TYPE Original Research
PUBLISHED 06 April 2023
DOI 10.3389/feduc.2023.1154592

Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence

Prosper Kissi[1], David Baidoo-Anu[2]*, Eric Anane[3] and Ruth Keziah Annan-Brew[4]

[1]Directorate of Academic Planning and Quality Assurance, University of Cape Coast, Cape Coast, Ghana, [2]Faculty of Education, Queen's University, Kingston, ON, Canada, [3]Institute of Education, University of Cape Coast, Cape Coast, Ghana, [4]Department of Education and Psychology, University of Cape Coast, Cape Coast, Ghana

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Stefan Daniel Keller,
University of Teacher Education Zuerich, Switzerland
Fernando José Sadio-Ramos,
Instituto Politécnico de Coimbra, Portugal

*CORRESPONDENCE
David Baidoo-Anu
✉ baidooanu.david@queensu.ca

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 30 January 2023
ACCEPTED 10 March 2023
PUBLISHED 06 April 2023

CITATION
Kissi P, Baidoo-Anu D, Anane E and Annan-Brew RK (2023) Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence.
Front. Educ. 8:1154592.
doi: 10.3389/feduc.2023.1154592

This study explored the relationship between multiple choice test construction competence and the quality of multiple-choice tests among senior high school teachers in Ghana. In all, 157 teachers were selected from four senior high schools in the Kwahu-South District. Participants responded to self-designed questionnaire developed to assess teachers' multiple-choice items construction competencies. A three-factor structure emanated from the exploratory factor analysis on teachers' multiple choice test construction competence—content validity, item "options" handling, and test items assembling. Teachers in this study perceived more competence in ensuring content validity, followed by test item assembling, and handling of "options" (that is, alternatives) of the test items. The study also found serious problems with copies of multiple-choice items teachers have constructed for the students. Findings from this study provide unique and compelling evidence regarding teachers' perceived test construction competence and analysis of their multiple-choice tests. Implications for policy and practice are discussed.

KEYWORDS
assessment, test construction, Multiple-choice test, teachers, Ghana, senior high schools

Introduction

Classroom assessment plays an instrumental role in supporting and improving teaching and learning (Black and Wiliam, 1998, 2010). As part of the tools used in classroom assessment, teacher-made tests play a crucial role in the assessment process. That is, teacher-made tests aid in pre-assessment (the assessment of what students already know before teaching), formative assessment (the assessment of student performance incorporated into the act of teaching), and summative assessment (the assessment of student learning at the end of some instructional period) of students' learning outcomes (Gareis and Grant, 2015), which, in turn, informs relevant educational decisions. The need for teachers to understand and use teacher-made tests to improve students' learning is increasingly becoming important in the field of education (Guskey, 2003; Guskey and Jung, 2013). Teachers must be proficient and competent in the area of assessment as they have traditionally been in the areas of curriculum and instruction (Gareis and Grant, 2015). Extant literature on test construction competencies and the quality of teacher-made tests showed that test construction competencies are related to the quality of test items (Marso and Pigge, 1989; Dosumu, 2002; Magno, 2003; Agu et al., 2013; Kinyua and Okunya, 2014). Thus, a teacher's competence in constructing test items is directly related to achieving good quality test instruments (Chau, as cited in Hamafyelto et al., 2015). When classroom teachers have limited test construction skills, the quality of the tests they construct is reduced. Tests that are poor in quality negatively affect the assessment validity (Amedahe and Asamoah-Gyimah, 2016). School teachers and administrators are not able to provide support and educational opportunities that meet each student's needs when the assessment tools constructed by teachers are low in quality (Agu et al., 2013). In other words, the lack of or low degree of validity of the test leads to undependable inferences about student learning (Gareis and Grant, 2015; Amedahe and Asamoah-Gyimah, 2016). Based on this, educational decisions such as the selection of students for educational opportunities would be wrongfully made.

The Ghanaian context

In Ghana, questionnaires as a self-report measure have been a common instrument that has been used to investigate test construction competencies or practices of classroom teachers (see Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Armah, 2018). Accordingly, Wiredu (2013) suggested investigating teachers' responses to questionnaire items by directly examining samples of tests developed by the teachers for construction flaws. This will help provide qualitative information concerning the quality of the teacher-made tests. Previous studies on teachers' test construction competencies did not involve the analysis of samples of teacher-made tests in understanding the relationship between teachers' responses to a questionnaire on test construction knowledge and test construction practices (Oduro-Okyireh, 2008; Anhwere, 2009). To go beyond just relying on the responses of teachers on self-report measures, Oduro-Okyireh (2008) recommended that research should be conducted to understand teachers' actual test construction competencies (the quality of teacher-made tests). One of the measurement theories that call for the need to use item analysis (quantitative and qualitative methods) to evaluate the quality of teacher-made tests is the classical true-score theory.

It is imperative to note that most norm-referenced achievement tests are commonly designed to differentiate examinees with regard to their competence in the measured areas (Nitko, 2001). That is, the test is designed to yield a broad range of scores, maximizing discrimination among all examinees taking the test. This is based on the crucial assumption that psychological differences exist and can be detected through a well-designed measurement process (Furr and Bacharach, 2014). The well-designed measurement process is a question of the quality of the test constructed to detect individual differences in a given psychological construct such as achievement in mathematics. Therefore, constructing a test of good quality largely depends on an individual's ability to quantify the differences among people (Furr and Bacharach, 2014). For example, in educational settings, the onus rests on the teacher's ability (competence) to construct a measuring instrument that would help detect students who have gained mastery in a given content area and those who have not. However, test-related factors (format and construction flaws) which are attributed to the test construction competence of the classroom teachers affect how well their tests can detect high achievers from low achievers in a given subject area. Item analysis procedures, based on the assumptions of the classical true-score theory, create an avenue to validate teachers' responses to any self-report measure used in assessing their test construction competence.

Quantitative item analysis is a numerical method for analyzing test items' difficulty and discrimination indices employing student-response alternatives or options (Kubiszyn and Borich, 2013). The indices from the quantitative item analysis communicate the presence of problem items and errors that minimize the tests' utility in separating high achievers from low achievers. Accordingly, performing qualitative item analysis helps reveal more specific problems that contribute to unacceptable difficulty and discrimination indices (Nitko, 2001).

Given that the educational system in Ghana is examination-oriented (Baidoo-Anu and Ennu Baidoo, 2022), teachers are expected to develop competencies in test construction to be able to construct sound and quality tests (tests that are useful for measuring differences in students' achievement in a given subject area). However, with the large class size in Ghanaian classrooms, teachers are mostly forced to rely on multiple-choice tests to assess their students (Kissi, 2020). Despite the predominant use of multiple-choice items in Ghanaian classrooms, attention has not been given to teachers' multiple-choice test construction competence and the quality of the multiple-choice tests they construct. Accordingly, the crux of our study is to explore senior high school teachers' perceived multiple-choice test construction competence and the quality of multiple-choice tests they construct. Based on the objective of our study, the following research questions were developed to guide the study.

1. What is the perceived multiple-choice test construction competence of teachers in senior high schools in Ghana?
2. What are the characteristics of the multiple-choice test items constructed by the teachers based on the following criteria: difficulty index and discrimination index?
3. What are the common types of error associated with teacher-made multiple-choice tests among senior high school teachers in Ghana?

