

Evaluating the Quality of Multiple-Choice Assessment: Evidence from Bhutan

Yangdon¹, Tshering Dorji², Tashi², Sonam Zangmo²

Samtse College of Education, Royal University of Bhutan

yangdon.sce@rub.edu.bt

ABSTRACT

The study explored the quality of multiple-choice items used in written tests in the Postgraduate Diploma Programme at Samtse College of Education, Royal University of Bhutan. Although multiple-choice items are widely used in the programme, their quality and tutors' understanding of effective item construction have not been systematically examined. This gap is important because poorly constructed items may affect the accuracy, fairness, and effectiveness of assessment. Using a convergent parallel mixed-methods design, quantitative data were obtained through item analysis of MCIs using difficulty and discrimination indices, while qualitative data were collected through semi-structured interviews and document analysis and analysed thematically. The findings showed that participants demonstrated limited understanding of effective multiple-choice item construction. The items also showed low discrimination and limited adherence to prescribed construction principles. The study recommends targeted training to strengthen tutors' knowledge and skills in writing effective multiple-choice items. The study contributes to assessment literature by linking multiple-choice item quality and educators' item-writing knowledge to professional development.

Keywords: assessment, multiple-choice items, semi-structured interview, document analysis

INTRODUCTION

Multiple-choice items (MCIs) are used in the Postgraduate Diploma in Education (PgDE) programme at Samtse College of Education (SCE), Royal University of Bhutan (RUB). A review of question papers from the previous two semesters showed that twelve modules included MCIs, with weightings ranging from 10% to 20% of overall assessment. MCIs consist of a stem and a set of options from which students select the best or correct answer (Torres et al., 2011). They are attractive to instructors because they can be graded quickly and can cover a broad range of course content. When well-constructed, they can also produce scores as reliable as those from constructed-response tests (Bacon, 2003). However, the quality MCIs used in the PgDE semester examinations has not been systematically evaluated.

Previous research has examined MCI quality in relation to item construction, difficulty, discrimination, distractor functioning, and adherence to established item-writing principles. These studies indicate that poorly constructed items can affect the validity, fairness, and interpretation of assessment results. However, limited evidence is available on the quality of MCIs used in teacher education in Bhutan, particularly in the PgDE programme at SCE. More specifically, there is limited evidence on the difficulty and discrimination of these items, their adherence to item-writing principles, and tutors' understanding of MCI construction. This gap highlights the need for the present study to generate evidence on MCI quality, identify areas for improvement, and inform professional development and assessment practices.

This study explored the module tutors' understanding and practices concerning the construction and utilisation of MCIs. Additionally, it assessed the quality of MCIs used in the semester examinations of the PgDE programme. The anticipated outcomes of this research include gaining insights into the quality of MCIs used in the examinations and identifying gaps in the module tutors' knowledge and abilities in constructing MCIs. Furthermore, the findings have the potential for broader applicability across other colleges within the RUB. Ultimately, the results will aid faculty members in enhancing their expertise in MCI construction, leading to improvements in assessment practices. It will also add to the literature on assessment both nationally and internationally.

The study, therefore, addresses two connected problems: whether the multiple-choice items used in PgDE semester examinations function effectively as assessment tools, and whether module tutors have the knowledge required to construct items that are valid, discriminating, and aligned with established item-writing principles. Clarifying these issues is important because poorly constructed items may distort evidence of student learning and weaken assessment decisions. By examining MCI quality and educators' item-writing knowledge, the study offers insights beyond Bhutan for strengthening the quality, fairness,

and effectiveness of multiple-choice assessment through systematic item analysis and professional development.

RESEARCH QUESTIONS

Primary question

What is the quality of multiple-choice items developed and used by the module tutors in the semester-end examinations of the Postgraduate Diploma in Education programme?

Sub-questions

1. How do module tutors understand multiple-choice items?
2. What are the difficulty indices of the multiple-choice items?
3. What are the discrimination indices of the multiple-choice items?
4. What is the construct validity of the multiple-choice items as reflected in their adherence to established item-writing principles?

LITERATURE REVIEW

Multiple-choice Items

Multiple-choice items are widely used in college and university assessments and often contribute substantially to students' course grades (Mavis et al., 2001; McDougall, 1997). An MCI typically consists of a stem, one correct answer, and several distractors. Because MCIs can be designed to assess cognitive processes ranging from recall to higher-order thinking, their usefulness depends heavily on the quality of item construction (Torres et al., 2011).

