EDUTREND
Journal of Emerging Issues and Trends in Education

Volume-3 | Issue-2| May-2026 | https://doi.org/10.59110/edutrend.881

Research Article

Document-Based Assessment of Item-Writing Skills: Implications for Policy Formulation in Teacher Education

Abraham Gyamfi[1*], Abraham Yeboah[1], Eric Atta Quainoo[2], Rosemary Acquaye[3]
[1]Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development , Kumasi, Ghana
[2]Wesley College of Education, Kumasi, Ghana
[3]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
*agyamfi@aamusted.edu.gh

ABSTRACT
Multiple-choice (MC) items are widely used in teacher education because they are efficient, objective, and suitable for assessing large groups of students. However, poorly constructed MC items may undermine the validity and reliability of assessment results. This study examined the extent to which College of Education teachers adhere to established principles of MC item construction and evaluated the quality of the items they develop and administer. A quantitative document analysis design was employed. Using simple random and stratified sampling procedures, 120 assessment instruments were selected from three Colleges of Education affiliated with Kwame Nkrumah University of Science and Technology. Only Section A of each instrument, consisting of 20 MC items, was analyzed, yielding 2,400 items. Four assessment experts evaluated the items against 21 established principles of MC item construction, and the data were analyzed using frequencies and percentages. The findings showed that only two of the 21 principles, representing 9.52%, reached the excellent adherence category. In addition, only 3.58% of the 2,400 items met all criteria for high-quality MC items, indicating that most items contained construction flaws. These findings suggest that weaknesses in MC item construction may threaten the validity, reliability, and fairness of assessment outcomes. The study contributes empirical evidence on actual item-writing practices among College of Education teachers and highlights the need for strengthened item moderation, assessment literacy training, and continuous professional development policies in teacher education institutions.

Keywords: *Assessment quality; College of Education teachers; Item-writing principles; Multiple-choice item construction; Teacher education policy*

ARTICLE HISTORY
Received: 25.01.2026
Accepted: 26.05.2026
Published: 30.05.2026

1. Introduction

Among objective item formats, the multiple-choice (MC) item is the most widely used in educational assessment. Nearly all forms of testing, including teacher-made tests and standardized assessments, employ MC items. Many people even mistakenly equate multiple-choice questions with objective tests. MC items, which generally consist of a stem and a set of response options, are popular among educators and test developers for several reasons (Scully, 2017; Kissi, Baidoo-Anu, Anane, & Annan-Brew, 2023). They can be administered quickly and scored by machine, making them suitable for large groups of students or test candidates (Magno, 2013; Wiredu, 2013). They are also applicable to examinees at different educational levels, unlike constructed-response items such as essay or short-answer questions. Moreover, previous studies have shown

that MC items have strong concurrent validity with other achievement measures (Oduro-Kyireh, 2008; Scully, 2017; Gyamfi, 2023), are associated with greater objectivity and reliability (Armah, 2018; Hamafyelto, Hamman-Tukur, & Hamafyelto, 2015), and allow broader sampling of content within a limited testing time (Nitko, 2012; Zhang, Z., & Burry-Stock, 2003). Incorporating MC items into assessments also enables test developers to evaluate item performance and use the results to improve subsequent assessments.

Despite these benefits, multiple-choice items have drawn considerable criticism. In the context of medical education, Awofala and Babatunde (2013), for instance, argued that professional competence requires the ability "to perform in a real-life setting that does not offer shortlists." Since the correct answer is provided among the response options, MC items are often criticized for being less capable of assessing cognitive processes beyond recall or recognition of knowledge (Brennan, 2006; Cohen & Wollack, 2006; Crocker & Algina, 2008). Nevertheless, MC items remain one of the most favored and widely used assessment formats. To fully realize their benefits, educators must be able to construct high-quality MC items that are valid, reliable, and aligned with intended learning outcomes (Erzoah, Gyamfi, Yeboah, & Langee, 2022).

However, constructing effective MC items remains a challenge for many teachers. Although MC items appear simple, their quality depends on teachers' ability to formulate clear stems, develop plausible distractors, avoid grammatical and structural clues, ensure alignment with learning outcomes, and maintain fairness in assessment. This issue is particularly important in colleges of education because teachers in these institutions serve as "trainers of trainers." They train prospective teachers who will later be responsible for classroom assessment in basic and secondary schools. Therefore, college of education teachers are expected to demonstrate strong competence in constructing MC items so that they can model and transfer appropriate assessment practices to teacher trainees (Baidoo-Anu, Asamoah, Quainoo, Gyamerah, Amoateng, & Sasu, 2023; Ministry of Education, Ghana, 2017).

Previous studies have contributed to understanding teachers' test construction competence, but important gaps remain. Much of the existing research has focused on basic and secondary school teachers, while limited attention has been given to college of education teachers as "trainers of trainers." In addition, many previous studies have relied on self-reported or survey-based measures of competence. Such approaches may not fully reflect teachers' actual item-writing practices because perceived competence does not necessarily translate into the quality of test items constructed and administered. Therefore, limited document-based evidence exists on the actual competence of college of education teachers in constructing multiple-choice items.