Literature review

Classical true-score theory

The theory conceptualizes any observed score on a test as the composite of two hypothetical components–a true score and a random error component. Mathematically, this is expressed in the form "$X = T + E$", where X represents the observed test score; T is the individual's true score; E is the random error component (Crocker and Algina, 2008). Thus, the theory is a simple mathematical model that describes how measurement errors can influence observed scores (Allen and Yen, 2002). The theory states that for every observed score, there is a true score or true underlying ability that can be observed accurately if there were no measurement errors (Allen and Yen, 2002). The observed score refers to a value that is obtained from the measurement of some characteristic of an individual. A true score is a theoretical idea that refers to the average score taken over repeated independent testing with the same test or alternative forms. It is also the real or actual level of performance on the psychological attribute being measured by a test (Furr and Bacharach, 2014). Apart from the influence of true scores, probable factors that affect observed scores are described by the theory as errors of measurement (Furr and Bacharach, 2014). True scores and error scores are unobservable theoretical constructs while observed scores are observable in nature (Nitko, 2001).

Regarding the definition of observed scores in the theory, in a situation where there are no errors of measurement in observed scores, one can greatly and confidently depend on observed scores for relevant decisions. This is because, from repeated independent testing, all observed scores reflect the true ability of the candidate who is assessed. Also, supposing that a core mathematics achievement test is administered to a group of students who differ in ability and their observed scores are without measurement errors, the teacher would place his or her confidence in the assessment results because differences (variability) in the students' test scores accurately reflect the differences in their true levels of knowledge in mathematics.

Nevertheless, the existence of errors of measurement results in deviations of observed scores from the true score(s) (Bhattacherjee, 2012), and this minimizes one's confidence and dependability on the assessment results. How much confidence one can place in test results is a question of two main concepts pertaining to the quality of assessment procedures: (a) reliability and (b) validity. Therefore, the classical true-score theory provides an understanding of factors (measurement errors) that influence observed scores' reliability and validity. Examples of such factors include ambiguous items, poor instructions on a test, fatigue, and guesswork because of item difficulty (Crocker and Algina, 2008; Amedahe and Asamoah-Gyimah, 2016). To contribute to the reliability and validity of assessment results by ensuring test quality, the theory ties a good test to the test construction competence of the test constructor. Accordingly, the theory emphasizes some procedures and principles for test construction and evaluation to aid in the effective control and reduction of the impact of measurement errors related to a given test.

Principles, guidelines, or suggestions for constructing and improving the quality of multiple-choice tests

Errors associated with multiple-choice tests negatively affect the reliability and validity of the entire assessment results. To help improve the quality of the multiple-choice test, some principles, guidelines, or suggestions have been given by researchers, professionals, and experts in the educational assessment of students and psychological testing. In constructing multiple-choice tests, it is quintessential to follow the general principles of test construction and specific item format test construction principles. The outlined general test construction principles and specific principles for the construction of multiple-choice tests are organized as indicated by Nitko (2001), Joshua (2005), Kubiszyn and Borich (2013), and Etsey (as cited in Amedahe and Asamoah-Gyimah, 2016).

General principles for test construction

a. Begin writing items far enough in advance so as to have time to revise them.
b. Align the content of the test with instructional objectives.
c. Include items or questions with varying difficulty levels.
d. Match test items to the vocabulary level of the students.
e. Be sure that the item deals with an important aspect of the content area.
f. Write or prepare more items than are actually needed.
g. Be sure that the problem posed is clear and unambiguous.
h. Be sure that each item is independent of all other items. That is, the answer to one item should not be required as a condition for answering the next item. A hint to one answer should not be embedded in another item.
i. Be sure the item has one correct or best answer on which all experts would agree.
j. Prevent unintended clues to the answer in the statement or question. Grammatical inconsistencies such as "a" or "an" give clues to the correct answer to those students who are not well prepared.
k. Give specific instructions on the test. For example, instructions should be given as to how students are required to answer the questions.
l. Give the appropriate time limit for the completion of the test.
m. Appropriately assemble the test items. For example, use a font size that students can see and read, properly space the items, and arrange test items according to difficulty level (that is, from low to high); number the items one after the other without interruption, and appropriately assign page numbers.
n. Use an appropriate number of items to test students' achievement.
o. Review items for construction errors.
p. Evaluate the test items for clarity, practicality, efficiency, and fairness.

Specific principles for constructing multiple-choice test

a. Present the stem as a direct question.
b. Present a definite, explicit, and singular question or problem in the stem.
c. Eliminate excessive verbiage or irrelevant information from the stem.
d. Include in the stem any word(s) that might otherwise be repeated in each alternative.
e. Use negatively stated stems carefully (by underlining and/or capitalizing or bolding the negative word in the stem).
f. Make alternatives grammatically parallel with each other and consistent with the stem.
g. Make alternatives mutually exclusive or independent of each other.
h. Avoid the use of "none of the above" as an option when an item is of the best answer type.
i. Avoid the use of "all of the above" as part of the options to the stem of an item.
j. Make alternatives approximately equal in length.
k. Present alternatives in a logical order (for example, chronological, most to least, or alphabetical) when possible.
l. Keep all parts of an item (stem and its options) on the same page.

m. Arrange the alternatives in a vertical manner.
n. Use plausible distractors/options/alternatives.

Though the classical true-score theory has been described as a weak theory, its application to examine the quality of test items was of particular interest as it helps to understand the question, "Why is there a need for teachers to be competent in applying the principles of test construction?" It also endorses the use of quantitative methods of evaluating the quality of test items based on test scores and complements such evaluation with qualitative item analysis.

Assessment competence and assessment practice

The construction of tests for assessment is an aspect of classroom assessment practices that requires some level of assessment competence. Kissi, (2020) defined assessment competence as "an acquired, modifiable, and unobservable but demonstrable ability which is an integration of an individual's knowledge, skills, attitudes, and values in/on assessment" (p. 70). In light of this, assessment competence refers to an individual's ability to use or demonstrate the knowledge and skills acquired through assessment training in order to assess students' learning (Kissi, 2020). In contrast, assessment practice is the process of acquiring, analyzing, and interpreting data regarding student learning. It entails making crucial decisions on the student and the procedures involved in imparting knowledge to the learner (Nitko, 2001). Assessment competence in the view of Kissi, (2020):

> *answers the question: how well do classroom teachers employ their ability (which is an integration of their knowledge, skills, attitudes, and values in/on assessment) to successfully carry out those activities to match expected standards or to ensure improvement in their assessment activities? Since assessment competence in itself cannot directly be observed, such a construct can be inferred from what teachers do in terms of how well they go about their assessment practices (p. 71).*

Assessment competence and multiple-choice test construction competence

In the view of Gareis and Grant (2015), classroom teachers should be able to apply adequate knowledge and skills in assessment as they have usually been doing when it comes to activities involved in the transfer of knowledge to students. According to Nitko (2001), because the activities involved in the assessment are relevant to making relevant educational decisions, teachers have to be competent in choosing and using assessment tools. As stipulated in the standards for teacher competence in the educational assessment of students, for teachers to function effectively in assessment, they should be competent in assessment. According to the American Federation of Teachers (AFT), National Council of Measurement in Education (NCME), and National Education Association (NEA) requirements (as cited in Nitko, 2001), teachers should be capable of selecting, creating, administering, scoring, interpreting, and utilizing assessment data for pertinent educational decisions following legal and ethical norms (Kissi, 2020). From the foregoing, it can be seen that one of the standards for teachers' assessment competencies is their capacity to design and create tests. This standard indicates the test construction skills they need to have (Kissi, 2020).

One factor that directly affects the test quality, in Chau's view (as cited in Hamafyelto et al., 2015), is the proficiency of classroom teachers in constructing assessment tools. To gather information to improve teaching and learning, it is possible to identify students' areas of weakness and instructional issues given that the assessment tools used are of good quality (Nitko, 2001). According to McMillan (2000), understanding how general, fundamental assessment guidelines and ideas may be applied to improve student learning and teacher effectiveness is what is most important about assessment.