Construction of MCIs

Ngaji et al. (2021) contend that specific sets of skills are required to construct MCIs while Gajjar et al. (2014) report that MCIs that accurately measure students' knowledge and differentiate their abilities are imperative. Effective MCIs measure higher-level thinking, improve student learning and performance, strengthen tutor efficiency, and offer a better understanding of assessment outcomes (Simkin & Kuechler, 2005). Several guidelines are required to construct effective MCIs. For instance, Begum and Slavin (2012) propose a clear alignment between the stem and the expected learning outcomes. The test items should be grammatically consistent, maintain focus, and provide adequate information for students to choose the correct answer without relying solely on the distractors. Additionally, MCIs must be framed as direct questions or incomplete statements and use positive language (Sireci et al., 1998). Once the stem is created, the role of distractors becomes crucial (Miller & Erickson, 1985). Distractors should effectively divert uncertain students, with one choice clearly representing the correct response while the remaining alternatives appear plausible. The arrangement of answer choices should follow alphabetical, chronological, or numerical

order (Haladyna et al., 2002). According to Bloom's taxonomy, a well-constructed MCI measures different cognitive skills such as comprehension, application, analysis, and synthesis (Mahjabeen et al., 2017). By conforming to these guidelines and addressing the various cognitive levels, MCIs can effectively assess students' understanding and critical thinking.

Principles of MCI Constructions

A set of principles govern the construction of MCIs (Downing & Haladyna, 1997). Non-adherence to the principles can lead to decreased validity evidence for tests (Downing, 2006) and are often artificially more difficult, leading to lower passing rates and the introduction of construct-irrelevant variance (CIV) into the test (Downing, 2005). CIV refers to the inclusion or exclusion of test scores due to the measurement of constructs or traits not directly targeted by the test itself (Messick, 1989). According to Tara et al. (2022), it is crucial to possess an understanding of the principles for constructing MCIs to ensure that they accurately test candidates' knowledge. One way to equip educators with the necessary knowledge is through training (Emily et al., 2015). Jia et al. (2020) further note that teachers should be trained to develop and construct MCIs as strong writing skills are essential for improving the quality of these items. Their study emphasises the importance of teacher commitment and efforts in quality assurance to meet professional standards and obligations. Haladyna et al. (2002) discovered that flawed item writing can significantly impact students' performance on MCIs, making the items either more or less difficult to answer accurately.

Advantages and Disadvantages of MCIs

The use of multiple-choice items as an assessment tool has both advantages and disadvantages. Parmenter (2009) highlights students' preference for MCIs due to the opportunities they provide for higher scores and their perceived validity compared to other assessment methods. A notable advantage of MCIs is the quick and easy grading process, making them appealing for courses with large enrollments. Well-constructed MCIs can yield reliable test scores while allowing for broader coverage of course topics (Bacon, 2003). Furthermore, MCIs can measure a variety of educational objectives and subject matter at different cognitive levels, ease of administration with large groups, consistent scoring, reliable assessment of student progression, reduced grading bias and test-taking anxiety, and easy implementation with computer-based testing, providing immediate results or feedback (Begum & Slavin, 2012; Boland et al., 2010).

However, MCIs are criticised for their focus on memorisation rather than evaluating higher-order cognitive skills (Walsh & Seldomridge, 2006). Framing MCIs that accurately assess higher-order thinking can be challenging (Tulving, 1967). Additionally, MCIs do not measure students' creativity, articulation of ideas, explanations, or the ability to

justify answer responses (Burton et al., 1991). Guessing and ruling out answers without substantive knowledge are also concerns (Funk & Dickson, 2011). However, thoughtfully written MCIs can be used to assess higher-level cognitive processes, although creating such items requires more skill than writing memory-based items (Buckles & Siegfried, 2006). Another criticism is that the format of MCIs allows for guessing, even when students lack substantive knowledge of the topic being assessed (Biggs, 1999). In contrast, Downing (2003) argues that blind guessing is rare on well-written classroom tests, and informed guessing, based on critical consideration of the question and available options, can provide a valid measure of student achievement. Carneson et al. (2016) express that MCIs with a single correct answer are an effective way to assess students' cognitive knowledge. Despite the merits and demerits, MCIs continue to be widely used to assess student achievement in postsecondary education globally, including across academic programmes at SCE.

Item Difficulty Index

The item difficulty index refers to the percentage of the total number of students who answered the test question correctly (Tavakol & Dennick, 2011). Similarly, Laxman and Vasant (2018) defined the difficulty level/index as the percentage of the total number of students from both the groups, high achievers and low achievers opting for key, i.e. answering correctly representing the difficulty index. It allows for differentiation between students who are high performing and those who are not, or between 'strong' and 'weak' students (Tavakol & Dennick, 2011). However, according to Gajjar et al. (2014), an ideal item will be the one that has average difficulty ranging between 31 and 60 percent with high discrimination index ≥ 0.25 . Menon and Kannambra (2017) pointed out that an item with average difficulty ranging from 30-70 percent and high discrimination 0.2 is considered an ideal MCI.

Construct Validity of MCIs

Construct validity refers to the extent to which a test measures what it claims to measure (Anastasi, 1986). Julie et al. (2005) emphasise that construct validity is closely linked to whether the questions effectively measure the domain of knowledge under examination. To ensure construct validity, several criteria have been identified. Mahmoud (2017) has established 22 item construction principles to evaluate the structure of MCIs, while Marie et al. (2006) provide nineteen guidelines for effective MCI construction. Brady (2005) emphasises that MCIs consistent with educational outcomes can effectively assess students' performance, offer timely feedback, and contribute to independent and self-directed learning. However, it is crucial to note that the construction of high-quality MCIs requires a significant commitment and alignment with curriculum objectives. Brown and Abdulnabi (2017) found that instructor-constructed MCIs in higher education courses often exhibited poor quality, impacting students' overall grades, especially in determining pass/

fail decisions. Hence, it is essential to evaluate MCIs before finalising grading decisions, and eliminate low-quality items prior to the test.