To address this gap, this study sought to examine the competence of college of education teachers in constructing multiple-choice items through a document-based assessment of actual test instruments. Specifically, it quantified teachers' adherence to established MC item-writing principles and evaluated the quality of the items they constructed and administered. By providing empirical evidence on actual item-writing practices, this study contributes to assessment quality assurance and offers practical implications for item moderation, continuous professional development, and teacher education policy.

2. Literature Review

2.1 Test Quality and Validity

Validity refers to the appropriateness of the interpretation and use of assessment outcomes, rather than merely the accuracy of test scores (American Educational

Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Gyamfi, 2022). It ensures that the correct information is gathered to make informed decisions. Judgments made about pupils based on assessment data must be accurate and appropriate. The test constructor's skills determine the quality of the items when using multiple-choice questions (Lambert & Lines, 2000; Magno, 2013; Malone, 2013), laying a solid foundation for how important it is for the constructor to have relevant skills or adhere to the appropriate guidelines when constructing MC items. As a result, the test constructor must evaluate the quality of the multiple-choice items. This implies that subpar test instruments result from poorly designed multiple-choice questions. One of the factors that can compromise a test's validity is the use of a subpar instrument (Gyamfi & Yeboah, 2022; Gyamfi & Acquaye, 2023), confirming the role of quality items in assessment. The appropriateness of the interpretation and use of assessment is evaluated by combining sources of validity evidence. The validity of test results is judged based on three sources of evidence: content-related evidence, criterion-related evidence, and construct-related evidence (Gyamfi & Yeboah, 2022). Without them, the use and interpretation of the results will be questionable.

Nitko (2012) states that tasks or items on an instrument that are relevant and reflective of the topic are considered to be content-related evidence. Whether the assessment tasks provide a representative sample from a larger domain of performance is the fundamental consideration in judgments of content representativeness. This suggests that assessment items not aligned with the syllabus or intended content domain may weaken content validity. When standardized tests are utilized, assessments of content relevance focus on whether the assessment tasks fall within the test user's domain definition. Evidence related to criteria is the kind that deals with the empirical method of examining the connection between test results or other measures (predictors) and independent external measures (criteria), such as grade point average at university and results from intelligence tests. The ability of assessment results to predict or infer an individual's position on one or more outcomes beyond the assessment process itself is known as criterion-related validity (Gyamfi & Yeboah, 2022), meaning that every valid assessment result should be able to predict future performance on a similar construct. The criterion refers to the result. Asamoah-Gyimah and Anane (2018) define construct-related evidence as the kind of evidence that relates to how well the assessment results may be understood as reflecting an individual's status regarding an educational or psychological feature, attribute, or mental process. Constructs include traits such as friendliness, honesty, creativity, reading comprehension, and mathematical thinking.

It is anticipated that the construct being tested will be the only invariant element affecting test scores based on construct-related evidence of validity. That is, the test scores should reflect an individual’s status on the construct. Construct-related factors that affect validity include but are not limited to the following, as listed by Asamoah-Gyimah and Anane (2018):

1. Unclear directions.
2. Too difficult reading vocabulary and sentence structure tends to reduce validity.
3. Ambiguous statements in assessment tasks and items
4. Inadequate time limits.
5. Inappropriate level of difficulty of the test items
6. Poorly constructed test items.
7. Test items being inappropriate for the outcomes being measured lowers validity.

8. Improper arrangement of items

It is therefore recommended, among other things, by assessment experts that items should be of high quality (Amedahe, 2014; Amedahe & Asamoah-Gyimah, 2016). In the case of multiple-choice items, adherence to the suggested principles ensures that items students are free from any invariant factors that affect the validity of assessment results (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014).

2.2 Principles of Constructing Multiple-Choice Items

To ensure that multiple-choice items are of high quality and produce valid and reliable results, assessment experts and educational measurement scholars (Brennan, 2006; Nitko, 2012; Asamoah-Gyimah & Anane, 2018) have proposed several principles for MC item construction. These principles can be grouped into several domains, including stem clarity, option quality, grammatical and structural consistency, avoidance of clues, item independence, and fairness in assessment. Table 1 summarizes the domains and principles used as the conceptual basis for evaluating MC item quality in this study.

Table 1. Principles of Multiple-Choice Item Construction

Domain	Principles of MC Item Construction
Stem clarity	The central issue of the item should be clearly stated in the stem; the stem should be concise, easy to read, and easy to understand.
Plausibility of options	Distractors should be plausible and attractive to students who have not mastered the content.
Homogeneity of options	Options should be homogeneous in content and grammatical structure.
Syntax and punctuation	All options should follow appropriate syntax and punctuation rules.
Avoidance of repetition	Repetition of words in the options should be avoided.
Avoidance of clues	Specific determiners or clues to the correct answer should be avoided.
Placement of correct answers	The placement of correct options should vary, and no predictable pattern should appear.
Originality of items	Items should not be copied directly from textbooks, previous tests, or other sources.
Logical arrangement	Options should be arranged alphabetically, numerically, or sequentially where appropriate.
Objectivity of answer	Opinion-based items should be avoided, and each item should have one clearly correct or best answer.
Item layout	Options should be arranged vertically rather than horizontally.
Parallel structure	Options should be approximately parallel in form and length.
Distinct alternatives	Options should be distinct, and overlapping alternatives should be avoided.
Use of “all/none of the above”	“All of the above” should be avoided, while “none of the above” may be used sparingly and appropriately.
Negative stems	Negative stems should be used sparingly, and words such as “not” should be emphasized when used.
Item independence	Each item should be independent, and the answer to one item should not depend on another.

