice test items. ed. T. M. Haladyna. *3rd* ed (Mahwah, New Jersey: Lawrence Erlbaum Associates).

Hamafyelto, R. S., Hamman-Tukur, A., and Hamafyelto, S. S. (2015). Assessing teacher competence in test construction and content validity of teacher-made examination questions in commerce in Borno state, Nigeria. *J. Educ.* 5, 123–128. doi: 10.5923/j.edu.20150505.01

Howell, D. C. (2013). Statistical methods for psychology. *8th* Edn. Belmont, CA: Wadsworth Cengage Learning.

Iddrisu, I., Appiahene, P., Appiah, O., and Fuseini, I. (2023). Exploring the impact of students demographic attributes on performance prediction through binary classification in the KDP model. *Knowl. Eng. Data Sci.* 6:24. doi: 10.17977/um018v6i12023p24-40

Kankam, B., Bordoh, A., Eshun, I., Kweku Bassaw, T., and Yaw Korang, F. (2014). Teachers' perception of authentic assessment techniques practice in social studies lessons in senior high schools in Ghana. *Int. J. Educ. Res. Inf. Sci.* 1, 62–68. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/5454

Kinyua, J., and Okunya, E. (2014). Validity and reliability of teacher-made tests: a case study of year 11 physics in Nyahururu District of Kenya. *African Educ. Res. J.* 2, 61–71. Available at: https://eric.ed.gov/?id=EJ1216886

Kissi, E., Baidoo-Anu, D., Anane, E., and Annan-Brew, R. K. (2023). Teachers' test construction competencies in an examination-oriented educational system: exploring teachers' multiple-choice test construction competence. *Front. Educ.* 23, 1–14. doi: 10.3389/feduc.2023.1154592

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, R., and Prasad, R. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Med J Armed Forces India.* 77, S85–S89. doi: 10.1016/j.mjafi.2020.11.007

LeFevre, J., and Dixon, P. (1986). Do written instructions need examples? *Cogn. Instr.* 3, 1–30. doi: 10.1207/s1532690xci0301_1

Licona-Chávez, A. L., Montiel Boehringer, P. K. T. P., and Velázquez-Liaño, L. R. (2020). Quality assessment of a multiple-choice test through psychometric properties. *MedEdPublish* 9, 1–12. doi: 10.15694/mep.2020.000091.1

Magnusson, D. (1967). An analysis of situational dimensions. *Percept. Mot. Skills* 32, 851–867. doi: 10.2466/pms.1971.32.3.851

Mannion, C. A., Hnatyshyn, T., O'Rae, A., Beck, A. J., and Patel, S. (2018). Nurse educators and multiple-choice examination practices. Available online at: http://hdl.handle.net/1880/108887 (Accessed December 6, 2024).

Metropolitan Directorate of Education. (2023). Dataset on mathematics teachers in the Sekondi-Takoradi Metropolis. Unpublished data set.

Mills, E. D., and Mereku, D. K. (2016). Students' performance on the Ghanaian junior high school mathematics national minimum standards in the Effutu municipality. *African J. Educ. Stud. Math. Sci.* 12, 25–34. Available at: https://www.ajol.info/index.php/ajesms/article/view/169003

Morizot, J., Ainsworth, A. T., and Reise, S. (2007). "Toward modern psychometrics: application of item response theory models" in Handbook of research methods in personality psychology. eds. R. W. Robins, R. C. Fraley, & R. Krueger (New York, NY: Guilford Press). 407–423.

Nitko, A. J. (2001a). Educational assessment of students. New Jersey: Prentice Hall.

Nitko, A. J. (2001b). Conceptual frameworks to accommodate the validation of rapidly changing requirements for assessments. *Curr. Assess.* 1, 143–163. Available at: https://repository.bbg.ac.id/bitstream/565/1/Curriculum_and_Assessment.pdf#page=151

Odukoya, J. A., Adekeye, O., Igbinoba, A. O., and Afolabi, A. (2017). Item analysis of university-wide multiple-choice objective examinations: the experience of a Nigerian private university. *Eur. J. Methodol.* 52, 983–997. doi: 10.1007/s11135-017-0499-2

Oduro-Kyireh, G. (2008). Testing practices of senior secondary school teachers in the Ashanti region of Ghana. [Unpublished master's thesis]. Cape Coast, Ghana: University of Cape Coast.