Przymuszała et al. (2020) suggested that providing teachers with a concise guideline document can help reduce the number of item-writing flaws in MCIs. They further recommend continuous professional development to enhance teachers' knowledge and awareness of good practices in MCI writing and item-writing flaws. Similarly, Tarrant and Ware (2008) emphasise the significance of providing training in writing high-quality test items. They advocate for thorough examination and review of MCIs both before and after their administration. Standardising MCIs is considered a mandatory initial step in achieving quality assessment, and one effective approach involves regular evaluation of questions through item and test analysis.

Analysis of the Quality of MCIs

Analysing the quality of MCIs can be approached in various ways. In this study, the approach adopted is based on Item Response Theory (IRT), which assesses the difficulty level and discrimination power of items to determine their quality. Additionally, the study employed well-established principles of MCI development to examine the construct validity of the items. Sajjad et al. (2020) conducted a study that revealed the impact of item writing errors on the efficiency of distractors in MCIs. Analysing MCIs can help identify such errors and improve the overall quality of the examination by enhancing the efficiency of distractors. Mehmood et al. (2021) also highlight the importance of post hoc analysis of MCIs, which allows examiners to identify any shortcomings or flaws encountered during the construction process. By applying the IRT-based analysis and considering established principles of MCI development, this study aimed to assess the quality and construct validity of MCIs.

METHODOLOGY

This study is grounded in pragmatism, a research paradigm that evaluates ideas and beliefs in terms of their practical consequences (Dieleman et al., 2017). A pragmatic orientation supported the selection of methods, techniques, and research procedures suited to the study purpose (Creswell & Creswell, 2018). Guided by this paradigm, the study employed a convergent parallel mixed-methods design. Quantitative and qualitative data were collected concurrently, analysed separately, and then integrated to develop a fuller understanding of the quality of MCIs and tutors' item-construction practices. The quantitative findings were presented alongside qualitative findings to support interpretation, comparison, and validation across data sources.

The study examined module tutors' understanding of multiple-choice items and analysed the quality of MCIs used in PgDE semester examinations. Semi-structured interviews

were conducted with 11 participants who had teaching experience in the programme. The interview questions focused on tutors’ conceptual understanding of MCIs, their uses, and principles of item construction. The study also conducted document analysis to examine the extent to which examination items adhered to established item-writing principles.

MCIs used in 12 PgDE semester examinations papers, comprising 240 multiple-choice items, were analysed. The quantitative data consisted of 72 corresponding students answer scripts for these examinations. For item analysis, students were first ranked according to their total examination scores and then divided using Kelley’s (1939) 27% upper-lower group method. The highest-performing 27% of students formed the upper group while the lowest-performing 27% formed the lower group. The middle 46% were excluded from the discrimination. This method is recommended since it maximises the distinction between the high and low performing students while maintaining adequate sample sizes for item analysis (Tavakol & Dennick, 2011). The upper and lower groups were subsequently used to calculate the Difficulty Index (P) and Discrimination Index (D) for each multiple-choice item.

The quality of the multiple-choice items was analysed by calculating the Difficulty Index (P), Discrimination Index (D), and adherence to item-writing guidelines. The Difficulty Index (P) represents the proportion of students who answered an item correctly and indicates the relative ease or difficulty of the item (Tavakol & Dennick, 2011). The P value was calculated using the formula:

$$P= Ri/N,$$

Where, P is the fraction of the correct answer

R is the total number of students who answered the question correctly

N is the total number of responses (correct + incorrect + blank response) and

i refers to the question number.

The interpretation of the Difficulty Index (P) is outlined in Table 1. In order to determine the P value for each module, the responses of the top and bottom groups using 27% method, the data were computed in an Excel sheet. The P value was calculated separately for the top and bottom groups to assess the difficulty level of the modules for both groups.

Table 1: Interpretation of Item Difficulty (P) Index

P score	Interpretation
< 30%	Difficult
= 30- 70%	Ideal/Acceptable
>70%	Very Easy

Note. Adapted from the study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot by Chauhan et al. (2013, p. 38) and Item Analysis of Multiple-choice Questions: Assessing an Assessment Tool in Medical Students by Chandrika (2016, p. 203).

The item-discrimination index is a value of how well a question is able to differentiate between students who are high performing and those who are not, or between ‘strong’ and ‘weak’ students (Tavakol & Dennick, 2011). According to Kumar et al. (2020), item-discrimination index is interpreted as: $DI \leq 0.20$ (poor), 0.21- 0.24 (acceptable), $DI 0.25-0.34$ (good) and $DI \geq 0.35$ (excellent). This study adopted the formula given by Zaman and Faize (2010) to calculate the Discrimination Index (D).

$$DI = \frac{Ru - Rl}{N}$$

where, DI = discrimination index

Ru = the number of students in the upper group who responded correctly

Rl = the number of students in the lower group who responded correctly.