Stem structure	The expected response should not be placed at the beginning of the stem.
Single correct answer	Each item should be checked to ensure that it has only one correct or best response.
Consistency of options	The number of options should be consistent across items; four or five options are suitable for higher education students.
Cross-item cueing	Items should be reviewed to ensure that the answer to one question is not revealed in another.

2.3 Teachers' competence in Constructing Multiple-Choice items and Item Quality

Ankomah (2020) examined 346 SHS instructors in the Sekondi/Takoradi Metropolis and assessed their test construction abilities, work ethics, and attitude towards test construction as indicators of adherence to test construction standards. A descriptive cross-sectional survey design was used to conduct the study. The teachers were moderately committed to their work and demonstrated a relatively high level of test construction competence.. The teachers' attitudes towards creating tests were likewise good. It was also discovered that the association between test construction skills and adherence to test construction principles is not serially mediated by job commitment or attitude towards test construction. The association between test construction skills and adherence to test construction is, however, mediated by attitude towards test construction rather than dedication to work alone. It was concluded that a positive attitude towards test construction is a requirement for adherence to test construction principles. Similar to the findings of Ankomah (2020), several studies have also investigated the test-construction skills of practicing teachers in various educational settings. For example, it has been revealed that teachers with high skills in test construction are in harmony with several studies (Afemikhe & Imobekhai, 2016; Adamu et al., 2015; Agu et al., 2013). Some other studies, however, disconfirm earlier findings (Quansah et al., 2019; Marmah & Impraim, 2013; Anhwere, 2009), suggesting that teachers lack adequate skills in test construction. The test construction skills of Senior High School (SHS) teachers in the Cape Coast Metropolis were investigated by Quansah, Amoako, and Ankomah (2019). Samples of End-of-Term Examination papers from three selected SHS in the Cape Coast Metropolis in the subjects of social studies, core mathematics, and integrated science, were chosen at random using qualitative document analysis. Experts in the field of educational measurement and evaluation rigorously analyzed the assessment tasks on the selected instruments. The findings showed that teachers' proficiency in creating end-of-term exams is lacking. This was made clear when problems were discovered with the test's representativeness and the relevance of its content, as well as the fairness and dependability of the assessment tasks. Even though the study examined the test construction skills in general, the skills of constructing MC items are embedded in the skills of constructing a test. This study focused specifically on MC items, as they are primarily used.

In a different study, 157 senior high school teachers in Ghana were asked to rate the quality of multiple-choice examinations and their ability to construct them. Kissi, Baidoo-Anu, Anane, and Annan-Brew (2023) investigated this association. The participants completed a self-created questionnaire designed to evaluate the construction competencies of multiple-choice items as taught by teachers. In this study, teachers felt more competent when it came to guaranteeing content validity than when it came to putting together test items and managing the "options" (alternatives) of the test items. Serious issues with copies of multiple-choice questions that teachers created for the pupils were also discovered by the investigation. Teachers' expression of their skills in constructing MC items would not necessarily translate into their actual competence in constructing MC items. Unlike Kissi, Baidoo-Anu, Anane, and Annan-Brew (2023), this

study sought to critically examine MC test instruments developed by teachers rather than relying on teachers' self-reported competence.

Furthermore, evidence from broader educational and professional assessment contexts indicates similar deficiencies in MC item construction. Tarrant, Knierim, Hayes, and Ware (2006), for example, examined 2,770 MCQs used in high-stakes nursing assessments and found that 46.2% of the questions contained violations of accepted item-writing guidelines, while more than 90% were written at low cognitive levels. Similarly, DiBattista and Kurzawa (2011) reported that classroom-based multiple-choice items often contained weak distractors and structural flaws that could reduce item quality. Downing (2005) also emphasized that violations of standard item-writing principles may negatively affect test scores and the validity of decisions made from assessment results. These findings suggest that weaknesses in MC item construction are not limited to teacher education institutions but represent a broader challenge in assessment practice. They also reinforce the importance of assessment literacy, systematic item review, and faculty training in improving the quality of multiple-choice assessments (Brookhart, 2011; Haladyna, Downing, & Rodriguez, 2002).

Other studies have attempted to establish whether there is a relationship between test construction skills and test quality. For instance, Zhang and Burry-Stock (2003) found a significant positive relationship between assessment skills and assessment practices among teachers. Assessment skills explained 50% of the variances in assessment practices. On the contrary, Shukla (2014) found no relationship between commitment and competency in assessment. Based on the results, Shukla concluded that competent teachers do not need to be equally committed to adopting better assessment practices. This implies that commitment is not a requirement for better assessment practices.