As one of the assessment competence criteria, test construction competence requires teachers to be adept at adhering to specific principles while creating assessment instruments or procedures that are suitable for instructional decisions. AFT, NCME, and NEA (as cited in Nitko, 2001) indicate that instructors who are proficient in this area will have the following conceptual and application skills in (a) planning the construction of assessment tools that help to inform decisions about students and instructional procedures; (b) selecting an appropriate technique which meets the intent of their instruction; (c) adhering to appropriate principles for developing and using assessment methods or techniques in their teaching, and avoiding common mistakes in student assessment; (d) using student data to examine the quality of each assessment technique used. In order to effectively assess pupils or students in accordance with the instructional objectives presented in class, teachers must select item format(s) that is or are suitable to the intent of their instruction. In Ghanaian senior high schools, the predominant item formats used in constructing end-of-term examination questions or tests are the essay and the multiple-choice item formats (Kissi, 2020). Hence, understanding teachers' multiple-choice test items' construction is needed within Ghana's educational system.

What is a multiple-choice item?

A multiple-choice item is an item that is made up of one or more introductory sentences followed by a list of two or more suggested responses (Nitko, 2001). The student is required to choose the correct answer from among the responses the teacher gives (Nitko, 2001). The part of the item that asks the question is called the stem. Instead of asking a question, it may set the task a student must perform or state the problem a student must solve. The list of suggested responses to the stem is called options. The options are also known as alternatives, responses, or choices (Morrow et al., 2000; Nitko, 2001). Usually, only one of the options is the correct or best answer to the question or problem the teacher poses. This is called the keyed answer, keyed alternative, or simply the key. The remaining incorrect options are called distractors or foils (Nitko, 2001; Joshua, 2005). To ensure that the assessment task neither prevents nor inhibits a student's ability to demonstrate attainment of the learning target, care should be taken to follow the guidelines for constructing multiple-choice tests. For instance, avoiding ambiguous and imprecise items, inappropriate and unfamiliar vocabulary, and poorly worded directions. After the first draft of the items, the items should be reviewed and edited. Moreover, the marking scheme should be prepared in conjunction with drafting the items (Etsey, as cited in Amedahe and Asamoah-Gyimah, 2016).

Quality of assessment procedures

The quality of assessment procedures is of great concern when it comes to the assessment of student learning. Ghanaian classroom teachers (trained and untrained) from the basic level to the university level construct, administer, and score classroom achievement tests regardless of whether they have had training in measurement and evaluation or not (Anhwere, 2009). When classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, the quality of the tests they construct is questionable. This is because, according to Chau (as cited in Hamafyelto et al., 2015), a teacher's test construction competence is directly related to ensuring the quality of a test. Poor test quality negatively affects the validity of assessment results (Amedahe and Asamoah-Gyimah, 2016). From the aforesaid, by implication, when teacher-made tests are low in quality, school administrators and teachers will not be able to make available support and educational opportunities that each student needs (Agu et al., 2013). In other words, a lack of or low degree of validity of test results leads to undependable inferences about student learning (Gareis and Grant, 2015; Amedahe and Asamoah-Gyimah, 2016) based on which educational decisions such as promotion and selection of students for educational opportunities would be wrongfully made. To avoid or minimize the negative effects of assessment procedures that are low in quality, the onus rests on classroom teachers to ensure the quality of the assessment procedures they employ. However, in Ghana, since classroom teachers hardly engage in quantitative item analysis as a way of assessing the utility of their multiple-choice test items, this study was relevant in (a) measuring their perceived competence in test construction; (b) evaluating their perceived competence in terms of the difficulty and discrimination indices of their multiple-choice test items; (c) employing qualitative item analysis to examine what test-related errors affected some of the observed indices.

Examining test construction competence through quantitative and qualitative item analysis

Test tryout, administration, and quantitative evaluation of the test

Quantitative evaluation (or item analysis) is a numerical method for analyzing test items employing student response alternatives or options (Kubiszyn and Borich, 2013). Before one would be able to conduct quantitative item analysis, the test should be administered to a sample with similar characteristics as the actual group who will be taking the final test (Shillingburg, 2016). This is called a test tryout. According to Cohen and Swerdlik (2010), for classroom teachers, test tryouts (pilot work) need not be part of the process of developing their tests for classroom use. However, the classroom teacher can engage in quantitative evaluation of test items after a test has been administered. The technique will enable them to assess the quality or utility of the items. It does so by identifying distractors or response options that are not doing what they are supposed to be doing. Quantitative evaluation of test items is ideally suited for examining the usefulness of multiple-choice formats (Kubiszyn and Borich, 2013).

Qualitative evaluation of the test

This method is used to review items on printed copies for test construction errors (Kubiszyn and Borich, 2013). Again, it is required for assessing the worth of the test before it is produced in large numbers to be administered (Amedahe and Asamoah-Gyimah, 2016). It is also done after a test is administered following quantitative item analysis and its purpose is to find out qualitative information about what led to unacceptable indices from quantitative item analysis. Hence, qualitative evaluation (or item analysis) is a non-numerical method for analyzing test items not employing student responses but considering content validity, clarity, practicality, efficiency, and fairness (Amedahe and Asamoah-Gyimah, 2016). Content validity, as one of the qualitative evaluation criteria, answers the questions: Are the items representative samples of the instructional objectives covered in a class? Does the test genuinely reflect the level of difficulty of the materials covered in a class? If the answer is "Yes," then content-related validity evidence is established (Amedahe and Asamoah-Gyimah, 2016). Clarity as another measure of evaluating the worth of the test refers to how the items are constructed and phrased while simultaneously judging them against the ability levels of the students. That is, the test material should be clear to students as to what is being measured and what they are required to do in attending to the questions (Nitko, 2001).

Practicality is concerned with the adequacy of the necessary materials and the appropriateness of time allocated for the completion of the test (Brown, 2004). The efficiency of a test seeks information as to whether the way the test is presented is the best to assess the desired knowledge, skill, or attitude of examinees in relation to instructional objectives (Amedahe and Asamoah-Gyimah, 2016). Conversely, fairness refers to the freedom of a test from any kind of bias. The test should be judged as appropriate for all qualified examinees irrespective of race, religion, gender, or age. The test should not disadvantage any examinee or group of examinees on any basis other than the examinee's lack of knowledge and skills the test is intended to measure (Nitko, 2001). Since the study focused on examining the characteristics of the teacher-made end-of-term multiple-choice tests, after the quantitative item analysis, the qualitative item analysis was used to identify possible test-related factors that affected the psychometric properties of the tests (in terms of difficulty and discrimination indices). Though qualitative evaluation is wide in scope, for the purpose of the study, it was operationalized as the deviations observed with respect to the principles of test construction using the multiple-choice test error analysis checklist (see Appendix A).

Methods

The study employed descriptive research design to understand senior high school teachers' multiple-choice test construction competence and the quality of multiple-choice test items they constructed. The study was done in two phases. The first phase was to obtain information on the multiple-choice test construction competence of the teachers. The second phase was to help validate the perceived multiple-choice test construction competence of the teachers through quantitative and qualitative item analysis.

Participants

Participants' selection was done in two phases.

Phase one

We selected 157 teachers from four senior high schools in the Kwahu-South District in the Eastern Region of Ghana. These participants were form one, form two, and form three teachers distributed across seven subject teaching areas (Financial Accounting, Cost Accounting, Business Management, Economics, English Language, Integrated Science, and Core Mathematics). The 157 participants responded to the self-designed questionnaire developed to assess teachers' multiple-choice test construction competence.