The study reviewed a list of guidelines and principles for effective MCI item construction used in different studies conducted by Downing (2004), Mahmoud (2017) and Marie et al. (2006). From the list, the study adapted 12 guidelines for constructing effective MCIs items. The MCIs of all the 12 modules were evaluated and coded 1 for adherence to the guidelines and 0 for non-adherence of the guidelines. To enhance the construct validity of the process, the evaluated MCIs were cross-checked among researchers before entering into the final result sheet.

The qualitative data were analysed thematically using the method developed by Braun and Clarke (2018). The detailed procedures are discussed below.

- i. All transcripts were read and reread multiple times. In the process of reading and re-reading, potential codes for each topic were created and recorded.
- ii. Before they were finalised, a rigorous exercise of adding, deleting, renaming, and combining codes was performed. In addition, the codes have been reviewed to see if they are related to the themes and if there are clear differences between them.
- iii. Relevant categories for the codes were created and entered after reading and rereading all transcripts several times.
- iv. All relevant information/data segments were entered into the codes.
- v. A report was generated on the analysis, which also included the data extract.

Although the analysis process was described as linear, it was an iterative process, with stages moving back and forth as needed. Finally, the data obtained from all sources were merged and triangulated to confirm the merging of findings and it aims to reduce biases (Denzin, 1989) and strengthen reliability or dependability of the study.

ETHICAL CONSIDERATIONS

The ethical clearance from the Dean of Research and Industrial Linkage's Office was sought by submitting duly completed ethical clearance application form. All the ethical standards as outlined in the application form were adhered to during the course of the study. In addition, administrative clearance was sought from the Dean of Academic Affair for the use of the semester question papers and students' answer scripts. In the data presentation and discussions, the identity of the module tutors, question papers and students' answer scripts are strictly protected. To maintain anonymity of the participants, pseudonyms were used: PA, PB, PC etc., where P stands for participant.

FINDINGS AND DISCUSSION

This section presents the analyses of the data that are collated, merged, and triangulated to better understand the stated issue.

Understanding of Multiple-choice Items

The qualitative data analysis revealed a diverse range of interpretations among participants regarding their understanding of multiple-choice items. For instance, PA described MCIs as consisting of a stem, correct answer, and distractors, emphasising the need for strong distractors that closely resemble the correct answer to ensure the right answer is not easily revealed. Congruently, literature defines multiple-choice items as assessment items comprising a stem followed by a list of possible responses, where one option is correct while the others act as distractors (Torres et al., 2011). According to PH, MCIs are an effective tool for assessing students' understanding across various topics. Other participants highlighted additional perspectives on MCIs: PA considered them the best form of objective items; PJ noted their inclusivity and ability to cater to individual needs; PE regarded them as a means to assess different levels of thinking; and PD saw them as a summative assessment tool and a suitable substitute for essay-type questions. Literature supports the idea that MCIs measure a wide range of educational objectives and subject matter at different cognitive levels (Begum & Slavin, 2012; Boland et al., 2010). While a few participants acknowledged the difficulty and time-consuming nature of constructing MCIs, they also recognised that MCIs are relatively easy to assess. It was asserted by some participants that well-constructed MCIs can effectively measure higher-level thinking abilities such as comprehension, analysis, and cognitive reasoning. Overall, while most participants recognised the general characteristics and uses of MCIs, their responses revealed gaps in understanding effective item construction. Knowledge of general features does not necessarily imply the ability to construct high-quality items. Effective construction requires understanding how stems, alternatives, distractors, and cognitive demands contribute to valid evidence of learning. The quantitative findings further highlight these gaps, particularly in item difficulty, discrimination, and adherence to item-writing guidelines.

Difficulty Level of MCIs

The difficulty level of multiple-choice items plays a crucial role in accurately assessing students' performance and effectively distinguishing between high and low achievers. One key factor that influences MCI difficulty is the cognitive complexity of the questions. Sato (2015) suggests that questions involving higher-order thinking skills, such as analysis and evaluation, tend to be more challenging compared to those focused on recall or comprehension. This distinction in difficulty allows for the differentiation of high achievers from low achievers (Gierl & Haladyna, 2013). To analyse the quantitative data, the difficulty index (P) from Chauhan et al. (2013) was utilised, as presented in Table 2.