2.4 Implications of MCQ Construction Skills on Assessment Quality and Teacher Education

MCQ construction skills have important implications for assessment quality, teacher education, and educational policy because poorly constructed items may weaken the validity, reliability, and fairness of assessment results. When teachers possess strong MCQ construction skills, assessment instruments are more likely to reflect intended learning outcomes, appropriate cognitive levels, and accurate measures of student achievement. High-quality MCQs can also assess higher-order thinking skills such as application, analysis, and evaluation, thereby encouraging students to engage with learning beyond memorization.

In the context of teacher education, limited competence in MCQ construction among teacher educators may indicate inadequate assessment literacy within teacher preparation programmes. Since College of Education teachers serve as "trainers of trainers," weaknesses in their item-writing practices may be transferred to pre-service teachers and later reproduced in classroom assessment practices. This situation highlights the need for stronger integration of assessment literacy, item-writing practice, and test moderation within teacher education curricula.

Furthermore, deficiencies in MCQ construction skills have policy implications. They suggest the need for targeted continuous professional development, systematic moderation of teacher-made tests, and clearer institutional standards for assessment quality assurance. Strengthening MCQ construction skills is therefore not only a technical matter of item writing, but also a strategic effort to improve the credibility of assessment practices and support evidence-based decision-making in teacher education institutions.

3. Method

The quantitative document analysis design was used for the study. The purpose of this study was to analyze and evaluate documents (Quansah, Amoako, & Ankomah, 2019; Creswell, 2014; Kissi, Baidoo-Anu, Anane, & Annan-Brew, 2023). Concerning document analysis and this study, test instruments developed by teachers in the colleges of education affiliated with the Kwame Nkrumah University of Science and Technology (KNUST) were randomly selected and examined. Specifically, the test instruments were evaluated to assess the degree to which the teachers adhere to the principles of constructing multiple-choice items, as well as the quality of the items built.

A random sample was used to select three out of the five colleges affiliated with KNUST (Ary, Jacobs, Sorensen, & Razavieh, 2010). The selection of the test instrument was based on a stratified sampling approach. The department served as the strata. The items were selected from the major departments within the colleges: Education, Science, Languages, Social Studies, and Mathematics.

In each of the departments within each college, test instruments developed for a particular course in Levels 100 to 400 between 2021 and 2023 were selected. For each course, at least two instruments were selected from different years. A specific course in a discipline may have two or three additional instruments in the sample. For example, Research Methods, an educational course in Education, test instruments for 2021, 2022, and 2023 could be selected. The test instruments for the mid-semester examinations were administered internally. All test instruments had the same format: Section A for 20 multiple-choice items, Section B for four fill-in items, and Section C for a short essay. Consent of the teachers, together with the examination committee, was sought before the data was gathered. The scoring rubrics were also requested and taken from the Heads of Department. Courses were randomly selected because every teacher at the College of Education has taken at least one course in assessment, either a bachelor's degree, a master's, or both. With this, teachers know how to construct multiple-choice items.

In all, 40 test instruments were selected from each of the three colleges over five disciplines. That is, 120 test instruments were critically examined using the principles of constructing multiple-choice items as the standard. The instruments were reviewed by four assessment experts with a minimum of a master's degree in measurement and evaluation. The binary evaluation blueprint of "present and not present' was used. Each MC item was matched against all the principles as to whether the principle is present in the item or not. A moderation exercise was done, where the three assessment experts sampled some items, and all experts were made to evaluate the same items. A discussion was held to ensure agreement before the actual evaluation of all items was done. For this study, only Section A, comprising 20 multiple-choice items, was examined. A lot of effort was made to ensure the schools' anonymity, confidentiality, and privacy in the data gathered. The summary of the selection of test instruments is presented in Table 2.

Table 2. Summary of test instruments selected for analysis

College	Area	Number of Test Instruments	Number of Courses	Years Covered
College 1	Mathematics	8	4	2
	Education	10	5	2
	Science	6	2	3
	Social Studies	6	2	3
	English	10	5	2
	Total	**40**	**18**	

College 2	Mathematics	10	5	2
	Education	9	3	3
	Science	6	3	2
	Social Studies	6	2	3
	English	9	3	3
	Total	**40**	**16**	
College 3	Mathematics	10	5	2
	Education	12	4	3
	Science	6	3	2
	Social Studies	6	2	3
	English	6	3	2
	Total	**40**	**17**	
Grand Total		**120**	**51**	

Table 2 presents a summary of the test instruments selected and examined in this study. The table indicates that 40 test instruments were chosen across 18 courses, spanning Levels 100 to 400, at College 1. Five educational and English courses each were selected. Each of them had two instruments for two years, making a total of 10 instruments included in the pool. From College 2, 40 test instruments were selected across 16 courses. Five mathematics courses, each with two sets (for two different years), were selected, making it 10 instruments. At College 3, 40 instruments were chosen across 17 courses. Major were educational course. Four educational courses, each with a set from each of the three years, were selected, making a total of 12 instruments. In all, 30 of the instruments were mathematics, 31 were educational courses, 18 were science, 18 were social studies, and 25 were English courses.