Phase two

The 157 participants who responded to the self-designed questionnaire were asked if they were willing to provide (a) copies of their latest end-of-semester self-constructed and administered multiple-choice test, (b) marking scheme, and (c) students' responses on the administered end-of-semester multiple-choice test items. Out of 157 participants, 47 teachers (across all the subject areas) provided these documents for further analysis. Accordingly, the 47 teachers were selected for the item analysis of the multiple-choice items they have constructed. Out of the 47 teachers, 68.09% had a first degree with education, 23.40% had a first degree without education, 4.26% of the participants had a master of philosophy, and 4.26% had completed a master of education programme. It is evident from the results that most of the participants were first-degree holders with a background in education. In the pursuit of a first degree, master of education, and master of philosophy, one is introduced to courses related to educational assessment of students' learning outcomes. From the cumulative percent, most of the participants (76.60%) possessed basic competence in the assessment of students (Kissi, 2020).

Instruments

The two instruments used for the data collection exercise were questionnaires and document examination. A 20-item self-designed instrument titled *Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC)* was used to assess teachers' multiple-choice test construction competence. The instrument was developed based on a comprehensive literature review on test construction competence. The instrument is made up of two sections namely "Section A" and "Section B." Section A is made up of items that help to obtain demographic information on teachers and Section B is made up of items that help to assess teachers' multiple-choice test construction competence. The scale of measurement that is used for the items under Section B is a 4-point Likert-type scale on a continuum of strongly disagree (SD), disagree (D), agree (A), and strongly agree (SA). The content validity of the research instrument was established by making sure that it objectively, fairly, and comprehensively covered the domain that it purports to cover. The instrument which was used for the study was initially made-up of 23 items. However, the items were reduced to 20 after experts' and teachers' judgment and pre-testing of the instrument among 130 teachers in a different district with similar characteristics as the study area. Concerning the experts' and teachers' judgment of the items on the questionnaire, items that were ambiguous and difficult to understand were rephrased so that the respondents could easily read and understand.

Document examination in this study covered students' responses on multiple-choice test items for end-of-semester administered teacher-made tests, copies of the marking schemes, and end-of-semester teacher-made tests. Using the marking schemes and students' responses on multiple-choice test items administered by the research participants, quantitative item analysis was performed to assess the characteristics of the multiple-choice test items for each of the classroom teachers. The assessment criteria used in assessing the characteristics of the items are based on the following item analysis descriptive statistics indices: (a) difficulty index (p-value), and (b) discrimination index (DI).

Based on the literature reviewed, the criteria suggested by Allen and Yen (1979) in terms of acceptable difficulty indices ranging from 0.30 to 0.70 and Kubiszyn and Borich's (2013) recommendation of at least a positive discrimination index for norm-reference tests, the following criteria were used in determining the characteristics of the teacher-made test items:

1. An item is judged as a good item if it is within the range of 0.30 to 0.70 and has a positive discrimination index.
2. An item is a problem item if it is within the range of 0.30 to 0.70 but has a zero discrimination index.
3. An item is a problem item if it is within the range of 0.30 to 0.70 but has a negative discrimination index.
4. An item is a problem item if it falls outside the range of 0.30 to 0.70 but has a positive discrimination index.
5. An item is a problem item if it falls outside the range of 0.30 to 0.70 and has a zero discrimination index.
6. An item is a problem item if it falls outside the range of 0.30 to 0.70 and has a negative discrimination index.

In addition, with regard to a qualitative evaluation of the teacher-made tests for format and construction flaws, the participants' end-of-semester administered Business Management and Core Mathematics multiple-choice tests were assessed for errors using the "Multiple-Choice Test Error Analysis Checklist" (see Appendix A).

Data analysis

Research question one sought to explore and describe the multiple-choice test construction competence of teachers in assessing students' learning outcomes at the senior high school level in the Kwahu-South District. The scoring of items based on the 4-point Likert scale of measurement was strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1. After the scoring, exploratory factor analysis was performed to determine the factor structure of the multiple-choice test construction competence of the teachers, and the factors were ranked based on their respective explained variance and mean. The standard deviations associated with each mean were also provided. Concerning the use of mean, a criterion score (CS) of 2.50 (that is, $[1+2+3+4]/4=2.50$) using the item's mean was established to determine the level of the respondents' agreement or disagreement towards their perceived test construction competence. An item mean score of 2.50 or above indicates teachers' positive attitudes, while a mean below 2.50 indicates teachers' negative attitudes which are embedded in each indicator of how well they employ their competence in constructing multiple-choice tests. After obtaining the difficulty and the discrimination indices for each set of items constructed by the teachers, means, standard deviations, and sum were used to analyze data collected on research question two. In addition, the mean of the problem items and the good items were compared using MedCalc's comparison of means calculator after meeting the assumptions that permit such analysis. Concerning research question three, "common format and construction flaws" is a categorical variable; therefore, frequency count was reported.

Results

Research question one

To answer research question one, there was a need to understand the structural patterns from teachers' responses to the TTCCQ-MC; thus, exploratory factor analysis was conducted using principal component analysis (PCA) with orthogonal rotation (varimax). Prior to starting the factor analysis, data were checked to ensure appropriateness for factor analysis. Exploratory factor analysis was conducted using 20 items that assess teachers' multiple-choice test construction competence. In testing the assumptions for PCA, the determinant of the correlation matrix as an indicator of multicollinearity was 0.015, which was substantially greater than the minimum recommended value of 0.00001. This meant that multi-collinearity was not a problem in conducting PCA. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = 0.70, and all KMO values for individual items were > 0.50, which was above the acceptable limit of 0.50 (Field, 2018). This meant that the sample size was adequate for PCA. Bartlett's test of sphericity was significant (χ^2 (190) = 619.939, $p < 0.001$). This indicated that correlations between items were good for PCA.

After satisfying the assumptions for PCA, an initial analysis was run to obtain the eigenvalue for each component in the data. Seven components had eigenvalues over Kaiser's criterion of one and in combination explained 60.65% of the total variance. The scree plot (see Figure 1) showed a point of inflexion that would justify retaining three components. Given the sample size of 157 and 20 items, the Kaiser's criterion on seven components, and convergence of the scree plot on three components, parallel analysis (PA) was conducted in addition to examine the appropriate number of components to maintain. Hayton et al. (2004) have pointed out that PA helps to identify the meaningful number of emerging factors from the set of items that are to be maintained. The PA also endorsed maintaining three factors. The results from PCA and reliability analysis endorsed 19-item TTCCQ-MC. That is, considering the absolute cut-off value of 0.40 for factor loadings, one of the items did not load on any of the factors since their loadings were below the cut-off value. Therefore, the final 19-item questionnaire with an overall reliability coefficient of 0.73 was considered valid for assessing teachers' multiple-choice test construction competence.

Based on the exploratory factor analysis, a three-factor structure emanated to help us understand teachers' multiple-choice construction competence. The first factor was termed test item assembling. Seven items were loaded into this factor, explaining 13.43% of the variance. These teachers prioritized ensuring proper spacing of test items for easy reading, keeping all parts of an item (stem and its options) on the same page, making sure options are approximately equal in length, and appropriately assigning page numbers to the test with clear specific instructions on the test. Factor two was named test content validity. Six items were loaded into this factor, explaining 12.50% of the variance. The items focused on teachers' priority in making sure that test items are matched to instructional objectives (intended outcomes of the appropriate difficulty level), preparing the marking scheme while constructing the items, and ensuring that each item deals with an important aspect of the content area and pose clear and unambiguous items. The third and final factor was named item "options" handling. Six items were loaded into this factor, explaining 11.77% of the variance. These teachers focused on ensuring that item options (i.e., alternatives) are approximately equal in length, options are presented in some logical order (e.g., chronological, most to least, or alphabetical) when possible, options are made independent of each other, and they also avoided the use of "none of the above" as an option when an item is of the best answer type. Exploratory factor analysis of teachers' multiple-choice test construction competence is presented in Table 1.