Table 2: Difficulty Index (P) of 12 Modules Offered at SCE

Module		p Value	Average p Value
Top Group		Bottom Group	
M1	85.71	81.43	83.57
M2	82.90	75.00	78.90
M3	70.70	57.90	64.30
M4	81.40	77.90	79.60
M5	87.10	65.70	76.40
M6	90.00	62.86	76.43
M7	98.57	82.14	90.35
M8	82.86	69.29	76.07
M9	78.57	70.00	74.29
M10	72.50	62.50	67.50
M11	51.67	38.33	45.00
M12	78.57	57.86	68.21

As evident from the table, the majority of the modules did not meet the desired difficulty level for multiple-choice items. A total of 240 MCIs were analysed, and it was discovered that 229 items had plausible distractors, with each MCI having four distractors. Out of 12 modules, only four modules (M3, M10, M11, and M12) accounted for 33.33% of the total and fell within the acceptable or ideal category. These modules demonstrated a difficulty index ranging from 45 to 68.21. Literature suggests that questions that are excessively difficult or easy can have a negative impact on student motivation and engagement (Alqahtani, 2015). On the other hand, the remaining eight modules (M1, M2, M4, M5, M6, M7, M8, and M9) accounted for 66.67% of the total and were categorised as "very easy". This classification implies that these modules failed to effectively differentiate between high achievers and low achievers.

The qualitative finding showed that almost all participants reported limited knowledge and skills required to accurately construct multiple-choice items. Consequently, this could result in the creation of MCIs that are either too easy or too difficult. Furthermore, participants expressed their significant challenges in constructing MCIs. This idea is encapsulated in the following quote by PF: "I find the construction of MCIs very challenging." Given their reported inadequate knowledge and skills, participants sought assistance from senior colleagues or referred to past question papers to construct MCIs. They attributed their lack of knowledge to lack of training on the development of MCIs. Therefore, PI suggested that opportunities for professional development would be beneficial in helping them construct robust MCIs. This suggestion aligns with the findings of Emily et al. (2015), who emphasised the importance of training educators in order to equip them with the necessary knowledge to construct MCIs. Simkin and Kuechler (2005) elaborated that proper training in formatting and creating high-quality MCIs can lead to improved student learning and performance, increased instructor efficiency, and a better understanding of assessment outcomes. Hence, constructing effective MCIs that accurately assess students' knowledge and differentiate their abilities is essential (Gajjar et al., 2014). It is evident from the prior discussion that 66.67% of MCIs fell into the "very easy" category, which can be attributed to participants' inadequate knowledge and skills in developing them.

Therefore, the findings highlight the need for tutors to develop items across an appropriate range of difficulty and cognitive demand rather than relying predominantly on items that assess relatively straightforward knowledge and understanding.

Discrimination Power of the MCIs

The discrimination index of 12 modules is shown in Table 3.

Table 3: Discrimination Index (D) of 12 Modules at SCE

Module		pValue	D value (Discrimination Index)
Top Group		Bottom Group	
M1	85.71	81.43	0.043
M2	82.90	75.00	0.079
M3	70.70	57.90	0.129
M4	81.40	77.90	0.035
M5	87.10	65.70	0.214
M6	90.00	62.86	0.271

Module		pValue	D value (Discrimination Index)
Top Group		Bottom Group	
M7	98.57	82.14	0.160
M8	82.86	69.29	0.136
M9	78.57	70.00	0.085
M10	72.50	62.50	0.100
M11	51.67	38.33	0.117
M12	78.57	57.86	0.207

As shown in Table 3, most modules demonstrated weak discrimination. Nine modules, representing 75%, had D values at or below 0.20 and therefore fell into the poor category. Two modules were in the acceptable range, and one module (M6) reached the good category with a D value of 0.271. Overall, the mean D value of 0.13 indicates that the MCIs generally had limited capacity to distinguish between higher- and lower-performing students. This finding is consistent with the high average P value of 73.39%, which suggests that many items were too easy to discriminate effectively among students.

Literature indicates that an effective measure of multiple-choice items is their ability to assess higher cognitive skills, such as data interpretation, synthesis, and knowledge application, rather than solely testing factual recall (Elgadal & Mariod, 2021). However, the qualitative findings showed that the MCIs constructed by the participants focused mainly on lower levels of thinking. For example, one participant mentioned, "When I prepare MCIs, it is mostly from the lower order thinking." The participants justified by stating that questions based on lower order thinking are easier to create, while constructing questions that require higher order thinking is more difficult and challenging. Providing questions that only assess lower order thinking skills will fail to effectively differentiate between high performing and low performing students. The ability of a question to differentiate high achievers from low achievers is crucial for the validity of assessments and decision-making regarding students' growth. Failure to employ testing strategies that accurately represent students' knowledge and skills can lead to flawed interpretations of their performance (Hoban et al., 2017). Thus, the low discrimination observed across most modules represents more than a statistical weakness; it raises concerns about the extent to which examination results provide precise and meaningful evidence of differences in student achievement.

Item construction Guideline Adherence of MCIs

The item construction guideline adherence of MCIs is provided in Table 4.