4. Results

Quantitative analysis was conducted to determine the extent to which College of Education teachers adhered to established principles of constructing multiple-choice items. In total, 120 test instruments were examined, with each instrument containing 20 multiple-choice items. This resulted in a total of 2,400 items analyzed in the study.

4.1 Adherence to the Principles of Constructing Multiple-Choice Items

The first research question sought to determine the extent to which test developers adhered to the principles of multiple-choice item construction. Each of the 2,400 items was evaluated against 21 principles of MC item construction identified from the literature. For each principle, the number and percentage of items that met the criterion were calculated. A percentage-based rating scale was adopted to classify the degree of adherence into five categories, namely very low, low, moderate, high, and excellent. Table 3 presents the rating scale used in the analysis.

Table 3. Rating Scale of Test Items

S/N	Number of Items Meeting Criteria	Percentage	Interpretation
1	0–480	0–20%	Very low
2	481–960	21–40%	Low
3	961–1440	41–60%	Moderate
4	1441–1920	61–80%	High
5	1921–2400	81–100%	Excellent

All 21 principles of MC item construction were used as the criteria for evaluating the test items. Tables 4 and 5 present the extent to which the evaluated items adhered to these principles.

Table 4. Adherence to Principles of Constructing Multiple-Choice **Items**

Principles	Total	Observed	Percent	Rating
Each option must be distinct. Overlapping alternatives should be avoided.	2400	2090	87.08	Excellent
Items measuring opinions should not be included. One option should clearly be correct or the best.	2400	2004	83.50	Excellent
Create independent items.	2400	1921	80.04	High
Specific determiners which are clues to the best/correct option should be avoided.	2400	1902	79.25	High
The responses in agreement must be itemized vertically and not horizontally.	2400	1576	65.67	High
Read through all items carefully to ensure that the answer to one question is not revealed in another.	2400	1451	60.46	Moderate
The central issue of the item should be in the stem.	2400	1421	59.21	Moderate
The expected response should not be put at the beginning of the stem.	2400	1256	52.33	Moderate
Be consistent in the number of options used.	2400	1051	43.79	Moderate
Vary the placement of the correct options.	2400	987	41.13	Moderate
Check each item to make sure that there is only one correct or best response to the item.	2400	982	40.92	Low
Avoid using “all of the above” as an option, while “none of the above” may be used sparingly.	2400	980	40.83	Low
All options must follow syntax and punctuation rules.	2400	975	40.63	Low
The responses in agreement must be parallel in form; that is, sentences should be approximately the same length.	2400	784	32.67	Low
All options for a given item should be homogeneous in content.	2400	712	29.67	Low
Repetition of words in the options should be avoided.	2400	672	28.00	Low
Sentences should not be copied from textbooks or previous test items. Original items should be developed.	2400	523	21.79	Low
The options should be plausible.	2400	485	20.21	Very low
All options for a given item should be homogeneous in grammatical structure.	2400	472	19.67	Very low
Negative stems should be used sparingly, and the word “not” should be emphasized when used.	564	43	7.62	Very low
The responses/options in agreement must be in alphabetical or sequential order.	2400	140	5.83	Very low

Table 5. Summary of Rating of Adherence to Principles

Rating	Number of Principles	Percentage
Excellent	2	9.52

High	3	14.29
Moderate	5	23.81
Low	7	33.33
Very low	4	19.05
Total	21	100

Tables 4 and 5 show that only two of the 21 principles, representing 9.52%, were adhered to at an excellent level. These principles were related to ensuring that each option was distinct and avoiding opinion-based items where no clearly correct or best answer was provided. Three principles, representing 14.29%, were adhered to at a high level. These included creating independent items, avoiding specific determiners that could serve as clues to the correct answer, and arranging response options vertically.

Five principles, representing 23.81%, were adhered to at a moderate level. These included ensuring that the answer to one item was not revealed in another, placing the central issue in the stem, avoiding the placement of the expected response at the beginning of the stem, maintaining consistency in the number of options, and varying the placement of correct answers. However, the majority of the principles fell within the low and very low categories. Seven principles, representing 33.33%, were rated low, while four principles, representing 19.05%, were rated very low. The weakest areas included the plausibility of options, grammatical homogeneity of options, proper use of negative stems, and alphabetical or sequential arrangement of response options. Overall, the findings indicate that adherence to MC item construction guidelines among College of Education teachers was generally poor.

4.2 Quality of Multiple-Choice Items

The second research question sought to determine the quality of the multiple-choice items constructed and administered by College of Education teachers. Item quality was determined by **calculating** the number of principles met by each item. Items that met all 21 principles were classified as perfect-quality items, while those meeting fewer principles were categorized according to the number of criteria satisfied. The results are presented in Table 6.

Table 6. Quality of Multiple-Choice Items

Quality of Item	Number of Items	Percentage
Met all 21 principles	86	3.58
Met 16–20 principles	166	6.92
Met 11–15 principles	843	35.13
Met 6–10 principles	966	40.25
Met 1–5 principles	339	14.13
Met 0 principle	0	0.00
Total	2400	100.00

Table 5 shows that only 86 out of the 2,400 evaluated items, representing 3.58%, met all 21 principles and could therefore be classified as perfect-quality items. A further 166 items, representing 6.92%, met 16–20 principles and were classified as relatively high-quality items. Meanwhile, 843 items, representing 35.13%, met 11–15 principles, indicating moderate quality.