Rankings based on the percentage of explained variance indicated that the teachers perceived much more competence in assembling test items (13.43%) followed by competence in ensuring content validity (12.50%). Items options handling was the least perceived competence (11.77%) by teachers in our study. The results as presented in Table 2 confirm the preceding observations.

As seen in Table 2, comparing the mean of means for each component [competence in test item assembling (MM = 3.37, SD = 0.37), competence in ensuring content validity (MM = 3.24, SD = 0.41), and competence in handling items' options (MM = 2.80, SD = 0.48)] to the criterion score of 2.50, it can be said that, generally, for each component, most of the teachers perceived their competence as high. However, based on the rankings, it can be said that most of the research participants found it very easy to exhibit competence in assembling test items (MM = 3.37, SD = 0.37, R = 1st), easy to demonstrate competence in achieving content validity (MM = 3.24, SD = 0.41, R = 2nd), and quite difficult to demonstrate competence in handling the items' alternative (MM = 2.80, SD = 0.48, R = 3rd).

Research question two

The result of the characteristics of the multiple-choice items developed by the research participants is presented in Table 3.

As shown in Table 3, based on quantitative items analysis statistics, items that met both acceptable criteria for the discrimination index and difficulty index were judged as good items. Items that did not meet the set criteria were judged as problem items. The result showed that out of the total number of 2,325 items, 2,306 were deemed valid for item analysis (that is, multiple-choice items with four options). This means that 19 items were excluded from the items analysis. With respect to the set criteria for assessing the characteristics of the items, out of the total of 2,306 items, 1,199 items were described as good items, and 1,107 items were identified as problem items. This means that most of the test items constructed by the teachers are described as good items per their respective difficulty and discrimination indices. However, the 1,107 items identified as problem items might

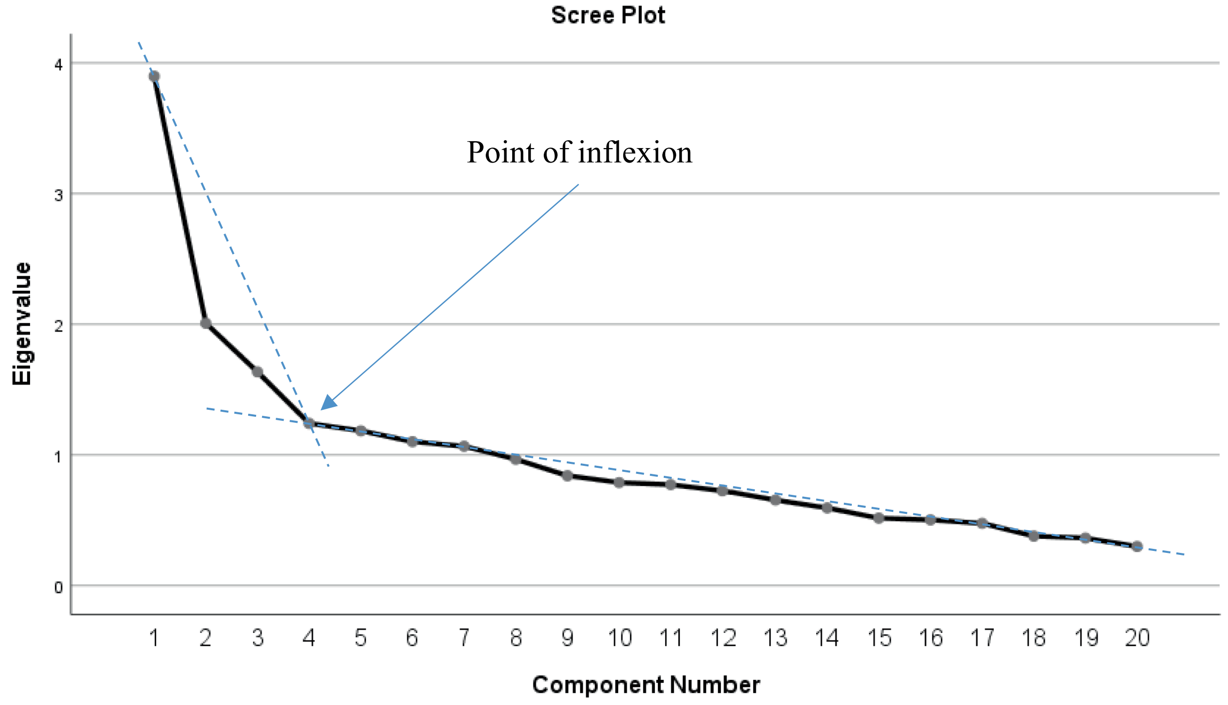

FIGURE 1
Scree plot for factor extraction.

have posed serious consequences for students who responded to these items.

Further analysis using "MedCalc's Comparison of means calculator" suggests that the average value for the number of good items produced by the classroom teachers (M = 25.51, SD = 8.51) was not statistically greater than the average value for the number of problems items produced (M = 23.55, SD = 8.98), t (92) = 0.03, p = 0.28, 2-tailed. Accordingly, it can be said that with respect to test characteristics, in general, the test items for assessing students' achievement lacked a suitable level of psychometric properties. This is attributable to the fact that the total number of good items produced by the teachers was not statistically different from the total number of problem items. Table 4 presents the result on problem items based on unacceptable difficulty indices that are less than 0.30, difficulty indices that are greater than 0.70, and discrimination indices that are less than or equal to 0.00.

Table 4 shows that out of the total number of 2,306 valid items for item analysis, 664 had difficulty indices less than 0.30 (difficult items) and 295 had difficulty indices greater than 0.70 (easy items). This means that most of the items were difficult. Further, in sum, the unacceptable number of items according to Allen and Yen's (1979) item evaluation criteria for item difficulty is 959 (that is, 664 + 295). On the other hand, out of the 2,306 valid items, 395 items had unacceptable discrimination indices less than or equal to zero based on Kubiszyn and Borich's (2013) recommendation that one can seriously consider any item with a positive discrimination index for the norm-referenced test(s). This means that most of the items had unacceptable difficulty indices as compared to the discrimination indices.

Research question three

The literature review on the use of quantitative item analysis in assessing items' characteristics revealed that the presence of problem items calls for qualitative evaluation of the multiple-choice test. Thus, research question three was formulated to help identify the multiple-choice format and item construction errors associated with teacher-made multiple-choice tests. In addressing this research question, the participants' end-of-semester administered Business Management (BM) and Core Mathematics (CM) multiple-choice tests were assessed for errors using the "Multiple-Choice Test Error Analysis Checklist" (see Appendix A). In all, 12 achievement tests (BM, 4; CM, 8) were qualitatively examined. The results are presented in Table 5.

As can be seen from Table 5, with specific reference to format errors, 11 out of 12 tests (BM, 3 out of 4; CM, 8 out of 8) were identified to have a detectable pattern of correct answers. Also, 6 out of 12 tests had items with a font size that some of the students could find more difficult to see and read (BM, 0 out of 4; CM, 6 out of 8). Therefore, it could be said that most of the tests were identified with the problem of a detectable pattern of correct answers as compared to the use of font size that students could find difficult to see and read.