OECD (2019). "An OECD learning framework 2030" in The future of education and labour (Cham: Springer International Publishing), 23–35.

Onyechere, I. (2000). New face of examination malpractice among Nigerian youths. *The Guardian*.

Oppong, S., Nugba, R. M., Asamoah, E., Quansah, N., and Ankoma-Sey, V. R. (2023). Teachers' confidence in classroom assessment practices: a case of basic schools in the upper Denkyira West District, Ghana. *Eur. J. Educ. Stud.* 10, 148–159. doi: 10.46827/ejes.v10i11.5063

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? *Theory Pract.* 48, 4–11. doi: 10.1080/00405840802577536

Quansah, F. (2018). Traditional or performance assessment: what is the right way to assessing learners? *Res. Humanit. Soc. Sci.* 8, 21–24. Available at: https://www.iiste.org/Journals/index.php/RHSS/article/view/40787

Quansah, F., and Amoako, I. (2018). Attitude of senior high school teachers toward test construction: developing and validating a standardised instrument. *Res. Humanit. Soc. Sci.* 8, 25–30.

Quansah, F., Amoako, I., and Ankomah, F. (2019). Teachers' test construction skills in senior high schools in Ghana: document analysis. *Int. J. Assess. Tools Educ.* 6, 1–8. doi: 10.21449/ijate.481164

Quansah, F., and Ankoma-Sey, V. R. (2020). Evaluation of pre-service education programme in terms of educational assessment. *Int. J. Res. Teacher Educ.* 11, 56–69.

Quansah, F., and Cobbinah, A. (2021). Equivalence of parallel tests in a basic statistics course in higher education using classical measurement theory. *Can. J. Educ. Soc. Stud.* 1, 13–28. doi: 10.53103/cjess.v1i2.11

Quaigrain, A. K. (1992). Teacher competence in the use of essay tests: A study of secondary schools in the Western Region of Ghana. *Unpublished Master's Thesis*, University of Cape Coast, Ghana.

Rao, M. C., Sreedhar, P., Bhanurangarao, M., and Sujatha, G. (2023). "Automatic multiple-choice question and answer (MCQA) generation using deep learning model" in International conference on information and management engineering (Springer Nature Singapore: Singapore), 1–8.

Rush, R. B., Rankin, C. D., and White, J. B. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med. Educ.* 16, 1–10. doi: 10.1186/s12909-016-0773-3

Schreier, M. (2012). Qualitative content analysis in practice. Thousand Oaks, CA: Sage.

Shaheen, S., Elmardi, A., and Ahmed, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine, Khartoum University. *Khartoum Med. J.* 11, 1477–1486.

Sofo, S., Ocansey, R. T., Nabie, M. J., and Asola, E. F. (2013). Assessment practices among secondary physical education teachers in Ghana. *Int. Online J. Educ. Sci.* 5, 274–281. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/7724

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan* 72, 534–539.

Tani, W. E. (2021). Test quality and students' performances: an appraisal of the national achievement test in English and mathematics for Cameroon primary schools. *Eur. J. Educ. Stud.* 8, 299–312. doi: 10.46827/ejes.v8i8.3861

Tarrant, M., Knierim, A., Hayes, S. K., and Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high-stakes nursing assessments. *Nurse Educ. Pract.* 6, 354–363. doi: 10.1016/j.nepr.2006.07.002

The Organisation for Economic Co-operation and Development. (2019). PISA 2018 assessment and analytical framework. Paris, France: OECD Publishing.

WAEC. (2016). Chief examiners' report on further mathematics. WAEC e-learning portal. Available online at: https://www.waeconline.org.ng/e-learning/Further/Furthmain.html (Accessed April 8, 2024).