The majority of the items fell within the lower quality categories. Specifically, 966 items, representing 40.25%, met only 6–10 principles, while 339 items, representing 14.13%, met only 1–5 principles. Although no item failed to meet all principles, the findings show that 96.42% of the evaluated items contained at least one item-writing

flaw. This indicates that most multiple-choice items constructed and administered by College of Education teachers were of low quality and may require systematic review, moderation, and improvement before being used for assessment purposes.

5. Discussion

5.1 Adherence to Principles of Constructing Multiple-Choice Items

The study found that College of Education teachers generally showed limited adherence to established principles of MC item construction. Only five of the 21 principles, representing 23.81%, were adhered to at high or excellent levels. This finding indicates that although some aspects of MC item construction were observed, many essential principles were not consistently applied. Such limited adherence suggests that item-writing competence among College of Education teachers remains a critical concern, especially because these teachers serve as “trainers of trainers” who are expected to model sound assessment practices for prospective teachers.

This finding is consistent with previous studies which reported inadequate test construction skills among teachers (Marmah & Impraim, 2013). Similarly, Kissi, Baidoo-Anu, Anane, and Annan-Brew (2023) and Quansah, Amoako, and Ankomah (2019) found that teachers demonstrated weaknesses in test construction. While Kissi, Baidoo-Anu, Anane, and Annan-Brew (2023) used a questionnaire to assess teachers’ MC item construction competence, Quansah, Amoako, and Ankomah (2019) employed document analysis to examine teacher-made tests. The present study extends these previous works by focusing specifically on MC items and by evaluating actual test instruments constructed and administered by College of Education teachers. This document-based approach provides more direct evidence of teachers’ actual adherence to item-writing principles than studies relying solely on self-reported competence.

The findings also align with evidence from broader educational and professional assessment contexts. Tarrant, Knierim, Hayes, and Ware (2006), for example, examined 2,770 MCQs used in high-stakes nursing assessments and found that 46.2% of the questions contained violations of accepted item-writing guidelines, while more than 90% were written at low cognitive levels. Similarly, DiBattista and Kurzawa (2011) reported that classroom-based multiple-choice items often contained weak distractors and structural flaws that could reduce item quality. Downing (2005) also emphasized that violations of standard item-writing principles may negatively affect test scores and compromise the validity of decisions made from assessment results. These findings suggest that weaknesses in MC item construction are not limited to teacher education institutions but represent a broader challenge in assessment practice. They also reinforce the importance of assessment literacy, systematic item review, and faculty training in improving the quality of multiple-choice assessments (Brookhart, 2011; Haladyna, Downing, & Rodriguez, 2002).

These findings suggest that the problem is not merely individual but systemic. Limited adherence to MC item construction principles may reflect gaps in assessment literacy, teacher education curricula, institutional support, and item review mechanisms. In particular, weaknesses such as implausible distractors, poorly structured options, lack of grammatical homogeneity, and inadequate arrangement of alternatives suggest that teachers may not be applying item-writing principles systematically during test development. This condition is concerning because MC item construction is not only a technical task but also a validity-related process that determines whether assessment results can be interpreted fairly and accurately.

However, evidence from intervention-based studies suggests that weaknesses in MC item construction can be improved through structured faculty development and targeted training. Abdulghani et al. (2015), for example, conducted a quasi-experimental study to examine the effect of long-term faculty development programs on the quality of MCQ item writing. Their findings showed that post-training MCQs demonstrated improvement in difficulty index, higher cognitive levels based on Bloom's taxonomy, and fewer item-writing flaws and nonfunctioning distractors. This suggests that limited competence in MC item construction is not necessarily permanent but can be addressed through systematic training, continuous professional development, and structured item review. Similarly, Ankomah (2020) reported that teachers possessed a high level of test construction skills, while Afemikhe and Imobekhai (2016) and Adamu et al. (2015) also reported relatively positive findings. Nevertheless, the positive findings from self-reported or questionnaire-based studies should be interpreted with caution because such approaches may capture teachers' perceived knowledge of item construction principles rather than the actual quality of the items they construct and administer. Studies based on document analysis or direct item evaluation tend to provide more concrete evidence of teachers' actual adherence to test construction principles.

5.2 Quality of Multiple-Choice Items

The study also revealed that the quality of MC items constructed by College of Education teachers was generally low. Only 86 out of the 2,400 items, representing 3.58%, met all 21 principles and could be classified as excellent-quality items. This finding indicates that the majority of the items contained at least one item-writing flaw. Therefore, the low quality of the items appears to be closely connected to teachers' limited adherence to MC item construction principles.

This finding supports Zhang and Burry-Stock (2003), who found a significant positive relationship between assessment skills and assessment practices among teachers. Their findings suggest that teachers with weak assessment skills are more likely to demonstrate poor assessment practices. This interpretation is also supported by Armah (2018) and Gyamfi, Loganathan, and Acquaye (2023), who emphasized the importance of teachers' assessment competence in producing quality assessment outcomes. In the context of the present study, the low proportion of excellent-quality items suggests that many teachers may not be translating assessment knowledge into effective item-writing practice.