To examine construction flaws associated with the tests, problem items were qualitatively examined. From Table 5, each of the following errors was observed with the problem items across 9 out of the 12 tests: (a) clues to the correct answer, (b) instruction-related issues (no and/or incomplete instruction), and (c) time for completion of items not indicated on the test. These observed errors are followed by other errors such as the use of implausible distractors (that is, 8 out of 12 tests) and ambiguous items/more

TABLE 1 Exploratory factor analysis of the teachers' multiple-choice test construction competence.

Description		Factors		
		1	2	3
Percentage of variance explained (after rotation)		13.432	12.502	11.766
Initial eigenvalue		3.898	2.007	1.635
Parallel Analysis (Random eigenvalues)		1.687	1.556	1.463
Q/N	When constructing multiple-choice tests, I:	1 Test Items Assembling	2 Content Validity	3 Item Options Handling
1	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)		0.699	
2	make sure each item deals with an important aspect of the content area		0.757	
3	prepare the marking scheme while constructing the items		0.528	
4	pose clear and unambiguous items		0.454	
5	give specific instructions on the test	0.481		
6	include in the stem any word(s) that might otherwise be repeated in each option			0.569
7	make the options grammatically consistent with the stem			0.544
8	make options independent of each other			0.604
9	avoid the use of "none of the above" as an option when an item is of the best answer type			0.529
10	make options approximately equal in length			0.717
11	present options in some logical order (e.g., chronological, most to least, or alphabetical) when possible			0.535
12	include questions of varying difficulty		0.492	
13	match items to the vocabulary level of the students	-	-	-
14	give appropriate time for completion of the test		0.464	
15	use the appropriate number of test items	0.632		
16	number the test items one after the other	0.584		
17	appropriately assign page numbers to the test	0.475		
18	properly space the test items for easy reading	0.683		
19	keep all parts of an item (stem and its options) on the same page	0.502		
20	review test items for construction errors	0.571		

Extraction Method: Principal Component Analysis.

than one correct answer (that is, 7 out of 12 tests). On the contrary, 1 out of the 12 tests was identified with clueing and linking items (that is, BM, 1 out of 4; CM, 0 out of 8). Thus, the result suggests that most of the tests examined with reference to construction errors associated with problem items had the following issues: (a) clues to the correct answer, (b) instruction-related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to clueing and linking items.

Discussion

The purpose of this study was to explore senior high school teachers perceived multiple-choice test construction competence and the quality of their multiple-choice tests. To understand the teachers' multiple-choice test construction competence in the Kwahu-South District, a three-factor structure emanated from the factor analysis - content validity, item "options" handing, and test item assembling. Findings showed that, generally, most of the teachers judged themselves as competent in constructing multiple-choice tests. In other words, they perceived themselves as possessing competence in achieving content validity, handling the options to the item's stems, and assembling the test. This observation could be related to the fact that most of the respondents had a background in education.

In the educational assessment of student learning outcomes, ensuring content validity and appropriately handling options of the item stems are more relevant competence areas as compared to competence in assembling test items. However, the teachers perceived more competence in test item assembling than ensuring content validity of the test and using appropriate "options or alternatives" of the test items. For example, the teachers perceived more competence in the proper spacing of test items for easy reading, keeping all parts of an item (stem and its options) on the same page rather than ensuring that test items are matched to instructional objectives

TABLE 2 Ranks of the teachers' multiple-choice test construction competence.

Component	Number of items	Mean of means (MM)	Std. deviation of MM	Ranks (R)
Competence in test item assembling	7	3.37	0.37	1st
Competence in Ensuring content validity	6	3.24	0.41	2nd
Competence in handling Items' options	6	2.80	0.48	3rd

TABLE 3 Characteristics of the multiple-choice items developed by the teachers.

Description	Items constructed by the teachers	Valid items for item analysis	Problem items	Good items
Sum (total)	2325.00	2306.00	1107.00	1199.00
Mean	49.47	49.06	23.55	25.51
Std. Deviation	14.15	14.14	8.98	8.51

TABLE 4 Summary of Items based on unacceptable difficulty and discrimination indices.

Description	Sum	Mean	Std. Deviation
Difficulty indices less than 0.30	664	14.13	6.98
Difficulty Indices greater than 0.70	295	6.28	4.56
Discrimination indices less than or equal to 0.00	395	8.40	5.74

(intended outcomes of the appropriate difficulty level), making the options independent of each other, and crafting options that are grammatically consistent with the stem to avoid clues to the correct answer.

Findings from this study revealed that the teachers perceived themselves as competent in constructing multiple-choice tests. However, findings from the quantitative evaluation of the items revealed that there were serious problems with copies of the multiple-choice tests the teachers constructed for assessing their students. Thus, most of the teachers in this study perceived themselves as competent multiple-choice test constructors; however, an analysis of the sample of their actual test items showed otherwise. This study confirms the recommendation by Ary et al. (2010) that direct observation of the behavior of a random sample of respondents is a brilliant strategy to validate their responses to self-report measures. The problems observed through the direct analysis of items were unacceptable difficulty and discrimination indices. In relation to item difficulty, from Nitko's (2001) point of view, teachers should ensure that the test they construct contains items that are not too difficult or too easy for their students. However, many of the items were described as problem items with respect to the high and low difficulty indices. Consequently, the quality of the assessment results used in grading the students was questionable. The findings from the quantitative item analysis support prior work that found that teachers often have inadequate prerequisite skills to construct quality multiple-choice items that effectively assess the learning achievements of students (Rivera, 2011; Agu et al., 2013; Kinyua and Okunya, 2014; Hamafyelto et al., 2015; Tshabalala et al., 2015). To address the issue, Nitko (2001) calls on classroom teachers to develop competence in tailoring test items to each of the student's ability levels. This is necessary as the reliability of an assessment is affected when test difficulty is not matched to the ability of the students involved (Amedahe and Asamoah-Gyimah, 2016).

Concerning the discrimination indices, Furr and Bacharach (2014) have stated that it is the responsibility of the classroom teacher to construct test items that effectively discriminate those who have mastered a given content area from those who have not. Where deficiencies exist, in norm-referencing, these items should not be considered in terms of the total number of items that make up students' composite scores in a given achievement test (Nitko, 2001; Crocker and Algina, 2008; Kubiszyn and Borich, 2013). However, these items were considered in arriving at the composite scores based on which grades were assigned.

According to Hambleton and Jones (1993), classical true-score theory item analysis procedures have the potential to provide invaluable information concerning construction flaws such as implausible distractors and double negatives. Therefore, informed by this assertion, research question three was established to identify the associated multiple-choice format and item construction errors that contributed toward the poor difficulty and discrimination indices through qualitative evaluation of the tests. The qualitative analysis of teachers-constructed multiple-choice items revealed fundamental flaws in the items' write-up and format errors which might have explained the problem items as identified with the teacher-made multiple-choice tests.

Generally, findings in relation to research question three reveal that most of the tests were identified with the problem of a detectable pattern of correct answers as compared to the use of font size that students could find difficult to see and read. Moreover, most of the tests examined with reference to construction errors were associated with problem items that had the following issues: (a) clues to the correct answer, (b) instruction-related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to clueing and linking items. This supports Rivera's finding that classroom teachers do not possess adequate skills in constructing test items (Rivera, 2011).