Table 4: Item Construction Guidelines Adherence

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	Total	
Item Guideline	20	20	20	20	20	20	20	20	20	20	20	20	240	Mean
G1- Present a definite, explicit, and singular question or problem in the stem.	19	20	18	20	19	20	20	20	20	20	18	20	234	19.50
G2- When possible, state the stem as a direct question rather than as an incomplete statement.	7	9	7	6	14	6	5	4	17	18	18	14	125	10.42
G3- Provide clear directions; no ambiguity about what is being asked.	20	20	17	19	20	20	20	20	20	20	20	20	236	19.67
G4- Include in the stem any word(s) that might otherwise be repeated in each alternative.	14	20	14	16	19	14	13	12	20	13	20	16	191	15.92
G5- Use negatively stated stems sparingly, when used, underline and/or capitalise the negative word, capitalise the number.	14	18	18	19	20	7	5	5	16	20	12	16	170	14.17

G6- Avoid unintentional clues to the correct answer such as unmatched grammar, unrelated choices, use of "always/never," "all of the above," "none of the above".	13	6	19	19	18	0	5	6	7	15	10	19	137	11.42
G7- Make all alternatives plausible.	20	17	20	20	15	20	30	19	19	16	16	17	229	19.08
G8- Make alternatives approximately equal in length or arrange them logically, alphabetically, or numerically.	11	10	10	16	14	9	3	17	10	19	12	6	137	11.42
G9- Capitalise the beginning of each answer choice when the stem is an incomplete statement.	7	20	9	16	19	18	15	3	19	17	14	17	174	14.50
G10-End each answer choice that is a sentence with a period.	6	20	20	4	17	14	27	16	20	12	20	15	191	15.92
G11- Balance the placement of key/correct answer.	13	20	16	11	20	12	18	14	20	20	14	20	198	16.50
G12- Avoid questions of the form "Which of the following statements is correct?"	13	20	12	11	18	17	27	28	20	20	20	19	225	18.75
Total														
Total	157	200	180	177	213	157	188	164	208	210	196	199	2247	

The quantitative results indicated that a significant number of questions did not meet the guidelines for constructing MCIs. Among the guidelines, G3 (Provide clear directions; no ambiguity about what is being asked) had the highest guideline adherence (GA) with a mean value of 19.67 (98.3%), followed by G1 (Present a definite, explicit, and singular question or problem in the stem) with a GA value of 19.50 (97.5%). G4 (Include in the stem any word(s) that might otherwise be repeated in each alternative), G5 (Use negatively stated stems sparingly, when used, underline and/or capitalize the negative word, capitalise the number), G9 (Capitalise the beginning of each answer choice when the stem is an incomplete statement), and G10 (End each answer choice that is a sentence with a period) had average GA values of 15.92 (79.6%), 14.17 (70.8%), 14.50 (72.5%), and 15.91 (79.6%) respectively. However, G2 (When possible, state the stem as a direct question rather than as an incomplete statement), G6 (Avoid unintentional clues to the correct answer such as unmatched grammar, unrelated choices, use of "always/never," "all of the above," "none of the above"), and G8 (Make alternatives approximately equal in length or arrange them logically, alphabetically, or numerically) had the lowest GA values with 10.42 (52.1%), 11.42 (57.1%), and 11.42 (57.1%) respectively. This indicates that non-adherence to G2, G6, and G8 is common among the tutors when designing MCIs.

During the qualitative analysis, it was discovered that almost all participants were unaware of the principles involved in constructing multiple-choice items. One participant, PD, explicitly stated, "I have no idea of any principles." Some participants associated the principles with the test blueprint, while others identified the equal distribution of MCIs across the six levels of thinking as the principles. Literature suggests the use of specific research-based principles guide the development of effective MCIs (Downing & Haladyna, 1997) but deviating from established item-writing principles can lead to a decrease in validity evidence for tests (Downing, 2006). Additionally, items that violate one or more of the standard item writing principles tend to artificially increase the difficulty level, resulting in lower passing rates and introducing construct-irrelevant variance (CIV) to the test (Downing, 2005).

Although the participants admitted their lack of awareness regarding the principles of constructing MCIs, they unanimously acknowledged the benefits of employing such principles. They emphasised that following the principles ensures that the questions are flawless, objective, unbiased, and assess the intended content accurately. Clarity of the questions was also highlighted by most participants. Participants expressed the importance of adhering to the principles for the sake of objectivity and authentic assessment. Some participants mentioned that the guidelines help achieve the intended learning outcomes and enhance understanding of the module's content. The design and use of valid and reliable multiple-choice items are crucial for ensuring the validity of assessment results (Amit et al., 2003) so that they accurately test candidates' knowledge (Tara et al., 2022).

The participants also noted that failing to follow the principles would result in poorly constructed MCIs, which would have several negative impacts. These impacts include consuming more time, failing to achieve learning outcomes, increasing the likelihood of guessing correct answers, confusing students, and the questions not measuring what they are intended to measure. Furthermore, participants pointed out that not adhering to the principles increases the chances of students obtaining correct answers (PA) and engaging in malpractices and plagiarism during exams, such as copying, non-verbal communication, and sharing correct answers. Congruent with this finding, Biggs (1999) highlighted that multiple-choice item have the potential to encourage students to make guesses, even when they lack substantial knowledge about the subject being assessed.

Taken together, the quantitative and qualitative findings indicate that weaknesses in item-writing guideline adherence may affect the validity, fairness, and interpretability of MCI-based assessment. The findings therefore suggest that strengthening tutors' knowledge of item-writing principles should be accompanied by systematic review and moderation of assessment items.