However, the relationship between teachers' knowledge, commitment, and actual assessment practice may not always be straightforward. Although professional competence and commitment are important aspects of teacher effectiveness, they do not automatically guarantee high-quality assessment practices (Shukla, 2014). Therefore, teachers' perceived knowledge or commitment should be interpreted cautiously unless supported by direct evidence from the quality of the assessment items they construct. These findings imply that possessing knowledge of MC item construction principles does not automatically result in high-quality item writing. Although teachers may have taken assessment-related courses (Yeboah, Gyamfi, & Sam, 2019), their knowledge may not be adequately reflected in the actual items they construct. For this reason, Ankomah (2020), Baidoo-Anu (2022), and Baidoo-Anu and Ennu Baidoo (2022) argued that teachers' attitudes, commitment, and professional practices may also influence the quality of assessment items.

The poor quality of the MC items has important implications for assessment validity, reliability, and fairness. Poorly constructed items may fail to measure the intended learning outcomes, provide unintended clues to students, or assess irrelevant aspects of knowledge. As a result, students' scores may not accurately represent their actual

standing on the construct being measured. This threatens the validity of assessment results and may lead to inaccurate educational decisions. In teacher education institutions, such decisions may affect guidance and counselling, selection, placement, certification, and instructional management.

Furthermore, the findings point to the need for a more structured system of item review and moderation. Constructing high-quality MC items requires time, expertise, and careful review. If teachers are required to prepare assessment instruments without adequate time, institutional support, or expert moderation, the quality of the items may be compromised. Therefore, the low quality of MC items found in this study may reflect broader systemic gaps in assessment training, curriculum implementation, and institutional quality assurance. Strengthening item-writing competence should therefore be treated as both a professional development priority and a policy issue in teacher education. Targeted continuous professional development, regular item moderation, and the involvement of assessment experts in reviewing test instruments may help improve the quality, validity, and fairness of assessments administered to students.

6. Conclusion

The findings revealed that College of Education teachers showed limited adherence to established principles of multiple-choice item construction. Although some principles were followed at high or excellent levels, the overall level of adherence was generally low. The results also showed that only 3.58% of the 2,400 evaluated items met all 21 principles, while 96.42% contained at least one item-writing flaw. These findings indicate that most of the multiple-choice items constructed and administered by College of Education teachers were of low quality. The poor quality of the items has important implications for assessment validity, reliability, and fairness. If multiple-choice items are poorly constructed, students' scores may not accurately reflect their actual knowledge, skills, or achievement. This may lead to inaccurate decisions in areas such as guidance and counselling, selection, placement, certification, and instructional management. The study therefore contributes document-based evidence on actual item-writing practices among College of Education teachers and highlights the need for stronger assessment quality assurance in teacher education institutions.

Based on the findings, it is recommended that lessons on multiple-choice item construction should be integrated into Continuous Professional Development (CPD) programmes for College of Education teachers. In addition, assessment experts should be involved in academic affairs units to review and moderate test instruments before they are administered to students. Colleges of Education should also establish a structured item review system to ensure that assessment instruments meet acceptable standards of quality, validity, and fairness. This study was limited to document-based analysis of multiple-choice items and did not examine the underlying factors that may explain teachers' limited adherence to item-writing principles. Therefore, further empirical research is recommended to investigate possible contributing factors, such as assessment training, workload, institutional support, teachers' assessment literacy, and item moderation practices. Such studies would help inform more effective interventions and policies for improving assessment quality in teacher education institutions.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., Aldrees, A. A., Alrowais, N., & Haque, S. (2015). Faculty development programs

improve the quality of multiple choice questions items' writing. *Scientific Reports, 5*, Article 9556. https://doi.org/10.1038/srep09556

Adamu, G. G., Dawha, J. M., & Kamar, T. S. (2015). A scheme for assessing technical teachers' competencies in constructing assessment instruments in technical colleges in Gombe State. *ATBU Journal of Science, Technology & Education, 3*(2), 1–8.

Afemikhe, O. A., & Imobekhai, S. Y. (2016). *Nigerian teachers' utilization of test construction procedures for validity improvement of achievement tests* [Unpublished paper]. Institute of Education, University of Benin.

Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Educational Research and Reviews, 8*(8), 431–439. https://doi.org/10.5897/ERR12.219

Amedahe, F. K. (2014). Test construction practices in secondary schools in the Central Region of Ghana. *The Oguaa Educator, 2*, 52–63.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anhwere, Y. M. (2009). *Assessment practices of teacher training college tutors in Ghana* [Unpublished master's thesis]. University of Cape Coast.

Ankomah, F. (2020). *Predictors of adherence to test construction principles: The case of senior high school teachers in Sekondi-Takoradi Metropolis* [Unpublished master's thesis]. University of Cape Coast.

Armah, C. (2018). *Test construction and administration practices among lecturers and staff of examinations unit of the University of Cape Coast in Ghana* [Unpublished master's thesis]. University of Cape Coast.

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

Asamoah-Gyimah, K., & Anane, E. (2018). *Assessment in schools* [Unpublished mimeograph]. University of Cape Coast.