Researchers have stated that the presence of format and construction errors reduces the quality of assessment results (Morrow et al., 2000; Nitko, 2001; Joshua, 2005; Kubiszyn and Borich, 2013; Amedahe and Asamoah-Gyimah, 2016). Therefore, problem items identified with the tests can pose serious consequences for students who responded to these items because these examinations in Ghana are high-stake (Amoako, 2019; Baidoo-Anu et al., 2022; Baidoo-Anu, 2022, Baidoo-Anu and Ennu Baidoo, 2022). Test results are used to make high stake decisions about students, especially determining their progress in the educational system. The findings from this study are

TABLE 5 Format and construction errors identified with the business management and core mathematics tests.

Type of errors	BM	CM	Total
Test format errors	Freq.	Freq.	Freq.
Alternatives not presented in some logical order	4/4	3/8	7/12
A detectable pattern of correct answers	3/4	8/8	11/12
The horizontal arrangement of options	3/4	4/8	7/12
Options of items appearing in different columns/pages	3/4	5/8	8/12
Page numbers not assigned	4/4	6/8	10/12
Poor arrangement of items/spacing of test items	2/4	4/8	6/12
Use of font size that is difficult to see and read	0/4	6/8	6/12
Item construction errors	Freq.	Freq.	Freq.
Ambiguous items/more than one correct answer	2/4	5/8	7/12
The central theme, task, or problem is not presented in the stem	3/4	0/8	3/12
Clues to the correct answer	4/4	5/8	9/12
Heterogeneous options	2/4	0/8	2/12
Grammatical, punctuation, and spelling errors	4/4	0/8	4/12
Implausible distractors	2/4	6/8	8/12
Instructional-related issues (no/ incomplete instruction)	4/4	5/8	9/12
Clueing and linking items	3/4	0/8	1/12
No answer	1/4`	3/8	4/12
Wrong key to the item	3/4	4/8	5/12
Not emphasizing (e.g., bolding, underlining or capitalizing) negative word in the stem	3/4	0/8	3/12
Time for completion of items not indicated on the test	4/4	5/8	9/12
Wrong answer	1/4	4/8	5/12

Key: Freq. = Frequency; / = out of.

consistent with several studies that found flaws in the multiple-choice items of teachers (Downing, 2004; Tarrant and Ware, 2008; DiBattista and Kurzawa, 2011; Wiredu, 2013). For example, Downing, 2004 argued that too high or low item difficulty disadvantaged some students. While Downing (2004, 2005) was not explicit on the type of students who are affected, Tarrant and Ware (2008) were more explicit and argued that flaws in high-stakes multiple-choice questions did not only disadvantage borderline students, rather high-achieving students were more likely than borderline students to be penalized by flawed items.

In sum, the direct assessment procedure helped to validate teachers' responses to the self-report measure used in the assessment of their competence in constructing multiple-choice. Both quantitative and qualitative item analyses were employed to validate the self-reported competence of the teachers. These methods revealed that though the teachers reported high levels of competence in constructing multiple-choice tests, the validation of their perceived competence using quantitative item analysis revealed that generally across all the seven subject areas, the number of problem items raise a concern about what they perceived about themselves and what their competence produced. Burton et al. (1991) have indicated that good multiple-choice test items are more demanding and take a lot of time to craft as compared to other types of test items. Given that multiple-choice test construction has different stages with each stage playing a significant role in test quality, teachers' less competence in any of the stages has the potential to mar the quality of tests (Agu et al., 2013). Thus, there is a need to ensure classroom teachers are practically exposed to item writing skills, especially ensuring content validity and crafting options to a multiple-choice item stem with good quality. According to Rivera (2011), classroom teachers can master the writing of test items through practice. Maba (2017) has also indicated that competence as an ability is modifiable and new experiences can be integrated. For instance, faculty members' (teachers) competence in developing multiple-choice test items with acceptable difficulty and discrimination indices improved significantly through training in constructing multiple-choice tests (Abdulghani et al., 2015). Consequently, new experiences gained by teachers as a result of exposure to constant training and practice in ensuring the quality of multiple-choice tests can lead to the integration and modification of their multiple-choice test construction competence.

Implication for policy and practice

Findings from this study provide unique and compelling evidence in the Ghanaian context regarding teachers' perceived test construction competence and analysis of teachers-constructed multiple-choice tests. Examinations results in Ghana are used to make high stake decisions regarding schools, teachers, and students (Baidoo-Anu and Ennu Baidoo, 2022). For instance, exam results determine students' progress from one grade to the other. Failure to pass these exams has dire consequences sometimes, including being retained in their present grade until they have passed the exams. This delays their progress and costs the family an extra year or more of associated schooling costs. According to the Ghana Ministry of Education (2018), more than 12% of senior high school students are retained in each grade level. Unfortunately, multiple-choice test items are the predominant type of items that are used during almost every examination in Ghana largely due to large class sizes. Thus, poor multiple-choice test constructions do not only affect students but also their families and the country's quality of education. This is because teachers' decisions made from these low-quality multiple-choice items may lack valid evidence and may not represent the actual achievements of students. This implies that educational stakeholders will not be able to adequately provide support and educational opportunities that meet each student's needs. Therefore, Ghana Education Service should priorities providing in-service professional development training opportunities for teachers to develop the prerequisite skills needed to construct quality multiple-choice items. Professional development training of this nature is not a one-day workshop but demands ongoing long-term support and resources for teachers. Moreover, evidence (course outline) showed that teacher education programs have test construction as part of the topics in educational assessment courses; however, teaching this course is more theoretical and does not provide the opportunity to practically engage pre-service teachers. Hence, we recommend that teacher education programs in Ghana could also incorporate practical lessons or training in their curriculum to help pre-service teachers develop competence in test construction with specific emphasis on achieving content validity and effective handling of multiple-choice item stem options.

We want to highlight that findings from this study were shared with teachers and district education directors, especially those who participated in the study. The common problems identified including recommendations were also shared with them.

Limitations and suggestions for future research

The study employed 157 teachers to respond to the question and 47 teachers for the sample multiple-choice test analysis. Moreover, the sample multiple-choice test analysis was carried out on mathematics and business management test. Performing qualitative evaluation in other subject areas could have revealed more specific problems in all subject areas that contributed to unacceptable difficulty and discrimination indices. However, such general evaluation was not feasible in terms of easy access to subject area experts in English, Financial Accounting, Economics, Cost Accounting, and Integrated Science to help in qualitatively examining tests for construction flaws such as ambiguities, more than one answer, and clues to correct answers. Consequently, the conclusions based on the relatively small sample of teachers do not present a holistic view of the test construction competence of the entire population of teachers considered for the study. Given the significant nature of this study, future research could expand the scope and sample to allow the generalization of the findings across the country.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by University of Cape Coast-Ghana. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PK conceived of the presented idea. PK, DB-A, EA, and RA-B contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The work was supported by the School of Graduate Studies and Graduate Students' Association of Ghana, University of Cape Coast, through the Samuel and Emelia Brew-Butler -SGS/GRASAG, UCC Research Grant.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., et al. (2015). Faculty development programs improve the quality of multiple choice questions items' writing. *Sci. Rep.* 5:9556. doi: 10.1038/srep09556

Agu, N. N., Onyekuba, C., and Anyichie, C. A. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: need for a test construction skill inventory. *Educ. Res. Rev.* 8, 431–439. doi: 10.5897/ERR12.219

Allen, M. J., and Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press Inc.

Allen, M. J., and Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Amedahe, F. K., and Asamoah-Gyimah, K. (2016). *Introduction to measurement and evaluation* (*7th ed.*). Cape Coast: Hampton Press.