LIMITATIONS OF THE STUDY

The study was limited to the Postgraduate Diploma Programme at Samtse College of Education, which may restrict the generalisability of the findings to other programmes and institutions. It also focused on selected indicators of MCI quality and relied on interviews and document analysis. The study did not examine other psychometric indicators such as distractor efficiency and test reliability, which could provide more comprehensive assessment of item quality.

CONCLUSION AND RECOMMENDATION

The study explored the MCIs used in the test of the PgDE programme at SCE. Based on the findings, it is concluded that participants have limited knowledge and skills in framing effective MCIs. For the construction of effective MCIs, educators must be aware of construct validity indicators (e.g., difficulty index, discrimination value), high-quality distractors, and principles of item construction. The ability to construct MCIs that measure deeper thinking can enhance student learning and achievement, and increase teacher effectiveness. Therefore, educators must be proficient in developing effective MCIs which demands specific knowledge and skills. Hence training is recommended to enable them to frame effective MCIs. This training should focus on understanding the key components of MCIs, such as stems, correct answers, and distractors and aligning MCIs with desired learning outcomes and educational objectives. Overall, the findings address the research problem by demonstrating that weaknesses in MCI difficulty, discrimination, and adherence to item-writing principles limit the effectiveness of PgDE assessments, highlighting the

need to strengthen educators' knowledge and skills in MCI construction. The MCIs used in the PgDE assessment were predominantly in the "very easy" category and demonstrated low discrimination, indicating weaknesses in their ability to differentiate students' levels of understanding and adequately assess the intended learning outcomes. The findings also indicate a need to strengthen educators' knowledge and skills in constructing effective MCIs, particularly in developing appropriate item difficulty, plausible distractors, and items that assess both lower- and higher-order thinking. Therefore, targeted and regular professional development on MCI construction and item analysis is strongly recommended. Such professional development should be complemented by opportunities for educators to share good practices and provide constructive feedback on assessment items. Strengthening these competencies is essential to ensure that PgDE assessments are valid, reliable, fair, and aligned with intended learning outcomes. Limited familiarity of the guidelines in framing effective MCIs have resulted in non-compliance with the principles. Failure to follow these guidelines can undermine assessment quality by affecting validity, reliability and fairness and the ability of the MCIs in assessing higher order thinking abilities of the students. Moreover, it may also increase ambiguity and encourage guessing. Therefore, it is imperative for educators to integrate the guidelines into the MCIs to uphold validity, fairness, and reliability.

REFERENCES

- Alqahtani, M. (2015). The importance of motivation in learning English as a foreign language in Saudi Arabia. *English Language Teaching*, 8(3), 1–8. Canadian Center of Science and Education.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.
- Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marketing context. *Journal of Marketing Education*, 25(1), 31–36.
- Begum, S., & Slavin, H. (2012). Perceptions of caring in nursing education by Pakistani Nursing Students: An exploratory study. *Nurse Education Today*, 32, 332–336. <https://doi.org/10.1016/j.nedt.2011.10.011>
- Biggs, J., & Collis, K. (1999). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.
- Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple choice questions. *Academic Psychiatry*, 34(4), 310–316.

- Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, 5(4), 238–242.
- Braun, V., & Clarke, V. (2018). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. doi:10.1191/1478088706qp063oa
- Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, 2(24).
- Buckles, S., & Siegfried, J. J. (2006). Using in-depth multiple-choice questions to evaluate in-depth learning economics. *Journal of Economic Education*, 37(1), 48–57.
- Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services and the Department of Instructional Science.
- Carneson, J., Delpierre, G., & Masters, K. (2016). *Designing and managing multiple-choice questions*. University of Cape Town. <https://doi.org/10.13140/RG.2.2.22028.31369>
- Chauhan, C., Zrally, C. B., & Dingwall, A. K. (2013). The Drosophila COMPASS-like Cmi-Trr coactivator complex regulates dpp/BMP signaling in pattern formation. *Developmental Biology*, 380(2), 185–198. <http://dx.doi.org/10.1016/j.ydbio.2013.05.018>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approach*. Sage Publication.
- Dieleman, J. L., Squires, E., Bui, A. L., Campbell, M., Chapin, A., Hamavid, H., Horst, C., Li, Z., Matyas, T., Reynolds, A., Sadat, N., Schneider, M. T., & Murray, C. J. L. (2017). Factors associated with increases in US health care spending, 1996–2013. *JAMA*, 318(17), 1668–1678. <https://doi.org/10.1001/jama.2017.15927>
- Downing, S. M. (2003). Guessing on selected-response examinations. *Medical Education*, 37, 670–671. <http://dx.doi.org/10.1046/j.1365-2923.2003.01585.x>
- Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, 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. *Advances in Health Sciences Education*, 10, 33–143.
- Downing, S. M. (2006). Selected response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Lawrence Erlbaum Associates Publishers.

- Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10(1), 61–82.
- Elgadal, A. H., & Mariod, A. A. (2021). Item analysis of multiple-choice questions: Assessment tool for quality assurance measures. *Sudan Journal of Medical Sciences (SJMS)*, 16(3), 334–346. <https://doi.org/10.18502/sjms.v16i3.9695>
- Emily, M. W., Jonathan S. P., & David. M. (2015). Does educator training or experience affect the quality of multiple-choice questions? *Academic Radiology*, 22(10), 1317–1322.
- Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, 38(4), 273–277.
- Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Community*, 39, 17–20.
- Gierl, M. J., & Haladyna, T. M. (2013). The importance and use of the discrimination index for test item analysis. *Journal of Educational Research*, 59(3), 321–337.
- Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4), 999–1010.
- 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), 309–33. https://doi.org/10.1207/S15324818AME1503_5
- Hoban, G., Neil, S., & Larkin, S. (2017). Failure to employ testing strategies that accurately represent students' knowledge and skills can lead to flawed interpretations of their performance. *Journal of Educational Assessment*, 24(3), 230–245.
- Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. *Problems of Education in the 21st Century*, 78(4), 576–594.
- Julie, H. M., Todd, T. W., & Petersen, R. (2005). Construct validity in psychological research. *Journal of Personality and Social Psychology*, 89(5), 10–21.
- Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2020). *Item analysis of multiple choice questions: A quality assurance test for an assessment tool*. <https://doi.org/10.1016/j.mjafi.2020.11.007>
- Laxman, V. C., & Vasant, V. M. (2018). Item analysis to identify quality multiple-choice questions in pharmacology examination. *Journal of Chemical and Pharmaceutical Research*, 10(9), 50–54.

- Mahjabeen, W., Alam, S., & Hassan, U. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Ann Pak Inst Med Sci.*, 310–315.
- Mahmoud, B. H. (2017). Flaws of multiple-choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, 56(8), 490–496.
- Marie, B., Lepine, J. A., & Johnson, R. E. (2006). The construct validity of integrity tests: A review and illustration of key item and scale development principles. *Journal of Business and Psychology*, 21(2), 185–208. doi:10.1007/s10869-0069017-5
- Mavis, B. E., Cole, B. L., & Hoppe, R. B. (2001). A survey of student assessment in U.S. medical schools: The balance of breadth versus fidelity. *Teaching and Learning in Medicine*, 13(2), 74– 79.
- McDougall, P. P., & Oviatt, B. M. (1997). *International entrepreneurship literature in the 1990s and directions for future research*. In D. L. Sexton & R. W. Smilor (Eds.), *Entrepreneurship 2000* (pp. 291–320). Upstart Publishing.
- Mehmood, B., Sultan, S., Azhar, M., Shakoor, A., Khan, N. B., & Niazi, Z. (2021). *Quality analysis of multiple-choice items*, 40(4), 236–39.
- Menon, A. R., & Kannambra, P. N. (2017). *Item analysis to identify quality multiple choice questions*. DOI:10.7860/njlm/2017/25690:2215
- Miller, P. W., & Erickson, H. E. (1985). *How to write tests for students*. National Education Association.
- Ngaji, O. O., Ngaji, O., A. V., Eduwem, D. J., & Amba, J. N. (2021). An assessment of science teachers' utilization of the knowledge of test construction procedure in multiple choice objective tests in secondary schools in Yala LGA, Cross River State. *British Journal of Education*, 9(11), 54–62.
- Parmenter, D. A. (2009). Essay versus multiple-choice: Student preferences and the underlying rationale with implications for test construction. *The Academy of Educational Leadership Journal*, 13(2), 57–71.
- Przymuszała, P. (2020). *Guidelines on writing multiple-choice items: A well-received and effective faculty development intervention*. <https://doi.org/10.1177/2158244020947432>
- Sajjad, M., Iltaf, S., & Khan, R. A. (2020). *Non-functional distractor analysis: An indicator for quality of multiple-choice items*. doi:<https://doi.org/10.12669/pjms.36.5.2439>
- Sato, K. (2015). Effects of microglia on neurogenesis. *Glia*, 63(8), 1394–1405. <https://doi.org/10.1002/glia.22858>

- Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Journal of Innovative Education*, 3, 73–98.
- Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J. H. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.), *Test interpretation and diversity* (pp. 105–140). American Psychological Association.
- Tara, C. D., Sheryl, G. J., & Petra, J. L. (2022). Great question. The art and science of crafting high quality multiple-choice questions. *Journal of the American College of Radiology*, 19(6), 687–692.
- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2), 198–206. <https://doi.org/10.1111/j.1365-2923.2007.02957.x>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's Alpha. *International Journal of Medical Education*, 2, 53–55.
- Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review*, 1, 1–11.
- Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 6(2), 175–184.
- Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). Writing multiple-choice test items. In G. R. Walz & J. C. Bleuer (Eds.), *Assessment: Issues and challenges for the millennium* (pp. 215–222). Greensboro, NC: ERIC Clearinghouse on Counseling and Student Services.
- Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Nursing Education*, 45, 212–219.
- Zaman, A., & Faize, A. F. (2010). Analysis of multiple-choice items and the effect of items' sequencing on difficulty level in the test of mathematics. *European Journal of Social Sciences*, 17(1), 61–67.