Awofala, A. O. A., & Babatunde, V. F. T. (2013). Examining attitude towards continuous assessment practices among Nigerian pre-service STM teachers. *Journal of Education and Practice, 4*(13), 37–49.

Baidoo-Anu, D. (2022). Between-school streaming: Unpacking the experiences of secondary school teachers and students in category C schools in Ghana. *International Journal of Educational Research Open, 3*, Article 100188. https://doi.org/10.1016/j.ijedro.2022.100188

Baidoo-Anu, D., & Ennu Baidoo, I. (2024). Performance-based accountability: Exploring Ghanaian teachers' perception of the influence of large-scale testing on teaching and learning. *Education Inquiry, 15*(3), 333–350. https://doi.org/10.1080/20004508.2022.2110673

Baidoo-Anu, D., Asamoah, D., Quainoo, E. A., Gyamerah, K., Amoateng, E. Y., & Sasu, E. O. (2023). Emergency remote assessment practices in higher education in sub-

Saharan Africa during COVID-19. *Frontiers in Education, 8*, Article 1221115. https://doi.org/10.3389/feduc.2023.1221115

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). American Council on Education/Praeger.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). American Council on Education/Praeger.

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

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), Article 4. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

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. *Advances in Health Sciences Education, 10*(2), 133–143. https://doi.org/10.1007/s10459-004-4019-5

Erzoah, K. K., Gyamfi, A., Yeboah, A., & Langee, P. (2022). Teachers' knowledge and practices of classroom assessment in the Ellembelle District of Ghana. *Advances in Research, 23*(4), 1–10. https://doi.org/10.9734/air/2022/v23i430337

Gyamfi, A. (2022). Application of classical test theory to the validation of teacher-made mathematics multiple-choice test items. *Asian Journal of Advanced Research and Reports, 16*(11), 1–12.

Gyamfi, A. (2023). Differential item functioning of performance-based assessment in mathematics for senior high schools. *Jurnal Evaluasi dan Pembelajaran, 5*(1), 20–34.

Gyamfi, A., & Acquaye, R. (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis, 13*(3), 68–78. https://doi.org/10.2478/atd-2023-0022

Gyamfi, A., Loganathan, S., & Acquaye, R. (2023). Improving teachers' classroom assessment practices: Perceptions of teachers in the Ellembelle District of Ghana. *e-mentor, 4*(101), 56–62. https://doi.org/10.15219/em101.1625

Gyamfi, A., & Yeboah, A. (2022). The changing phase of validity: The past and now. *Global Scientific Journals, 10*(6), 1016–1023.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hamafyelto, R. S., Hamman-Tukur, A., & Hamafyelto, S. S. (2015). Assessing teacher competence in test construction and content validity of teacher-made examination questions in commerce in Borno State, Nigeria. *Journal of Education, 5*(5), 123–128.

Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in examination-oriented educational system: Exploring

teachers' multiple-choice test construction competence. *Frontiers in Education, 8*, Article 1154592. https://doi.org/10.3389/feduc.2023.1154592

Lambert, D., & Lines, D. (2000). *Understanding assessment: Purposes, perceptions, practice*. RoutledgeFalmer.

Magno, C. (2013). Standards of teacher competence on student assessment in the Philippines. *The Assessment Handbook, 10*, 42–53.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344. https://doi.org/10.1177/0265532213480129

Marmah, A. A., & Impraim, A. K. (2013). University lecturer's competence in the construction of multiple-choice test items: A case study of Coltek Kumasi. *The International Journal of Humanities & Social Studies, 1*(4), 1–9.

Ministry of Education, Ghana. (2017). *National teachers' standards for Ghana: Guidelines*. Ministry of Education.

Nitko, A. J. (2012). *Educational assessment of students* (6th ed.). Pearson/Prentice Hall.

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

Quansah, F., & Amoako, I. (2018). Attitude of senior high school teachers toward test construction: Developing and validating a standardised instrument. *Research on Humanities and Social Sciences, 8*(1), 25–30.

Quansah, F., Amoako, I., & Ankomah, F. (2019). Teachers' test construction skills in senior high schools in Ghana: Document analysis. *International Journal of Assessment Tools in Education, 6*(1), 1–8. https://doi.org/10.21449/ijate.481164

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation, 22*(1), Article 4. https://doi.org/10.7275/swgt-rj52

Shukla, S. (2014). Teaching competency, professional commitment and job satisfaction: A study of primary school teachers. *IOSR Journal of Research & Method in Education, 4*(3), 44–64. https://doi.org/10.9790/7388-04324464

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*(8), 662–671. https://doi.org/10.1016/j.nedt.2006.07.006

Wiredu, S. G. (2013). *Assessment practices of tutors in the nurses' training colleges in the Western and Central regions of Ghana* [Unpublished master's thesis]. University of Cape Coast.

Yeboah, A., Gyamfi, A., & Sam, N. I. (2019). Relevance of assessment course: A follow-up study of graduate teachers in Ghana. *Academia Journal of Educational Research, 9*(9), 302–310.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342. https://doi.org/10.1207/S15324818AME1604_4