Amoako, I. (2019). What's at stake in high stakes testing in Ghana: Implication for curriculum implementation in basic schools. *Int. J. Soc. Sci. Educ. Stud.* 5:72. doi: 10.23918/ijsses.v5i3p72

Anhwere, Y. M. (2009). Assessment practices of teacher training college tutors in Ghana. Available at: https://ir.ucc.edu.gh/xmlui/bitstream/handle/123456789/1690/ANHWERE%202009.pdf?sequence=1&isAllowed=y (Accessed May 23, 2020).

Armah, C. (2018). Test construction and administration practices among lecturers and staff of examinations unit of the university of Cape Coast in Ghana. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/3868 (Accessed November 10, 2022).

Ary, D., Jacobs, C. L., Sorensen, C., and Razavieh, A. (2010). *Introduction to research methods in education* (*8th ed.*). Belmont, CA: Wadsworth.

Baidoo-Anu, D. (2022). Between-school streaming: unpacking the experiences of secondary school teachers and students in category C schools in Ghana. *Int. J. Educ. Res. Open* 3, 100188–100189. doi: 10.1016/j.ijedro.2022.100188

Baidoo-Anu, D., and Ennu Baidoo, I. (2022). Performance-based accountability: exploring Ghanaian teachers perception of the influence of large-scale testing on teaching and learning. *Educ. Inq.* 1-18, 1–18. doi: 10.1080/20004508.2022.2110673

Baidoo-Anu, D., Gyamerah, K., and Chanimbe, T. (2022). Secondary school categorization in Ghana: silent plights of students and implications for equitable learning. *J. Hum. Behav. Soc. Environ.* 33, 348–365. doi: 10.1080/10911359.2022.2061665

Bhattacherjee, A. (2012). Social science research: principles, methods, and practices (2nd ed.). Available at: http://scholarcommons.usf.edu/oa-tesxtbook/3 (Accessed May 5, 2020).

Black, P. J., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* 92, 81–90. doi: 10.1177/003172171009200119

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., and Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Available at: https://testing.byu.edu/handbooks/betteritems.pdf

Cohen, R. J., and Swerdlik, M. J. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (*7th ed.*). New York: McGaw-Hill.

Crocker, L., and Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *CJSoTL* 2:4. doi: 10.5206/cjsotl-rcacea.2011.2.4

Dosumu, C. T. (2002). *Issues in teacher-made tests*. Ibadan: Olatunji and Sons Publishers.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Med. Educ.* 38, 1006–1012. doi: 10.1111/j.1365-2929.2004.01932.x

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv. Health Sci. Educ.* 10, 133–143. doi: 10.1007/s10459-004-4019-5

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (*5th ed.*). London: Sage.

Furr, R. M., and Bacharach, V. R. (2014). *Psychometrics: An introduction* (*2nd ed.*). London: Sage.

Gareis, R. C., and Grant, W. L. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (*2nd ed.*). New York: Routledge.

Ghana Ministry of Education (2018). The education strategic plan 2018–2030. Available at: https://www.globalpartnership.org/sites/default/files/2019-05-education-strategic-plan-2018-2030.pdf (Accessed May 20, 2022).

Guskey, T. R. (2003). How classroom assessments improve learning. *Educ. Leadersh.* 60, 6–11.

Guskey, T. R., and Jung, L. A. (2013). *Answer to essential questions about standards, assessments, grading, & reporting*. Thousand Oaks, CA: Corwin.

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

Hambleton, R. K., and Jones, R. W. (1993). An NCME instructional module on: comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.* 12, 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory analysis: a tutorial on parallel analysis. *Organ. Res. Methods* 7, 191–205. doi: 10.1177/1094428104263675

Joshua, M. T. (2005) *Fundamentals of test and measurement in education*. Calabar: University of Calabar Press.

Kinyua, K., and Okunya, L. O. (2014). Validity and reliability of teacher-made tests: case study of year 11 physics in Nyahururu District of Kenya. *Afr. Educ. Res. J.* 2, 61–71.

Kissi, P. (2020). *Multiple-choice construction competencies and items' quality: Evidence from selected senior high school subject teachers in Kwahu-South District*. Available at: http://hdl.handle.net/123456789/4641 (Accessed January 5, 2023).

Kubiszyn, T., and Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice* (*10th ed.*). Hoboken, NJ: Wiley.

Maba, W. (2017). Teachers' sustainable professional development through classroom action research. *Int. J. Res. Soc. Sci.* 7, 718–732.

Magno, C. (2003). The profile of teacher-made test construction of the professors of University of Perpetual Help Laguna. *UPHL Institut. J.* 1, 48–55.

Marso, R. N., and Pigge, F. L. (1989). "The status of classroom teachers' test construction proficiencies: assessments by teachers, principals, and supervisors validated by analyses of actual teacher-made tests." in *Paper presented at the annual meeting of the National Council on measurement in education, San Francisco, California*. (ERIC Document Reproduction Service No. ED306283).

McMillan, J. H. (2000). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin publishing company.

Morrow, J. R., Jr., Jackson, A. W., Disch, J. G., and Mood, D. P. (2000). *Measurement and evaluation in human performance* (*2nd ed.*). Champaign, IL: Human Kinetics.

Nitko, A. J. (2001). *Educational assessment of students* (*3rd ed.*). Upper Saddle River, New Jersey: Prentice-Hall.

Oduro-Okyireh, G. (2008). Testing practices of senior secondary school teachers in the Ashanti region. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/1324 (Accessed April 3, 2020).

Rivera, J. E. (2011). Test item construction and validation: developing a statewide assessment for agricultural science education. *Career Tech. Educ. Res.* 36, 69–80. doi: 10.5328/cter36.2.69

Shillingburg, W. (2016). Understanding validity and reliability in classroom, school-wide, or district-wide assessments to be used in teacher/principal evaluations. Available at: https://cms.azed.gov/home/GetDocumentFile?id=57f6d9b3aadebf0a04b2691a (Accessed April 8, 2020).

Tarrant, M., and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med. Educ.* 42, 198–206. doi: 10.1111/j.1365-2923.2007.02957.x

Tshabalala, T., Mapolisa, T., Gazimbe, P., and Ncube, A. C. (2015). Establishing the effectiveness of teacher-made tests in Nkayi district primary schools. *Nova J. Humanit. Soc. Sci.* 4, 1–6.

Wiredu, S. G. (2013). Assessment practices of tutors in the nurses' training colleges in the Western and Central regions of Ghana. Available at: https://ir.ucc.edu.gh/xmlui/bitstream/handle/123456789/2685/WIREDU%202013.pdf?sequence=1&isAllowed=y (Accessed May 6, 2020).

Appendix A

Multiple-Choice Test Error Analysis Checklist Instruction: Record once if each error has occurred several times or once for each test.

Q/N	Errors in constructing multiple - choice test	Number of occurrences across tests	Total
	Test format errors		
1.	Alternatives not presented in some logical order		
2.	A detectable pattern of correct answers		
3.	The horizontal arrangement of options		
4.	Options of items appearing in different columns/pages		
5.	Page numbers not assigned		
6.	Poor arrangement of items/spacing of test items		
7.	Use of font size difficult to see and read		
	Item construction errors		
8.	Ambiguous items/More than one correct answer		
9.	The central theme, task, or problem is not presented in the stem		
10.	Clues to the correct answer		
11.	Clueing and linking items		
12.	Grammatical, punctuation, and spelling errors		
13.	Heterogeneous options		
14.	Implausible distractors		
15.	Instructional-related issues (no/incomplete instruction)		
16.	No answer		
17.	Not emphasizing (e.g., bolding, underlining or capitalizing) negative word in the stem		
18.	Time for completion of items not indicated on the test		
19.	Use of "all of the above"		
20.	Wrong answer		
21.	Wrong key to the item		
22.	Wrong usage of "none of the above"		